

Omics technologies in livestock improvement: From selection to breeding decisions

Edited by

Mudasir Ahmad Syed, Marcos De Donato, Basharat Ahmad Bhat, Abdoulaye Baniré Diallo and Sunday O. Peters

Published in

Frontiers in Genetics
Frontiers in Veterinary Science



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83251-340-8
DOI 10.3389/978-2-83251-340-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Omics technologies in livestock improvement: From selection to breeding decisions

Topic editors

Mudasir Ahmad Syed — Sher-e-Kashmir University of Agricultural Sciences and Technology Shuhama Srinagar Jammu and Kashmir, 190006, India

Marcos De Donato — The Center for Aquaculture Technologies, United States

Basharat Ahmad Bhat — University of Otago, New Zealand

Abdoulaye Baniré Diallo — Université du Québec à Montréal, Canada

Sunday O. Peters — Berry College, Georgia, United States

Citation

Syed, M. A., De Donato, M., Bhat, B. A., Diallo, A. B., Peters, S. O., eds. (2023). *Omics technologies in livestock improvement: From selection to breeding decisions*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-340-8

Table of contents

- 07 **Editorial: Omics technologies in livestock improvement: From selection to breeding decisions**
Syed Mudasir Ahmad, Marcos De Donato, Basharat Ahmad Bhat, Abdoulaye Banire Diallo and Sunday O. Peters
- 13 **Genome-Wide Association Studies Reveal Susceptibility Loci for Noninfectious Claw Lesions in Holstein Dairy Cattle**
Ellen Lai, Alexa L. Danner, Thomas R. Famula and Anita M. Oberbauer
- 28 **Genomic Prediction of Average Daily Gain, Back-Fat Thickness, and Loin Muscle Depth Using Different Genomic Tools in Canadian Swine Populations**
Siavash Salek Ardestani, Mohsen Jafarikia, Mehdi Sargolzaei, Brian Sullivan and Younes Miar
- 41 **Omics Multi-Layers Networks Provide Novel Mechanistic and Functional Insights Into Fat Storage and Lipid Metabolism in Poultry**
Farzad Ghafouri, Abolfazl Bahrami, Mostafa Sadeghi, Seyed Reza Miraei-Ashtiani, Maryam Bakherad, Herman W. Barkema and Samantha Larose
- 57 **Genome-Wide Scanning for Signatures of Selection Revealed the Putative Genomic Regions and Candidate Genes Controlling Milk Composition and Coat Color Traits in Sahiwal Cattle**
Satish Kumar Illa, Sabyasachi Mukherjee, Sapna Nath and Anupama Mukherjee
- 71 **Changthangi Pashmina Goat Genome: Sequencing, Assembly, and Annotation**
Basharat Bhat, Nazir A. Ganai, Ashutosh Singh, Rakeeb Mir, Syed Mudasir Ahmad, Sajad Majeed Zargar and Firdose Malik
- 80 **Genes and Pathways Affecting Sheep Productivity Traits: Genetic Parameters, Genome-Wide Association Mapping, and Pathway Enrichment Analysis**
Seyed Mehdi Esmaeili-Fard, Mohsen Gholizadeh, Seyed Hasan Hafezian and Rostam Abdollahi-Arpanahi
- 95 **SNPs in Mammary Gland Epithelial Cells Unraveling Potential Difference in Milk Production Between Jersey and Kashmiri Cattle Using RNA Sequencing**
Syed Mudasir Ahmad, Basharat Bhat, Shakil Ahmad Bhat, Mifftha Yaseen, Shabir Mir, Mustafa Raza, Mir Asif Iqbal, Riaz Ahmad Shah and Nazir Ahmad Ganai
- 104 **Genome-Wide Association Studies in Indian Buffalo Revealed Genomic Regions for Lactation and Fertility**
Vikas Vohra, Supriya Chhotaray, Gopal Gowane, Rani Alex, Anupama Mukherjee, Archana Verma and Sitangsu Mohan Deb

- 120 **Insights Into mRNA and Long Non-coding RNA Profiling RNA Sequencing in Uterus of Chickens With Pink and Blue Eggshell Colors**
Siyu Chen, Kecheng Chen, Jiaming Xu, Fangwei Li, Jinlong Ding, Zheng Ma, Gen Li and Hua Li
- 132 **Genome-Wide Association Studies for Growth Curves in Meat Rabbits Through the Single-Step Nonlinear Mixed Model**
Yonglan Liao, Zhicheng Wang, Leonardo S. Glória, Kai Zhang, Cuixia Zhang, Rui Yang, Xinmao Luo, Xianbo Jia, Song-Jia Lai and Shi-Yi Chen
- 141 **Single-cell RNA Sequencing Reveals Heterogeneity of Cultured Bovine Satellite Cells**
Pengcheng Lyu, Yumin Qi, Zhijian J. Tu and Honglin Jiang
- 149 **FMixFN: A Fast Big Data-Oriented Genomic Selection Model Based on an Iterative Conditional Expectation algorithm**
Wenwu Xu, Xiaodong Liu, Mingfu Liao, Shijun Xiao, Min Zheng, Tianxiong Yao, Zuoquan Chen, Lusheng Huang and Zhiyan Zhang
- 159 **Circular RNA Expression and Regulation Profiling in Testicular Tissues of Immature and Mature Wandong Cattle (*Bos taurus*)**
Ibrar Muhammad Khan, Hongyu Liu, Jingyi Zhuang, Nazir Muhammad Khan, Dandan Zhang, Jingmeng Chen, Tengting Xu, Lourdes Felicidad Córdova Avalos, Xinqi Zhou and Yunhai Zhang
- 176 **Transcriptome-Wide Analyses Identify Dominant as the Predominantly Non-Conservative Alternative Splicing Inheritance Patterns in F1 Chickens**
Xin Qi, Hongchang Gu and Lujiang Qu
- 190 **Pleiotropic Loci Associated With Foot Disorders and Common Periparturient Diseases in Holstein Cattle**
Ellen Lai, Alexa L. Danner, Thomas R. Famula and Anita M. Oberbauer
- 201 **Significance and Relevance of Spermatozoal RNAs to Male Fertility in Livestock**
Bijayalaxmi Sahoo, Ratan K. Choudhary, Paramajeet Sharma, Shanti Choudhary and Mukesh Kumar Gupta
- 219 **Comprehensive Analysis of mRNA, lncRNA, circRNA, and miRNA Expression Profiles and Their ceRNA Networks in the *Longissimus Dorsi* Muscle of Cattle-Yak and Yak**
Chun Huang, Fei Ge, Xiaoming Ma, Rongfeng Dai, Renqing Dingkao, Zhuoma Zhaxi, Getu Burechao, Pengjia Bao, Xiaoyun Wu, Xian Guo, Min Chu, Ping Yan and Chunnian Liang
- 233 **Integrated Analysis of mRNA and MicroRNA Co-expressed Network for the Differentiation of Bovine Skeletal Muscle Cells After Polyphenol Resveratrol Treatment**
Dan Hao, Xiao Wang, Yu Yang, Bo Thomsen, Lars-Erik Holm, Kaixing Qu, Bizhi Huang and Hong Chen

- 246 **Characterization of XR_311113.2 as a MicroRNA Sponge for Pre-ovulatory Ovarian Follicles of Goats via Long Noncoding RNA Profile and Bioinformatics Analysis**
Hu Tao, Juan Yang, Pengpeng Zhang, Nian Zhang, Xiaojun Suo, Xiaofeng Li, Yang Liu and Mingxin Chen
- 258 **Function Identification of Bovine ACSF3 Gene and Its Association With Lipid Metabolism Traits in Beef Cattle**
Wei He, Xibi Fang, Xin Lu, Yue Liu, Guanghui Li, Zhihui Zhao, Junya Li and Runjun Yang
- 272 **Effects of Heat Stress on the Ruminal Epithelial Barrier of Dairy Cows Revealed by Micromorphological Observation and Transcriptomic Analysis**
Zitai Guo, Shengtao Gao, Jun Ding, Junhao He, Lu Ma and Dengpan Bu
- 281 **Skeletal Muscle Expression of Actinin-3 (ACTN3) in Relation to Feed Efficiency Phenotype of F₂ *Bos indicus* - *Bos taurus* Steers**
Robert N. Vaughn, Kelli J. Kochan, Aline K. Torres, Min Du, David G. Riley, Clare A. Gill, Andy D. Herring, James O. Sanders and Penny K. Riggs
- 292 **Genome Wide Scan to Identify Potential Genomic Regions Associated With Milk Protein and Minerals in Vrindavani Cattle**
Akansha Singh, Amit Kumar, Cedric Gondro, A. K. Pandey, Triveni Dutt and B. P. Mishra
- 300 **Whole Genome Sequencing Analysis to Identify Candidate Genes Associated With the rib eye Muscle Area in Hu Sheep**
Yuan Zhao, Xiaoxue Zhang, Fadi Li, Deyin Zhang, Yukun Zhang, Xiaolong Li, Qizhi Song, Bubo Zhou, Liming Zhao, Jianghui Wang, Dan Xu, Jiangbo Cheng, Wenxin Li, Changchun Lin, Xiaobin Yang, Xiwen Zeng and Weimin Wang
- 310 **Genome-Wide Association Study and F_{ST} Analysis Reveal Four Quantitative Trait Loci and Six Candidate Genes for Meat Color in Pigs**
Hang Liu, Liming Hou, Wuduo Zhou, Binbin Wang, Pingping Han, Chen Gao, Peipei Niu, Zongping Zhang, Qiang Li, Ruihua Huang and Pinghua Li
- 321 **Whole-Genome-Based Web Genomic Resource for Water Buffalo (*Bubalus bubalis*)**
Aamir Khan, Kalpana Singh, Sarika Jaiswal, Mustafa Raza, Rahul Singh Jasrotia, Animesh Kumar, Anoop Kishor Singh Gurjar, Juli Kumari, Varij Nayan, Mir Asif Iquebal, U. B. Angadi, Anil Rai, Tirtha Kumar Datta and Dinesh Kumar
- 334 **Improvement of Genomic Predictions in Small Breeds by Construction of Genomic Relationship Matrix Through Variable Selection**
Enrico Mancin, Lucio Flavio Macedo Mota, Beniamino Tulliozi, Rina Verdiglione, Roberto Mantovani and Cristina Sartori

352 Applications of Omics Technology for Livestock Selection and Improvement

Dibyendu Chakraborty, Neelesh Sharma, Savleen Kour, Simrinder Singh Sodhi, Mukesh Kumar Gupta, Sung Jin Lee and Young Ok Son

368 Locating a novel autosomal recessive genetic variant in the cattle glucokinase gene using only WGS data from three cases and six carriers

Geoffrey E. Pollott, Richard J. Piercy, Claire Massey, Mazdak Salavati, Zhangrui Cheng and D. Claire Wathes



OPEN ACCESS

EDITED AND REVIEWED BY
Martino Cassandro,
University of Padua, Italy

*CORRESPONDENCE
Syed Mudasir Ahmad,
✉ mudasirbio@gmail.com

SPECIALTY SECTION
This article was submitted
to Livestock Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 01 December 2022

ACCEPTED 13 December 2022

PUBLISHED 04 January 2023

CITATION
Ahmad SM, De Donato M, Bhat BA,
Diallo AB and Peters SO (2023), Editorial:
Omics technologies in livestock
improvement: From selection to
breeding decisions.
Front. Genet. 13:1113417.
doi: 10.3389/fgene.2022.1113417

COPYRIGHT
© 2023 Ahmad, De Donato, Bhat, Diallo
and Peters. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Editorial: Omics technologies in livestock improvement: From selection to breeding decisions

Syed Mudasir Ahmad^{1*}, Marcos De Donato^{2,3},
Basharat Ahmad Bhat⁴, Abdoulaye Banire Diallo⁵ and
Sunday O. Peters⁶

¹Division of Animal Biotechnology, Faculty of Veterinary Sciences and Animal Husbandry, SKUAST, Kashmir, India, ²Escuela de Ingeniería y Ciencias, Instituto de Tecnología y Educación Superior de Monterrey (ITESM), Monterrey, Mexico, ³The Center for Aquaculture Technologies, San Diego, CA, United States, ⁴Departments of Surgical Sciences, University of Otago, Dunedin, New Zealand, ⁵Department of Computer Science, Université du Québec à Montréal, Montreal, QC, Canada, ⁶Department of Animal Science, Berry College, Mount Berry, GA, United States

KEYWORDS

livestock, epigenomics, transcriptomics, GWAS, polymorphism, RNA sequencing

Editorial on the Research Topic

[Omics technologies in livestock improvement: From selection to breeding decisions](#)

Livestock rearing is the main component of global food production systems and contributes to nearly half of global agricultural production. The emerging demand for animal products is considered a food revolution in developing countries. Although livestock farming is an essential component of the global economy but faces an immense challenge to meet out the increased demand of the exploding human population and other environmental factors. As a result, understanding animal health and production is becoming mandatory. Advances in genetic and genomic technologies have played a key role in improving animal welfare and productivity for decades. Genetic development in livestock, both in terms of productivity and other functional trait complexes associated with health and animal welfare, has been greatly aided by genomic selection, which is regarded as a success story. Genomic breeding programmes offer the potential to improve cattle productivity through the utilization of molecular genetics, the detection of markers and chromosomal areas that contain quantitative trait loci, and genome mapping technology. Present breeding strategies not only account for phenotypic variance in traits but also other factors like epigenetics. Epigenetic mechanisms (DNA methylation, RNA methylation, histone modifications, chromatin remodelling, and non-coding RNA regulation) have been strongly found to influence livestock traits like growth, development, and other phenotypic effects. Technologies related to omics are developing more frequently in the area of animal production. Other omics topics like phosphoproteomics, peptidomics, or lipidomics are presently used in livestock production and disease management in addition to omics

methodologies such as transcriptomics, genomics, proteomics, and metabolomics. Animal breeding systems already in place now have additional dimensions due to the introduction of revolutionary ideas like animal cloning and genetic engineering. It provides a faster way to increase the frequency of desirable alleles in an animal population using genetically modified animals than traditional breeding procedures. With this background, the aim of the current research topic was to collect research data on the role of omics technologies in livestock improvement from selection to breeding decisions.

The special edition of *Frontiers in Genetics* i.e., “Omics Technologies in Livestock Improvement: From Selection to Breeding Decisions” published 26 original research articles, 2 review papers, and 01 methodology paper. All 29 articles were focused mainly on understanding the complex interplay of genes in mediating important traits, elucidating livestock genome/transcriptome, and improving livestock yield, nutrition values, and animal welfare.

Omics technologies are effective tools, especially when used in conjunction with advanced molecular and breeding methods. Genomic research in particular has the potential to improve the precision and efficacy of traditional breeding and advanced breeding approaches by enhancing consistency and predictability. Indeed, the information provided by genomics and other omics-technologies like epigenomics and proteomics is crucial for understanding genes and their function. Additionally, summarized the role of omics technologies in improving livestock (Chakraborty et al.).

Male infertility still poses a significant problem for animals, despite significant increases in reproductive efficiency brought about by artificial insemination procedures. Recent research indicates that the spermatozoa of fertile and infertile males have different populations of RNA. Studies have also shown that spermatozoal RNA (spRNA) is essential for spermatogenesis, fertilisation, and the early stages of embryonic development. Updated various spRNA species in livestock animals, including protein-coding and non-coding RNAs and their possible impact on quality of sperms, especially motility, freezability, and fertility (Sahoo et al.). The regulatory structure of circRNA in testis development and spermatogenesis in cattle bulls was evaluated using RNA sequencing by Khan et al., in immature and mature Wandong cattle bull testes. In calf and bull testes, 579 of the 17,013 circRNAs were elevated and 103 were downregulated. Bull spermatogenesis and the genes ATM, GSK3B, CCNA1, KMT2C, NSD2, KMT2E, QKI, SUCLG2, HOMER1, and SNAP91 were discovered to be strongly correlated. Through genetic selection, this information may aid in enhancing bulls' reproductive productivity (Khan et al.).

Deep RNA sequencing was utilised to discover probable single nucleotide polymorphisms (SNP) in mammary

epithelial cells from two different cow breeds (Jersey and Kashmiri) in order to study the alterations in coding regions that affect milk output differences. Ahmad et al., work seeks to find high-impact SNPs in Jersey and Kashmiri cows in order to uncover the main pathways controlling milk production features in each breed using RNA-Seq data (Ahmad et al.). There were discovered to be 684 (464 SNPs and 220 INDELs) and 607 (442 SNPs and 169 INDELs) high-impact variations unique to Kashmir and Jersey cattle, respectively. The RNA sequencing-based SNP research revealed a considerable difference between Kashmiri and Jersey cattle in terms of milk production attributes. The high-impact SNP variations in Kashmiri cattle contributed to adaptive immunity and tolerance to infectious diseases of the mammary gland. Contrarily, in Jersey cattle, enhanced pathways were mostly implicated in production and lactation maintenance. These findings offer information on breed-specific genetic diversity that can be applied to the genomic selection of animals. There are many studies like RNA sequencing leads and high-throughput genomic DNA data, which have identified the genes related to lipid storage in broilers. Using RNA-Seq and microarray data techniques, Farzad Ghafouri et al. identified and categorised candidate genes and miRNAs involved in lipid metabolism. Based on the amount of abdominal fat present, two broiler groups were selected: high and low. In the analysis, 34 genes and 19 miRNAs were detected as common. A total of seven genes were revealed as common, three being most important viz., REBF1, SREBF2, SCD, and FASN. These have a significant impact on fat metabolism, storage, and signalling pathways of endocrine glands that are triggered by PPAR, AMPK, and adipokines. This method improved our comprehension of the biological mechanisms affecting adipose tissues (Ghafouri et al.). In animal food industry, the genetics and physiological mechanisms that govern skeletal muscle mass production are of prime importance to study. Satellite cells are crucial for skeletal muscle growth and regeneration. scRNA-seq was used by Lyu et al. to analyse the makeup of bovine satellite cells. According to the findings, bovine satellite cells may be divided into subpopulations with different transcriptional statuses, rates of proliferation, and myogenic potential. The current study by Lyu et al. also indicates the existence of FAP cells in bovine skeletal muscle, which may help researchers devise new methods for enhancing these features or find the DNA sequences and variants linked to them in cattle (Lyu et al.).

Lameness in cattle is frequently brought on by two non-infectious claw lesions: sole ulcers (SU) and white line disease (WLD). Sensitivity to SU and WLD is influenced by both hereditary and non-genetic variables, and protection can be accomplished by genetic methods and herd management. Genome-wide association studies (GWAS) were conducted using generalised linear mixed model (GLMM) regression,

random forest (RF) and a chunk-based association testing (CBAT) technique to find susceptibility loci for WLD, SU, SU and/or WLD, and any sort of non-infectious claw damage. Potential genes for skeletal growth and mineralization, keratin, adipose tissue, wound healing and skin lesions were present in the linked areas on BTA8 and BTA13. The abundance of correlations found in this study conducted by Lai et al., as well as previous studies that showed the intricacy of the genetic background underlying non-infectious claw lesion vulnerability (Lai et al.). Lameness in an animal incurs financial burden on the animal owner, making it the third largest cause of culling an animal after mastitis and infertility. This condition is caused by foot lesions which can be either infectious or non-infectious in nature. Digital dermatitis (DD), WLD and SU are the most frequent causes. A study was performed by Lai et al. to find the loci influencing the association between the genetic causes of lameness and other prevalent health conditions. Estimates of the genetic connection between DD and SU, WLD and SU were all substantially different from zero ($p < 0.05$), whereas estimates of the genetic association between mastitis and DD, SU with metritis, and DD with milk fever were just suggestive ($p < 0.1$). These estimations of the five genetic correlations were all positive. Lai et al., observed that selection against disorders of foot may also reduce vulnerability to other health disorders because of the significant genetic association estimations between foot disorders and other illnesses (Lai et al.).

GWAS provide opportunities for comprehensive improvement of buffalo by finding genetic variants connected with lactation and reproductive features. The work conducted by Vohra et al., intends to find novel SNPs linked with milk output, lactation consistency, milk composition and fertility features at the genomic level in Murrah buffalo using the genotype-by-sequencing technique. To run GWAS on fat percentages and SNF percentages independently, more than 38,000 SNPs were employed. GWAS was also conducted on 305 days' worth of milk output to test lactation consistency. SNPs were found to be substantially linked with the first principal component, which explained the greatest fraction of variability in milk output. Breeding effectiveness, post-partum breeding interval, and age at sexual maturity were all taken into account as fertility features. Additionally, certain putative genetic areas that may play a part in controlling milk production and fertility in Murrah were found (Vohra et al.). Singh et al., conducted a GWAS to find substantially related SNPs between proteins in milk and minerals in cattle. Identifying potential genes connected to milk minerals and milk protein percentage in Vrindavani cattle was the goal of their study. Protein percentage Ca, Cu, Zn, P and Fe were found to have a strong association with BTA 7, 2, 3, 14, and 2, respectively (Singh et al.). Further, consumers mostly judge the quality of meat based on its colour. In order to aid in pig breeding programs, Liu et al. conducted a study to identify the

genes associated with meat colour in Suhui pigs. Additionally, it calculates the genetic correlation and heredity of meat colour. (Liu et al.). Using GWAS and FST testing on Suhui pigs, six potential genes (PIK3CG, HOMER1, VCAN, PIK3CA, FKBP1B, and FABP3) and 39 possible SNPs associated to flesh colour were found. These findings pave the way for genetic modification of pig flesh colour since they have functional implications for muscle growth, lipid binding and phosphatidylinositol phosphorylation. For sheep breeding industry, ewe productivity is considered the most important economic trait. Genes influencing fertility and lamb growth following parturition in Iranian Baluchi sheep were discovered using GWAS methods and gene set enrichment analysis (GSEA). The notable SNPs within or close to the genes RDX, ARHGAP20, FDX1, THBS1, ZC3H12C, and EPG5 were linked to composite features at birth. The identified pathways and genes have roles relevant to pregnancy, such as autophagy in the placenta, calcium ion transport, placental development and progesterone release and maternal immunological response. Genes NR2C1, HSD17B4, RSU1, VEZT, CUBN, PRLR, VIM, and FTH were discovered to be associated with composite features at weaning. According to Esmaeili-Fard et al., the findings imply that calcium ion transport throughout pregnancy and milk feeding lambs following parturition had the greatest influence on weight gain in comparison to other maternal origin effects (Esmaeili-Fard et al.). Liao et al. analysed the selectively bred group of meat rabbits and used a non-linear mixed model to fit growth curves and measure the effects of SNPs throughout the entire genome. The genetic organization of growth parameters, which is useful for applying genome selection, was also disclosed in this study, which is the first report of GWAS based on single-step NMM for longitudinal traits in rabbits. The logistic model, comprised of 87,704 SNPs in rabbits, was ideally selected and subjected to GWAS using this approach. The two growth indices mature weight (A) and maturity rate (K) were shown to be simultaneously impacted by a total of 45 important SNPs spread across 5 chromosomes. Seven positional genes were proposed as potential candidates influencing meat rabbit growth, including GBA3, KCNIP4, LDB2, PPARGC1A, GNA13, SHISA3 and FGF10 (Liao et al.).

One of the most important economic traits in swine industry which determines its profitability is the meat quality and genetics has an important role in determining such traits. Ardestani et al., conducted a study to evaluate the accuracy of standard BLUP prediction with several prominent genetic assessment techniques such as GBLUP, ssGBLUP, and BayesC for back-fat thickness (BFT), average daily gain (ADG) and loin muscle depth (LMD) parameters in Canadian swine populations. Despite the absence of any significant differences ($p > 0.05$) between the prediction accuracies produced from these genomic approaches in each scenario, ssGBLUP and BayesC techniques often demonstrated the maximum predictive performance and unbiasedness, respectively (Salek Ardestani et al.).

The milk cattle breed known as Sahiwal is indigenous to the Indian subcontinent. They are renowned for their high milk production, exceptional ability to adjust to the hot, humid conditions that prevails in their native territory, and resilience to parasites, ticks, and tropical diseases. This study uses a genotyping INDUS chip to investigate the signature of selection in the genome of Sahiwal cattle. The potential signs of selection in Sahiwal cattle were discovered in 14 important regions of the genome. The most prominent locations were mapped onto BTA 6, 20, and 23, and the p -values from several univariate statistics were combined into a composite signal using the DCMS methodology. The selection signatures, which include important candidate genes linked to features including coat colour, milk fat percentage, sperm membrane integrity, and carcass qualities, are present in BTA 6. There are genes related to bovine *tuberculosis* sensitivity and parasite infestation tolerance located in a significant area on BTA 23 at the 17-Mb region. Illa et al., came to the conclusion that the extremely relevant genetic areas contributed to Sahiwal becoming one of the best milk cow breeds for the tropics (Illa et al.).

Bioinformatic techniques were used to streamline and accelerate the process into a single step in order to identify the causative variants of full penetrance recessive genetic diseases using only nine whole genome sequenced animals. A novel mutation causing prenatal mortality in Irish moiled calves was found utilising whole genome sequencing of only three situations and six carriers. The variant call format (VCF) data file of these 9 animals was examined using four different techniques: autozygosity-by-difference (ABD), genotyping criteria (GCR), variant prediction scoring, and enrolled SNP information. Only one site (Chr4: g.77173487A>T (ARS-UCD1.2 (GCF 002263795.1)) out of around ten million variants in the VCF file was recognised by all these techniques. The glucokinase gene contained a splice acceptor variation at this location (GCK). This study by Pollott et al., showed that only a limited number of cases and controls are needed, and that controls should exceed cases by a ratio of 2:1 and should be less closely related to cases (Pollott et al.). The Pashmina goat genome published in the current edition by Bhat et al., may offers an opportunity to understand the biology of development of this pricy and luxurious fiber (Bhat et al.). This study is one of the first attempts at providing Pashmina genomics and transcriptomic information. Using the Illumina HiSeq 2500 sequencer, a total of 294.8 GB (>100X coverage) of the whole-genome sequence data was generated from a 26 months old male Changthangi Pashmina goat. Moreover, the genomes of a

wild goat and a Pashmina goat were compared, and the results showed a total of 2,823 high impact single nucleotide variants as well as minor insertions and deletions that may be related to the evolution of Pashmina goats. The complex traits of the Pashmina goat, such as annual fibre cycling, defence mechanisms against hypoxia, a method of surviving in extremely cold temperatures, adjustment to a meagre diet, and distinctions of Pashmina fibre from other fibres to avoid marketing practises, can also be understood with the aid of this study.

The majority of the effort in genomic selection has focused on computational efficiency and prediction accuracy, but computing restrictions are becoming an increasingly critical factor that must be considered. A Bayes-based genomic selection model named FMixFN was created in a study by Xu et al. It combines consistent prediction ability and computational effectiveness. The needs of large breeding enterprises or combined breeding programs could be satisfied by the reliable, big data-oriented genomic selection approach known as FMixFN. The FMixFN technique is publicly available at <https://zenodo.org/record/5560913> (Xu et al.). Vaughn et al., used a mixed linear model in fixed effects and random effects to investigate the expression of Actinin-3 (ACTN3) gene, which encodes a muscle-specific structural protein, in correlation to the feeding efficiency phenotype in *Bos taurus* - *Bos indicus* crossbred steers (Vaughn et al.). It was assumed that animals with a higher proportion of fast-twitch muscle fibres are comparatively feed inefficient because the ACTN3 is exclusive to fast-twitch muscle fibres and absent from slow-twitch muscle fibres.

The ovulation rate, a critical reproductive characteristic for goats, determines the top limit of the female's litter size. The rate of ovulation in the ovary is influenced by follicle growth and development. To anticipate how lncRNAs and miRNAs would interact, Tao et al., sequenced and analysed the mRNA expression profiles of pre-ovulatory follicles from goats. Out of the 895 lncRNAs that were found, 88 displayed a clear difference in expression, suggesting important impacts on the ovarian follicles in goats. lncRNA XR_311113.2 might operate as a chi-miR-424-5p sponge. This research demonstrates that lncRNAs may be a valuable new area of investigation in the context of ovarian follicular development (Tao et al.). Yak, a unique reservoir of genetic material that primarily lives on the Qinghai-Tibet Plateau. A clear heterosis in production performance may be shown in cattle-yak, the hybrid offspring of yak (*Bos grunniens*) and cattle (*Bos taurus*). Through RNA sequencing, 7,126 mRNAs, 791 lncRNAs, and 1,057 circRNAs that differ in expression between yaks and

cattle-yaks in the longissimus dorsi muscle were discovered. The ncRNAs were discovered to be associated with the proliferation and differentiation of myoblast cells, skeletal development, and signalling pathway of muscle growth using bioinformatics techniques. Huang et al. claim that this study can be utilised as a benchmark to establish the molecular framework for understanding muscle growth (Huang et al.). The blue egg is biologically interesting as well as economically significant for consumers, egg sellers, and scientists. Studies conducted in the past mostly focused on protein-coding genes to understand the genetic pathways behind pigment deposition. The underlying mechanism by which non-coding RNAs affect the pigment deposition in various eggshell colours is still unknown. Profiling the uterine gland transcriptome (lncRNA and mRNA) of 15 Changshun blue eggshell layers using RNA sequencing. 8 and 22 lncRNA-gene pairings were predicted by combining analyses of lncRNA and mRNA profiles. According to mRNA sequencing, the majority of pathways were mostly focused on lipid-related metabolisms. The found lncRNAs influence immunological and lipid-related metabolisms, as well as pigment disposition, to exert similar effects on colour creation (Chen et al.). The integrated co-expressed transcriptomes, i.e., mRNA and miRNA, were investigated in primary bovine myoblasts following Resveratrol treatment. According to Hao et al., this study improved our knowledge of the functions of RSV in inducing miRNA, the features of DE miRNAs in the important co-expressed component that regulate mRNAs, and it discovered new potential transcription factors and miRNAs for traits associated with meat quality (Hao et al.).

He et al., investigated the link between ACSF3 expression and cellular synthesis of triglycerides (TG) by silencing and over-expression of ACSF3. As a result of the discovery that ACSF3 regulates cytoplasmic triacylglycerol and long-chain PUFA levels, polymorphism may be used as a diagnostic biomarker for forthcoming marker-assisted selection in the production of elevated lipid accumulation traits in beef cattle (He et al.). Heat stress in cattle exhibit a direct impact on rumen health by affecting the processes of fermentation and metabolism of rumen papillae proliferating the rumen papillae. It may adversely affect the physical membrane of the rumen epithelium to a significant degree by increasing corneum loss. HS up-regulated biological processes such as sister chromatid segregation while down-regulating MAPK and NF- κ B cell signalling pathways, according to an examination of the rumen papillae's transcriptome profile. These pathways were linked to DNA replication and repair, amino acid metabolism, and other pathways that were problematic. TLR4 or Tight junction protein signalling expression did not change specifically in response to heat stress, according to Guo et al., indicating that HS had a limited detrimental impact on the ruminal epithelium's physical barrier but did not completely destroy it. The

development of mitigation techniques as well as the productivity of lactation cows is both impacted by the increase in amino acid metabolism in rumen papillae (Guo et al.).

A basic description of the inheritance patterns of alternative splicing in broiler and layer chickens is provided by Qi et al. to best explain post-transcriptional regulation during hybridization. The White Leghorn and Cornish Game chicken breeds, which have markedly different body types and reproductive characteristics, were crossed in an experiment, and the muscle, brain, and liver tissues were then sequenced to determine the inheritance patterns. High tissue and strain specificity is suggested by alternative splicing profiles. The majority of the alternative splicing genes had patterns that were conserved across all three organs, according to a study between parental strains and hybrid crossings. This study gives an overview of the alternative splicing inheritance patterns in these chicken breeds (Qi et al.). Khan et al. developed the first comprehensive buffalo user-friendly web genomic resource (BuffGR). It contained genetic research on five buffalo breeds with major commercial importance: the Mediterranean, Egyptian, Bangladeshi, Jaffarabadi, and Murrah (Khan et al.). The website's database contains information on 4504691 SSR markers across all breeds, 1458 distinct circRNAs, 37712 lncRNAs, and 938 miRNAs from the genomic sequences of the Mediterranean breed of buffalo. This data could be utilised to research the genetic diversity of many buffalo breeds. It is also possible to research post-transcriptional regulation and its function in several bovine diseases. The BuffGR can be found at <http://backlin.cabgrid.res.in/buffgr/>.

Precision when population size is small is a significant barrier to their genomic selection; during the past several years, variable selection approaches with diverse variance have been proposed to increase breeding value accuracy. While these models might be more accurate than conventional and genetic forecasts, they also carry a correspondingly higher breeding value bias and dispersion. Mancin et al., used a number of diverse techniques in his study to increase the precision of genomic selection in a small population. They were chosen using a variety of algorithms, including XGBoost, penalised regression, and recursive feature eliminations (Mancin et al.). Variable selection ssGBLUP, especially XGBoost, has prediction accuracy that is higher than other ssGBLUP methods without the exaggerated bias and dispersion that come with weighted ssGBLUP. In this study, machine learning algorithms are used, which may provide a solution to the problem of genomic selection in small populations, such as the local cow population. The techniques described in his study may help preserve indigenous cow breeds, study them, and boost their economic competitiveness. For carcass evaluation in meat industry, rib eye area is an important index used. Since this indicator is inherited and possesses genetic diversity, it can be used to improve the

genetic makeup of sheep. This study conducted KASPar genotyping on five SNPs, demonstrating that the rib eye region is highly linked with SNPs in LOC105611989, DPP6, and COL12A1. These SNPs could be used as genetic markers for rib eye area molecular breeding. The findings published by Zhao et al., offer genetic factors evaluated on the rib eye region and guidance for breeding Hu sheep based on carcass features (Zhao et al.).

Author contributions

SA initiated the Research Topic. MD, BB, AD, and SP co-edited the Research Topic with SA and all authors participated in the editorial process. SA drafted the manuscript. All authors revised and approved the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



Genome-Wide Association Studies Reveal Susceptibility Loci for Noninfectious Claw Lesions in Holstein Dairy Cattle

Ellen Lai, Alexa L. Danner, Thomas R. Famula and Anita M. Oberbauer*

Animal Science Department, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

Yachun Wang,
China Agricultural University, China
Christine Francoise Baes,
University of Guelph, Canada

*Correspondence:

Anita M. Oberbauer
amoberbauer@ucdavis.edu

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 22 January 2021

Accepted: 15 April 2021

Published: 28 May 2021

Citation:

Lai E, Danner AL, Famula TR and
Oberbauer AM (2021) Genome-Wide
Association Studies Reveal
Susceptibility Loci for Noninfectious
Claw Lesions in Holstein Dairy Cattle.
Front. Genet. 12:657375.
doi: 10.3389/fgene.2021.657375

Sole ulcers (SUs) and white line disease (WLD) are two common noninfectious claw lesions (NICL) that arise due to a compromised horn production and are frequent causes of lameness in dairy cattle, imposing welfare and profitability concerns. Low to moderate heritability estimates of SU and WLD susceptibility indicate that genetic selection could reduce their prevalence. To identify the susceptibility loci for SU, WLD, SU and/or WLD, and any type of noninfectious claw lesion, genome-wide association studies (GWAS) were performed using generalized linear mixed model (GLMM) regression, chunk-based association testing (CBAT), and a random forest (RF) approach. Cows from five commercial dairies in California were classified as controls having no lameness records and ≥ 6 years old ($n = 102$) or cases having SU ($n = 152$), WLD ($n = 117$), SU and/or WLD (SU + WLD, $n = 198$), or any type of noninfectious claw lesion ($n = 217$). The top single nucleotide polymorphisms (SNPs) were defined as those passing the Bonferroni-corrected suggestive and significance thresholds in the GLMM analysis or those that a validated RF model considered important. Effects of the top SNPs were quantified using Bayesian estimation. Linkage disequilibrium (LD) blocks defined by the top SNPs were explored for candidate genes and previously identified, functionally relevant quantitative trait loci. The GLMM and CBAT approaches revealed the same regions of association on BTA8 for SU and BTA13 common to WLD, SU + WLD, and NICL. These SNPs had effects significantly different from zero, and the LD blocks they defined explained a significant amount of phenotypic variance for each dataset (6.1–8.1%, $p < 0.05$), indicating the small but notable contribution of these regions to susceptibility. These regions contained candidate genes involved in wound healing, skin lesions, bone growth and mineralization, adipose tissue, and keratinization. The LD block defined by the most significant SNP on BTA8 for SU included a SNP previously associated with SU. The RF models were overfitted, indicating that the SNP effects were very small, thereby preventing meaningful interpretation of SNPs and any downstream analyses. These findings suggested that variants associated with various physiological systems may contribute to susceptibility for NICL, demonstrating the complexity of genetic predisposition.

Keywords: sole ulcer, pododermatitis circumscripta, white line disease, lameness, genome-wide association study, random forest, Bayesian regression, dairy cattle

INTRODUCTION

Lameness, or abnormal gait and/or posture, is a pathognomonic sign that the affected cow is in pain and frequently reflects claw damage. Many claw conditions can cause lameness, including injury, infectious claw lesions, and noninfectious claw lesions. The two most common noninfectious claw lesions causing lameness in dairy cattle are sole ulcers (SUs), also known as pododermatitis circumscripta, and white line disease (WLD) (Green et al., 2002; Shearer and van Amstel, 2017). These lesions are not only a welfare issue but are also associated with reduced milk production and decreased fertility (Green et al., 2002, 2010; Hernandez et al., 2005; Charfeddine and Pérez-Cabal, 2017). Consequently, SU and WLD represent a considerable financial burden, with the average costs associated with prevention, treatment, and losses from reduced productivity ranging from \$181 (Dolecheck et al., 2019) to \$258 (Cha et al., 2010) per case of SU and \$155 for WLD (Dolecheck et al., 2019) (adjusted to 2020 US dollars). Production losses from extended calving interval, increased culling, and decreased milk production increase greenhouse gas emissions by 33 (3.6%) and 39 (4.3%) kg CO₂ equivalents per ton of fat- and protein-corrected milk per case of SU and WLD, respectively (Mostert et al., 2018). Reducing the prevalence of SU and WLD would alleviate these welfare, economic, and environmental concerns and thereby improve the sustainability of dairy production.

Both genetic and non-genetic factors contribute to susceptibility to SU and WLD, and prevention can be achieved through genetic means and herd management. Current prevention methods focus on management control primarily through regular claw trimming (Shearer and van Amstel, 2001) and providing rubber flooring in stalls and alleys (Vanegas et al., 2006; Fjeldaas et al., 2011; Eicher et al., 2013). Although dairies have implemented these prevention methods, SU and WLD remain prevalent worldwide, with estimates ranging from 4.1 to 27.8% for SU and from 2.0 to 11% for WLD in Holstein cattle depending on parity and the housing style (Cramer et al., 2008; Bicalho et al., 2009; van der Linde et al., 2010; Oberbauer et al., 2013). Heritability estimates of susceptibility range from 0.01 to 0.3 for SU and from 0.017 to 0.26 for WLD (Van der Waaij et al., 2005; van der Linde et al., 2010; Häggman and Juga, 2013; Oberbauer et al., 2013; van der Spek et al., 2013, 2015a; Malchiodi et al., 2015a), implying that these non-genetic means to reduce prevalence could be bolstered by genetic selection against susceptibility to these claw lesions. Although many genome-wide association studies (GWAS) have been performed to identify the susceptibility loci, loci previously associated with SU and WLD are discordant (Malchiodi et al., 2015b; van der Spek et al., 2015b; Sánchez-Molano et al., 2019), and susceptibility to these claw lesions is believed to be a complex trait governed by loci of small effect (van der Spek et al., 2015b). Some have postulated that selection against susceptibility to SU,

WLD, and other noninfectious claw lesions could be achieved through indirect selection on body conformation traits or feet and leg traits (Van der Waaij et al., 2005; Häggman et al., 2013). However, the genetic correlation between the conformation traits and susceptibility to noninfectious claw lesions appears to be low (Häggman and Juga, 2013; Malchiodi et al., 2015b; Ring et al., 2018), further accentuating the need to identify loci associated directly with susceptibility to noninfectious claw lesions. Thus, the objective of this study was to identify the genomic regions associated with susceptibility to SU, WLD, SU and/or WLD, and noninfectious claw lesions using well-characterized herds under similar management practices: we hypothesized that we would identify small-effect loci associated with predisposition to noninfectious claw lesions in addition to those already identified.

MATERIALS AND METHODS

All procedures were conducted in accordance with the ethical standards set by the University of California, Davis, and approved by the Institutional Animal Care and Use Committee (protocol no. 22099).

Phenotypic Data

Dairies were selected to minimize environmental variations by including dairies in Central and Northern California using freestall housing, a flush system for waste removal, and diets balanced to meet the nutrition requirements from the National Research Council (National Research Council (NRC), 2001). Case/control phenotypes were defined using hoof trimming records. The hoof trimming records were generated by three hoof trimmers: one serviced dairies A, B, and C; one serviced dairy D; and the last trimmer serviced dairy E. Hoof trimmer qualifications were described in a previous paper (Lai et al., 2020), and the three trimmers employed common criteria in defining the lesions. Hoof trimming regimens varied among dairies: cows were trimmed at the beginning of and at mid-lactation, at dry off, and when lame (dairy A); at dry off and when lame (dairies B and C); only when lame (dairy D); and at mid-lactation, at dry off, and when lame (dairy E). The following claw lesions were documented in the hoof trimming records: SU, hemorrhage, sole fracture, sole abscess, wall abscess, white line abscess (WLD), heel abscess, laminitis, foot wart, and foot rot. Cows were phenotyped as cases or controls based on whether they had or lacked records of claw lesions, respectively. Four case/control datasets were generated based on the type(s) of claw lesions the cases had. For datasets 1 (SU) and 2 (WLD), cases were defined as cows with at least one record of SU or WLD, respectively. For dataset 3 (SU + WLD), cases included cows with either one or both of the claw lesions. Cases for dataset 4 (noninfectious claw lesions, NICL) included cows with at least one of the following noninfectious claw lesions: SU, hemorrhage, sole fracture, sole abscess, wall abscess, WLD, heel abscess, and/or laminitis. Cows with no claw lesions and that were at least 6.0 years old were considered sound controls. The age restriction was imposed to avoid misphenotyping younger cows that had insufficient time

Abbreviations: BTA, *Bos taurus* autosome; CBAT, chunk-based association testing; GLMM, generalized linear mixed model; GRM, genetic relatedness matrix; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAF, minor allele frequency; NICL, noninfectious claw lesions; PVE, proportion of phenotypic variance explained; RF, random forest; SUs, sole ulcers; WLD, white line disease.

to develop claw lesions. The same sound controls were used to compare against the cases in each of the four datasets.

Genotypes

Whole blood was collected from cows phenotyped as cases and controls. DNA was extracted from whole blood samples using the QIAGEN QIAamp DNA Blood Mini Kit (QIAGEN Inc., Valencia, CA) and quantified using the NanoDrop (ND-2000 v3.2.1) spectrophotometer (Thermo Scientific, Wilmington, DE, United States). DNA samples were genotyped on the BovineHD BeadChip [777K single nucleotide polymorphisms (SNPs), Illumina Inc., San Diego, CA, United States] by GeneSeek (Lincoln, NE, United States), and Illumina's GenCall algorithm was used to call genotypes. A portion of the controls used in this study were the same controls used in our previous study (Lai et al., 2020), for which raw and processed genotype data are publicly available at the NCBI Gene Expression Omnibus database (GEO series record GSE159157). Additional cows genotyped in this study are available in the GEO database (GEO series record GSE165945).

Genotypes were updated to the ARS-UCD1.2 assembly positions (Rosen et al., 2020) and quality filtered using PLINK 1.9 (Chang et al., 2015; Purcell and Chang, 2015) to remove from further analyses SNPs and cows with genotyping rates <95%, SNPs with significant deviation from Hardy–Weinberg equilibrium ($p < 1E-6$) to exclude systematic genotyping errors, and SNPs with minor allele frequencies (MAFs) < 5% to exclude rare variants. To visualize genetic similarity among the remaining cows, multidimensional scaling (MDS) analysis was performed, and the first two dimensions were plotted. Because the downstream programs for GWAS analysis [the generalized linear mixed model (GLMM) and random forest (RF)] required genotypes at each SNP, missing genotypes remaining after quality filtering were imputed using BEAGLE 5.1 (Browning et al., 2018) using the default parameters and an effective population size of 58 previously estimated for North American Holstein cattle (Makanjuola et al., 2020).

Generalized Linear Mixed Model GWAS

Because disease phenotype was binary (cases and controls), the model used for association testing needed to reflect this binary outcome. Accordingly, logistic regression was used to model the binary outcome for the power analysis and for association testing. Power analysis was conducted using the *genpwr* R package (Moore et al., 2019), assuming an additive genetic effect and a sample size and case rate similar to the sample population (sample size = 275, case rate = 0.6). Given these parameters, the smallest effect SNP that the GWAS was expected to detect would have an odds ratio of at least 1.7 and a MAF of at least 0.34. For association testing, a genetic relatedness matrix (GRM) and farms were included as covariates in the model to account for population stratification and relatedness as well as the effect of farm, respectively. The probability of disease was defined as p_{ijk} for the k -th cow on the i -th farm identified in the j -th SNP genotype class and the logit of this probability, as $\theta_{ijk} = \log[p_{ijk}/(1 - p_{ijk})]$. The logit of the probability of disease was

modeled as a function of the recorded explanatory variables (e.g., farm and SNP genotype) along with a presumed quantitative genetic contribution for each SNP:

$$\theta_{ijk} = \mu + F_i + S_j + a_k$$

where μ is an unknown constant common to all cows, F_i the contribution of i -th farm to the risk of disease, and S_j is the contribution of the j -th SNP genotype to the risk of disease. The additive genetic effect a_k is assumed to be drawn from the multivariate normal density $N(0, A\sigma_a^2)$, with A as the standardized GRM among the animals in the dataset calculated in GEMMA (Zhou and Stephens, 2012) and σ_a^2 is the unknown additive genetic variance of the disease risk. Model fitting and association testing via the score test (i.e., the Leverage multiplier test) were implemented with the generalized linear mixed model association test (GMMAT) R package (Chen et al., 2016).

The effective number of independent markers (M_e) was calculated as the number of SNPs remaining after linkage disequilibrium (LD) pruning using the Genetic Type I error calculator and used as the denominator for Bonferroni correction of the association p values (Li et al., 2012). Significant SNPs were defined as those with $p \leq 0.05/M_e$ and suggestive SNPs were defined as those with $p \leq 1/M_e$ (Lander and Kruglyak, 1995). Genomic inflation factors were calculated as the ratio of the median of the observed and expected p values. Quantile–quantile plots (qqplots) and Manhattan plots were plotted using the R package *qqman* (R Development Core Team, 2010; Turner, 2014).

Chunk-Based Association Testing

Chunk-based association testing (CBAT), also called set-based association testing, was performed to decrease multiple testing and, in turn, improve the power of detecting associated regions in the small sample size. In contrast to gene-based association testing, which jointly tests variants within genes for association with the phenotype (e.g., Xia et al., 2017), CBAT analyzes consecutive windows of variants (i.e., chunks) across each chromosome without prior filtering. Accordingly, CBAT includes variants in non-coding regions containing regulatory elements that could contribute to phenotypic variation in complex traits (Koufariotis et al., 2014, 2018). Quality-filtered SNPs were split into 100-kb chunks overlapping by 50 kb. Each chunk was LD-pruned to remove SNPs that were in strong LD ($R^2 > 0.98$) and then tested for association with the phenotype by determining whether the phenotypic variance explained (PVE) by the chunk was significantly greater than zero. Specifically, association testing for each chunk was performed by calculating a GRM using the SNPs in the chunk and regressing the phenotype on the GRM. In addition to the chunk-based GRM, a thinned GRM (from genome-wide SNPs) and farms were included as covariates in the model to adjust for population stratification and differences among farms. The thinned GRM was calculated using genome-wide LD-pruned SNPs: SNPs within a window of 1 Mb and a $R^2 > 0.5$ were pruned out such that only SNPs in linkage equilibrium were used in the GRM calculation. For each chunk of SNPs, the following linear model was used to define the

disease phenotype y for the k -th cow as a function of phenotypic contribution from the j -th chunk that comprised m SNPs and the i -th farm:

$$y_{ijk} = \mu + F_i + C_j + a_k + \varepsilon_{ijk}$$

where μ , F_i , and a_k are the same components outlined in the previous equation contributing to phenotype (coded as 0 for controls and 1 for cases), $C_j = \sum_{l=1}^m S_l$ is the contribution of the chunk to the phenotype in which S_l is the contribution of the l -th SNP in the chunk, and ε_{ijk} is the residual term. Estimates of PVE for each chunk were transformed to the underlying liability scale to adjust for ascertainment of cases using prevalence estimates from the literature: 4.08% for SU, 7.89% for WLD, 0.10 for SU + WLD, and 0.10 for NICL (DeFraain et al., 2013; Oberbauer et al., 2013). Calculating the thinned GRM, estimating PVE by each chunk, association testing with the likelihood ratio test, and p value estimation via 10 permutations for each chunk (Listgarten et al., 2013) were performed using the linkage disequilibrium-adjusted kinships (LDAK) program (Speed et al., 2012). For each dataset, the significance thresholds were adjusted using Bonferroni correction: chunks with $p \leq 0.05/(\text{number of chunks})$ were defined as significant and chunks with $p \leq 1/(\text{number of chunks})$ were defined as suggestive (Lander and Kruglyak, 1995). Manhattan plots and qqplots were plotted using the R package *qqman* (R Development Core Team, 2010; Turner, 2014).

Random Forest GWAS

A RF fits a model that includes all SNPs and does not require an assumption about the mode of inheritance (e.g., additive, dominant, and recessive), making RFs an appealing approach for complex traits such as susceptibility to claw lesions, in which the trait is highly polygenic and epistasis is present (Goldstein et al., 2010). Furthermore, RFs are insensitive to uneven sampling of cases and controls across different dairies, as RFs first build decision trees, then quantify the importance values afterward with data available in the trees.

Linkage disequilibrium pruning and RF analyses were performed as previously detailed (Lai et al., 2020) for each of the four datasets. Briefly, LD-pruned genotypes and farms were used as predictors for the RF analyses performed using the *caret* R package (Kuhn, 2008; R Development Core Team, 2010). For each dataset, the population was randomly divided into a training (two-thirds of the cows) and a test (one-third of the cows) population. Using the training population, the number of predictors considered at each node of each decision tree, *mtry*, was tuned using five values, $0.1p$, $0.2p$, $0.5p$, $0.8p$, and p , where p is the total number of predictors (Goldstein et al., 2010; Briec et al., 2018). The *mtry* resulting in the most accurate RF model was used for downstream analyses. The most important predictor was assigned a value of 100, and any other predictor's importance values was scaled accordingly (e.g., a predictor with an importance value of 50 is 50% as important as the most important predictor). Model validation was performed by using the predictors and their importance values to predict the case/control phenotype in the test population. To determine which SNPs were important

and worthy of further investigation, a scree plot was plotted and the second-order point of inflection was identified using the inflection R package (Christopoulos, 2016, 2017) (i.e., the "elbow method"). Predictors with importance values equal to or greater than the second-order point of inflection were defined as important SNPs and explored in downstream analyses if and only if the RF model was significantly more accurate at predicting phenotype in the test population than the non-information rate (i.e., the frequency of the more common phenotype).

Defining Associated Regions

For each of the four datasets, the top SNPs were defined as significant and suggestive SNPs from the GLMM regression or important SNPs from a significantly predictive RF model. Boundaries of the genomic regions of association were defined using SNPs in LD with top SNPs. Similar to the methodology of Richardson et al. (2016) and Twomey et al. (2019), the positions of SNPs within 5 Mb and with $R^2 \geq 0.5$ of each top SNP were determined using non-pruned imputed genotypes, and the furthest SNP upstream and downstream in LD with the significant or suggestive SNP defined the LD block boundaries. Overlapping LD blocks were combined. Using the same procedure outlined for CBAT, the PVE by the LD blocks defined from the GLMM and RF analyses was estimated and compared against the PVE by chunks of SNPs of the same size that overlapped by 50 kb from all chromosomes.

Bayesian Estimation of SNP Effects and Assessing Model Fit

A Bayesian approach was used to test the association of the top SNPs identified in the GLMM and the RF with the case/control phenotype for the four datasets. Bayesian methodology was selected because it allows multiple SNPs to be fitted jointly, recognizes that some SNPs are correlated and most likely have small effects on susceptibility (van der Spek et al., 2015b), and can account for the uneven sampling of cases and controls from dairies. Additionally, the effect size estimates obtained from Bayesian estimation are directly interpretable, and Bayesian model evaluation is extremely thorough. Because highly correlated predictors complicate Bayesian regression, the significant and suggestive SNPs detected in the GLMM GWASs were LD-pruned ($R^2 > 0.9$) using PLINK 1.9 (Chang et al., 2015; Purcell and Chang, 2015) prior to estimating effects to keep the most significant SNP in each LD block for inclusion into the Bayesian model. Estimation of SNP effects was performed using a Bayesian logistic regression model as described in Lai et al. (2020). The important SNPs from the RF did not need to be LD-pruned, as SNPs were LD-pruned prior to RF analyses. Briefly, each set of top SNPs (i.e., LD-pruned suggestive/significant SNPs from the GLMM analyses and important SNPs from RF analyses) was used as predictors along with farm as a covariable in a Bayesian logistic regression model, and the model was fitted via sampling the posterior using the Hamiltonian Monte Carlo algorithm in the R package *rstanarm* (Gelman et al., 2020; Goodrich et al., 2020). The same population was used in the GLMM and RF GWAS as for the SNP effect estimation, which

could lead to the inclusion of false-positive associations in the Bayesian model. Thus, to discern whether the included SNPs were false positives, the fit of the Bayesian model using the estimated parameters was evaluated using leave-one-out (LOO) cross-validation and posterior predictive checking (PPC) using the *loo* and *bayesplot* R packages (Vehtari et al., 2017, 2020; Gabry et al., 2019). Bayesian estimation of the SNP effects generated a distribution of where the true value of the SNP effect was, and this range was quantified in the 95% uncertainty intervals (UI), as opposed to a point estimate in frequentist methods. SNPs with 95% UIs that did not overlap zero were considered significantly associated with susceptibility to the respective claw lesion(s).

Functional Annotation of Associated Regions

Genes and previously defined quantitative trait loci (QTL) falling within or overlapping with the associated LD blocks and chunks were obtained using FAANGMine using the genomic regions search function (Functional Annotation of Animal Genomes (FAANG), 2019) and the CattleQTLdb (Hu et al., 2019). RefSeq genes were extracted from the resulting gene list and used in the pathway and gene ontology enrichment analysis in FAANGMine. Genes were searched in the Mouse Genome Informatics batch query database to find the associated mammalian phenotypes (Smith and Eppig, 2009). Genes were also queried in the Cattle Gene Atlas (Fang et al., 2020) to determine in which tissues they were expressed.

RESULTS

Descriptive Data

The percentage and count of cows with records of each claw lesion from each dairy are presented in **Table 1**. Of the cows that had hoof trimming records from the five dairies, 5.6 and 12.0% had records of SU and WLD, respectively, similar to previous prevalence estimates (Cramer et al., 2008; Bicalho et al., 2009; van der Linde et al., 2010; Oberbauer et al., 2013). For cows that were genotyped, cases were sampled from all five dairies, whereas controls were sampled from dairies A and D, which had cows that met our strict soundness and age criteria for controls. The dataset included 156 SU cases, 119 WLD cases, 203 SU + WLD cases (72 cows had both SU and WLD), 222 NICL cases, and 104 sound controls, for a total of 287 cows (**Table 2**). The average age of the controls sampled was 8.7 years old (SD = 1.4), and when compared to the average age of onset of 4.2 (SD = 1.7) for SU and 4.5 (SD = 2.6) years for WLD, it indicated that our age cutoff of 6.0 years old was sufficient to avoid misphenotyping control cows.

After quality filtering, ~556,000 SNPs for 152 SU cases, 117 WLD cases, 198 SU + WLD cases (71 cases had both SU and WLD), 217 NICL cases, and 102 sound controls remained for MDS, GLMM, Genetic Type I error calculation, CBAT, and RF analyses. The MDS plot showed some population stratification, with a prominent center cluster and two other sparse clusters, although clustering was not by farm or case/control phenotype (**Supplementary Figure 1**). Pairwise relationship coefficients calculated for the GRM ranged from −0.094 to 0.50, with

TABLE 1 | Percent (and number) of cows with records of foot lesions across the five dairies.

Farm	Percentage of cows with lesion (n)										Total no. of cows with records
	Sole ulcer	White line disease	Foot wart	Wall abscess	Sole abscess	Sole fracture	Foot rot	Bruise	Heel abscess	Severe laminitis	
A	6.9 (467)	15.6 (1050)	5.6 (376)	2.2 (146)	4.7 (319)	1.9 (125)	0.9 (62)	NR	NR	NR	6,734
B	1.2 (44)	2.6 (91)	8.5 (301)	0.2 (6)	0.6 (23)	1 (37)	0 (1)	NR	NR	NR	3,549
C	1.3 (35)	2.1 (57)	8.8 (236)	0.5 (13)	0.3 (9)	0.4 (10)	0 (1)	NR	NR	NR	2,676
D	5.5 (254)	12.8 (596)	5.8 (268)	NR	NR	NR	0.9 (42)	1.6 (73)	NR	NR	4,658
E	11.1 (380)	21.4 (733)	17.9 (614)	NR	NR	NR	1.1 (39)	NR	16.3 (559)	8.3 (284)	3,427
Total ^a	5.6 (1180)	12.0 (2527)	8.5 (1795)	1.3 (165)	2.7 (351)	1.3 (172)	0.7 (145)	1.6 (73)	16.3 (559)	8.3 (284)	21,044

NR, not recorded.

^aTotals were calculated across dairies that had records of the lesion; dairies that did not record the lesion were excluded from the calculation of totals.

negative values indicating that the two cows were less related to each other than other random pairs of individuals. The distribution of the pairwise relationship coefficients did not differ greatly between pairs of cows from the same farm and pairs from different farms (Supplementary Figure 2). The Genetic Type I error calculator determined that the effective number of markers on autosomal chromosomes for Bonferroni correction was $\sim 156,000$ SNPs for the four datasets, yielding a significance threshold of $p = 3.2 \times 10^{-7}$ [6.5 on $-\log_{10}(p)$ scale] and a suggestive threshold of $p = 6.4 \times 10^{-6}$ [5.2 on $-\log_{10}(p)$ scale]. The total number of 100-kb chunks used in CBAT was $\sim 51,730$ for the four datasets, yielding a significance threshold of $p = 9.7 \times 10^{-7}$ [6.0 on $-\log_{10}(p)$ scale] and a suggestive threshold of $p = 1.9 \times 10^{-5}$ [4.7 on $-\log_{10}(p)$ scale]. Linkage disequilibrium pruning at $R^2 > 0.90$ left 215,343–218,185 SNPs for RF analysis, depending on the dataset.

Generalized Linear Mixed Model GWAS and CBAT

The GLMM analyses detected a region of association on BTA8 for SU and BTA13 for WLD, SU + WLD, and NICL while sufficiently accounting for population stratification and relatedness, as indicated by the qqplots and the genomic inflation factors of 1.01, 1.02, 1.01, and 0.99 for SU, WLD, SU + WLD, and NICL, respectively (Supplementary Figure 3). The CBAT using 100-kb overlapping chunks across the genome also properly accounted for population stratification and relatedness (qqplots in Supplementary Figure 5) and identified the same regions as the single-marker GLMM GWAS for each of the four datasets, providing further support for these regions (Supplementary Table 1; Manhattan plots in Supplementary Figure 6). The SU CBAT also identified two suggestive chunks on BTA17 (Supplementary Table 1 and Supplementary Figure 6A). For the NICL CBAT, the reduction in the number of tests performed allowed the chunk at BTA13:46,450,001–46,550,001 to reach genome-wide significance ($p = 6.9 \times 10^{-7}$; Supplementary Table 1 and Supplementary Figure 6D). This significant chunk contained the most significant SNP from the single-marker GLMM GWAS and three suggestive SNPs downstream.

The GLMM association testing for SU susceptibility identified 12 suggestive SNPs on BTA8 falling in or directly upstream of the gene *DCAF12* (also known as *DDB1* and *CUL4*-associated

factor 12) (Tables 3, 4). The 12 suggestive SNPs collectively defined a 3.2-Mb LD block at BTA8:74,345,807–77,546,693 (Table 3 and Figure 1A) encompassing or overlapping with 60 genes: 52 protein-coding genes, four long non-coding RNA (lncRNA) genes, a transfer RNA (tRNA) gene, a microRNA (miRNA) gene, a small nuclear RNA (snRNA) gene, and a small nucleolar RNA (snoRNA) gene. Because the 12 suggestive SNPs from the SU GLMM were in strong LD ($R^2 > 0.9$), the most significant SNP, BovineHD0800023021, was selected to represent this LD block in the Bayesian logistic regression model. The minor allele at BovineHD0800023021 (T) had an effect that was significantly less than zero at 95% UI (Table 3 and Figure 2A), indicating that it was associated with reduced susceptibility to SU. The LOO analysis yielded acceptable Pareto k values ($k < 0.5$) for all cows, which indicated that the model was able to predict the phenotype of each cow with similar accuracy using the genotypes at BovineHD0800023021 from all other cows. Goodness-of-fit assessment via PPC also showed that the distribution of the phenotypes simulated using the estimated SNP effect closely aligned with that of the observed data (Supplementary Figure 7A), further validating the fit of the model. In addition to identifying suggestive chunks in the same regions on BTA8, CBAT for SU detected two significant chunks on BTA17 (Supplementary Table 1 and Supplementary Figure 6A) that both fell within *TMEM12* (transmembrane protein 132B).

For WLD, the GLMM association testing found a single suggestive intergenic SNP at BTA13:46,491,619 (BovineHD1300013725; Supplementary Figure 4A), which was also the most significant SNP identified by the GLMM analyses for SU + WLD and NICL (Table 3 and Supplementary Figure 4B and Figure 1B). In addition to detecting BovineHD1300013725, the GLMM analyses for the SU + WLD and NICL datasets detected eight other suggestive SNPs in the same LD block as BovineHD1300013725 (Table 3 and Supplementary Figure 4). These nine suggestive SNPs detected in the SU + WLD GWAS were slightly more significant in the NICL GWAS and defined a 2.4-Mb LD block at BTA13:45,283,136–47,676,681 containing 27 genes: 16 protein-coding genes, six lncRNA genes, two snRNA genes, two snoRNA genes, and one miRNA gene. For all four GLMM GWAS, the limited number of genes in the LD blocks defined from suggestive SNPs precluded pathway and gene ontology analyses.

Given that the GLMM GWAS for SU + WLD and NICL identified nine suggestive SNPs in the same LD block ($R^2 > 0.5$) on BTA13 (Figure 1B and Supplementary Figure 4) and the top SNP is the same as that in the WLD GWAS, only the NICL Bayesian SNP effect estimation results are presented. Eight of these suggestive SNPs were in strong LD ($R^2 > 0.9$), whereas the remaining suggestive SNP (BTB-00525539) was in weaker LD with the others ($R^2 = 0.7$). Consequently, the most significant SNP in the LD block of eight SNPs (BovineHD1300013725) and BTB-00525539 were included in the Bayesian logistic regression model. The minor allele at BovineHD1300013725 representing the eight SNPs in strong LD had an effect that was significantly greater than zero at 95% UI (Figure 2B), indicating that the minor allele (C) was associated with increased susceptibility to

TABLE 2 | Distribution of cases for sole ulcers (SU), white line disease (WLD), SU + WLD, noninfectious claw lesions (NICL), and sound controls after quality filtering across the five dairies.

Farm	Controls	Cases			
		SU	WLD	SU + WLD	NICL
A	81	44	48	75	87
B	0	8	13	17	23
C	0	4	7	9	10
D	21	71	33	72	72
E	0	25	16	25	25
Total	102	152	117	198	217

TABLE 3 | SNPs that were suggestive in the generalized linear mixed model association analysis and the linkage disequilibrium (LD) blocks they defined for sole ulcers (SU), white line disease (WLD), sole ulcers and/or white line disease (SU + WLD), and noninfectious claw lesions (NICL).

Dataset	BTA	SNP	SNP position (bp)	Minor/ major allele	Minor allele count		Minor allele frequency		Score ^a (variance)	p	SNP significance in Bayesian estimation ^b	LD block start (bp)	LD block end (bp)	LD block length (kb)
					Cases	Controls	Cases	Controls						
SU	8	BovineHD0800023014	75,489,164	T/C	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023015	75,490,011	T/G	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	ARS-BFGL-NGS-112795	75,490,692	A/G	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023016	75,491,531	C/T	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023017	75,492,307	G/A	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023018	75,493,464	T/C	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023019	75,494,163	C/T	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023021	75,496,244	T/C	77	97	0.253	0.476	−20.4 (18.8)	2.66E−06	*	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023022	75,496,918	A/G	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023023	75,497,471	C/T	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
WLD	8	BovineHD0800023024	75,498,118	A/G	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	8	BovineHD0800023025	75,501,482	T/C	75	94	0.247	0.461	−20 (18.1)	2.71E−06	−	74,345,807	77,546,693	3,200.9
	13	BovineHD1300013725	46,491,619	C/T	106	48	0.453	0.235	19.9 (19.4)	6.13E−06	*	46,307,416	47,584,595	1,277.2
	13	BovineHD1300013725	46,491,619	C/T	183	48	0.462	0.235	25 (25.5)	7.03E−07	*	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013733	46,526,509	C/T	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
SU + WLD	13	BovineHD1300013739	46,540,186	G/T	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013740	46,541,925	C/T	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013750	46,561,964	C/T	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013759	46,582,769	G/A	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013765	46,596,264	A/G	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013774	46,637,235	A/G	188	52	0.475	0.255	24.8 (25.6)	9.86E−07	−	45,283,136	47,676,681	2,393.5
	13	BTB-00525539	47,420,271	C/A	195	59	0.492	0.289	24.6 (27.8)	3.03E−06	ns	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013725	46,491,619	C/T	199	48	0.459	0.235	26.4 (27.2)	3.96E−07	*	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013733	46,526,509	C/T	204	52	0.470	0.255	26 (27.3)	6.68E−07	−	45,283,136	47,676,681	2,393.5
	13	BovineHD1300013739	46,540,186	G/T	204	52	0.470	0.255	26 (27.3)	6.68E−07	−	45,283,136	47,676,681	2,393.5

(Continued)

TABLE 3 | Continued

Dataset	BTA	SNP	SNP position (bp)	Minor/major allele	Minor allele count		Minor allele frequency	Score ^a (variance)	p	SNP significance in Bayesian estimation ^b	LD block start (bp)	LD block end (bp)	LD block length (kb)
					Cases	Controls							
13	BovineHD1300013740	46,541,925	C/T	204	52	0.470	0.255	26 (27.3)	6.68E-07	-	45,283,136	47,676,681	2,393.5
13	BovineHD1300013750	46,561,964	C/T	204	52	0.470	0.255	26 (27.3)	6.68E-07	-	45,283,136	47,676,681	2,393.5
13	BovineHD1300013759	46,582,769	G/A	204	52	0.470	0.255	26 (27.3)	6.68E-07	-	45,283,136	47,676,681	2,393.5
13	BovineHD1300013765	46,596,264	A/G	204	52	0.470	0.255	26 (27.3)	6.68E-07	-	45,283,136	47,676,681	2,393.5
13	BovineHD1300013774	46,637,235	A/G	204	52	0.470	0.255	26 (27.3)	6.68E-07	-	45,283,136	47,676,681	2,393.5
13	BTB-00525539	47,420,271	C/A	213	59	0.491	0.289	25.8 (29.3)	1.79E-06	ns	45,283,136	47,676,681	2,393.5

^aScores are for the minor allele generated from the generalized linear mixed model analysis. Negative scores indicate that the minor allele is associated with reduced susceptibility and positive scores indicate that the minor allele is associated with increased susceptibility.

^bSNPs used in the Bayesian model were either significantly different from zero at 95% uncertainty interval (*) or not significant (ns). Other SNPs were in LD with SNPs that were used in the model and were excluded from the model (-).

NICL (Table 3). In contrast, the effect of the minor allele at BTB-00525539 was not significantly different from zero (Figure 2B). Although the score variances of the suggestive SNPs were large (Table 3), possibly due to the sample cohort, Bayesian estimation was less affected than GLMM regression by these limitations and indicated that the SNP effects were significant for SU and NICL (Figure 2). For the LOO analysis of the model, the acceptable Pareto k values ($k < 0.5$) from all cows demonstrated that the model including BovineHD1300013725 and BTB-00525539 was able to predict the NICL phenotype of each cow based on the genotypes at these two SNPs from the other cows with similar accuracy. The PPC-simulated data based on the estimated SNP effects of these two SNPs were similar to the observed data, indicating good model fit (Supplementary Figure 7B).

To draw attention to the impactful SNPs shown in Table 3 and the LD blocks they defined in Table 4, the minor allele frequencies at the most significant SNP for SU (BovineHD0800023021) in cases and controls were 0.253 and 0.476, respectively. The GLMM output score was negative and Bayesian estimation indicated a significant negative effect on susceptibility; that is, the minor allele was associated with reduced susceptibility. In contrast, the MAF at the most significant SNP for NICL (BovineHD1300013725) was higher in cases (0.459) than in controls (0.235), indicating that the minor allele was associated with higher susceptibility. Likewise, the GLMM score was positive, and Bayesian estimation of the effect size resulted in a significant positive effect. Similar minor allele frequencies, scores, and significantly positive effect size estimates were observed at BovineHD1300013725 for WLD and SU + WLD. As seen in Table 4, the LD blocks defined by the suggestive SNPs had PVE between 0.06 and 0.08, depending on the dataset (SU, WLD, SU + WLD, or NICL), all of which were significantly greater than zero (permuted $p < 0.05$). In contrast, the genome-wide chunks with the same length as the LD blocks had an average PVE ~ 0.008 , with PVE increasingly slightly with increasing chunk size, and average permuted p values ~ 0.5 .

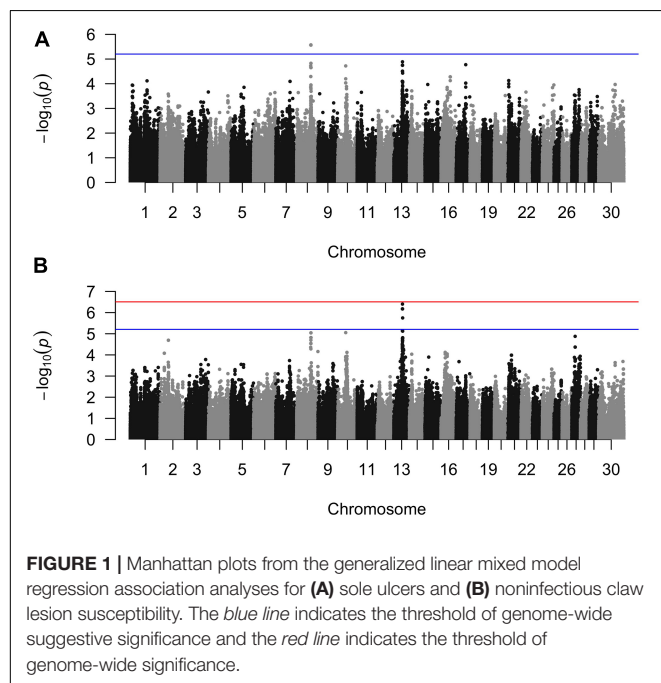
Random Forest GWAS

The RF models for all four datasets were not significantly more accurate at predicting the phenotype in the test population compared to the non-information rate (i.e., the frequency of the more common phenotype), indicating that the RF models were overfitted (Briec et al., 2018) such that the SNPs that passed the significance threshold were likely random noise. Because importance values are assigned and the importance threshold defined after fitting the RF model, some SNPs will always pass the importance threshold. Consequently, the value of these important SNPs and the likelihood that the important SNPs are truly trait linked must be gauged using model validation. In this case, the models were invalidated because of their poor phenotype prediction in the test population, indicating that the SNPs classified to be important were unlikely associated with the phenotype.

Additionally, the genomic regions identified by SNPs that passed the importance threshold did not overlap across the four datasets, despite their shared etiology, or with the genomic regions on BTA8 and BTA13 detected in the GLMM association

TABLE 4 | Proportion of phenotypic variance explained (PVE) by each linkage disequilibrium (LD) block defined from the generalized linear mixed model association analysis compared to the mean PVE of all chunks of genomic regions with the same length for sole ulcers (SU), white line disease (WLD), sole ulcers and/or white line disease (SU + WLD), and noninfectious claw lesions (NICL).

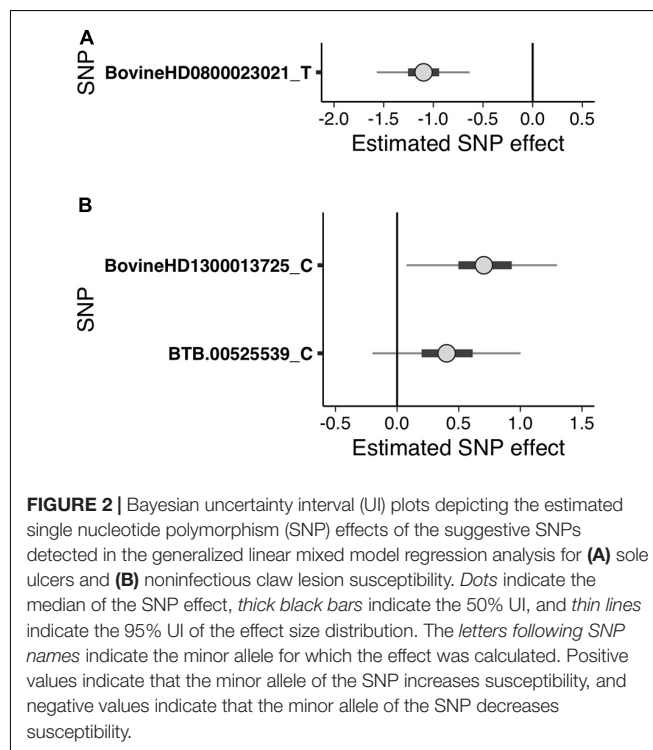
Dataset	LD block					Genome-wide mean of chunks with same length as LD block		
	BTA	Start (bp)	End (bp)	Length (kb)	PVE (SD)	PVE <i>p</i>	PVE (SE)	PVE <i>p</i> (SE)
SU	8	74,345,807	77,546,693	3,200.90	0.081 (0.054)	3.93E-04	0.00809 (0.0004)	0.478 (0.006)
WLD	13	46,307,416	47,584,595	1,277.20	0.061 (0.047)	2.93E-05	0.00794 (0.0002)	0.485 (0.004)
SU + WLD	13	45,283,136	47,676,681	2,393.50	0.071 (0.050)	1.05E-06	0.00873 (0.0004)	0.482 (0.006)
NICL	13	45,283,136	47,676,681	2,393.50	0.074 (0.051)	5.79E-09	0.00828 (0.0003)	0.484 (0.005)



analyses. Model overfitting combined with the lack of common genomic regions across the four datasets indicated that the RFs were unable to overcome the complex genetic architecture of noninfectious claw lesions and identify genomic regions of biological importance. Thus, downstream analyses to estimate SNP effects and conduct pathway and gene ontology analyses were not pursued.

DISCUSSION

Using GLMM regression, CBAT, and a RF approach to compare the SNP genotypes of sound controls and various types of noninfectious claw lesion cases, we identified genomic regions associated with susceptibility to these claw lesions. Given the overlapping etiology of the noninfectious claw lesion in this study, we expected that association testing would detect the genomic regions shared across some or all four datasets. Common genomic regions were identified from the GLMM and CBAT approaches, but not for the RF approach. Although RFs are a promising tool to identify loci associated with complex traits, the RF models in this study were overfitted, precluding



meaningful interpretation of the SNPs that passed the importance threshold. For GLMM testing and CBAT, the associated region detected on BTA8 for SU appeared to be specific for SU because the analyses for the other claw lesions did not detect this region; a SNP in this region (ARS-BFGL-NGS-108587) has previously been associated with SU (van der Spek et al., 2015b). The SNP detected on BTA13 for WLD increased in significance as cows with SU and other noninfectious lesions were added to the GLMM GWAS and CBAT analysis, implying that these lesions shared a genetic component that was less prevalent in SU cases. LD blocks defined by the top SNPs from the GLMM GWAS with nonzero effects from Bayesian estimation were explored further for candidate genes and previously defined QTL that were also functionally relevant to NICL etiology. The identification of promising candidate genes within the associated regions may lend more confidence to those regions; however, genetic selection does not require candidate gene identification and instead uses markers that are associated with, but not necessarily causal for, the trait. Thus, the candidate genes are presented below to

postulate their contribution to etiology rather than to inform genetic selection.

Sole ulcers and WLD are thought to result from increased laxity of the suspensory system from collagen breakdown and a thinner digital cushion, allowing the distal phalanx to rotate and sink within the claw (Lischer et al., 2002; Bicalho et al., 2009; Newsome et al., 2017a,b; Shearer and van Amstel, 2017; Stambuk et al., 2019). As the distal phalanx crushes the underlying corium, a hemorrhage develops at the pressure site and horn production through keratinization in the corium is disrupted, leading to horn thinning and, eventually, a hole in the horn through which the corium protrudes and develops into a SU (Greenough, 2007; Shearer et al., 2015). Similarly, WLD is thought to develop as a result of improper weight bearing and/or flooring causing defective horn production along the white line that is more prone to debris and bacteria infiltration, and when the bacteria reach the corium, an abscess forms (Shearer and van Amstel, 2017). It has been theorized that subclinical laminitis weakens the suspensory system and thereby predisposes the cow to SU and WLD (Thoenes et al., 2004), although evidence supporting this theory is limited (Danscher et al., 2010). New bone development on the third phalanx (Rusterholz, 1920; Blowey et al., 2000; Lischer et al., 2002) is associated with increasing age (Tsuka et al., 2012; Newsome et al., 2016) and is thought to contribute to a higher incidence of ulceration (Rusterholz, 1920; Tsuka et al., 2012). Because foot and leg conformation influences weight distribution within and between claws, the foot and leg conformation traits are thought to be correlated with SU + WLD susceptibility, although stronger evidence is needed to support the low to moderate phenotypic (Capon et al., 2008; Pérez-Cabal and Charfeddine, 2016) and genetic (Chapinal et al., 2013) correlations that were previously observed. Based on the etiology of noninfectious claw lesions and the possible genetic correlation of the susceptibility of these claw lesions with the conformation traits, genes and QTL related to collagen, keratinization, bone growth, adipose, and foot and leg conformation were considered functionally relevant.

For SU, the suggestive SNPs fell in or near *DCAF12* (DDB1 and CUL4-associated factor 12), an evolutionarily conserved apoptosis regulation gene involved in DNA repair and protein degradation that is required for tissue homeostasis under stress conditions, as demonstrated in *Drosophila* (Hwangbo et al., 2016). The metabolic stress associated with NICL could potentially disrupt the regulation of *DCAF12* and contribute to aberrant tissue homeostasis within the claw. Within the LD block, *APTX*, *AQP7*, *B4GALT1*, *ENHO*, *GALT*, *GULO*, and *UBAP2* had functions involved in wound healing, skin lesions, bone growth and mineralization, adipose tissue, and keratin summarized in **Table 5**. Notably, the LD block included a SNP that van der Spek et al. (2015b) had previously associated with SU susceptibility, ARS-BFGL-NGS-108587, supporting this SNP as a susceptibility locus for SU and the investigation into the region. No other previously defined QTL, physiologically relevant, or foot and leg conformation QTL were identified in the LD block. The two suggestive chunks on BTA17 both fell in *TMEM132B* (transmembrane protein 132B; **Table 5**), which, in humans, encodes a member of the TMEM132 family of evolutionarily ancient cell adhesion molecules that

connect the extracellular medium with the intracellular skeleton (Sanchez-Pulido and Ponting, 2018).

For NICL, all nine suggestive SNPs fell directly upstream or within introns of *DIP2C* (disco-interacting protein 2 homolog C), which is hypothesized to play a role in transcription and methylation regulation. *DIP2C* has been shown to regulate DNA methylation and the epithelial–mesenchymal transition in human cell lines (Larsson et al., 2017), and mutations in *DIP2C* have been associated with skeletal dysplasia affecting bone and cartilage development in humans (Maddirevula et al., 2018). The LD block contained three additional candidate genes with functions related to adipose tissue, bone growth, and bone mineralization (**Table 5**). The LD block on BTA13 did not overlap with previously defined QTL that were apparently related to NICL or foot and leg conformation traits. According to the Cattle Gene Atlas (Fang et al., 2020), some candidate genes were expressed ubiquitously (*DCAF12*, *APTX*, *GALT*, *UBAP2*, *DIP2C*, *PCNA*, and *WDR37*), and others were expressed more highly in specific tissues, such as adipose, cardiovascular, bone marrow, central nervous system, mammary, liver, or immune tissues (*AQP7*, *B4GALT1*, *ENHO*, *GULO*, and *RASSF2*; **Table 5**).

Prior GWAS studies of NICL, while having larger sample sizes, were sampled from larger geographical regions and used lower-density SNP panels. An acknowledged limitation of this study is the small sample size. However, previous GWAS with smaller sample sizes using the high-density SNP array were able to detect associated loci in Holstein populations for digital cushion thickness ($n = 502$) (Stambuk et al., 2020) and left displaced abomasum ($n = 406$) (Lehner et al., 2018), implying that locus detection is possible despite smaller sample sizes. By maintaining stringent phenotyping for sound controls, minimizing environmental and housing variability, and increasing SNP density, we aimed to optimize the ability to detect genomic variants at the expense of larger sample sizes. Additionally, the CBAT approach reduced the number of tests performed to increase power and found the same regions of association, providing further support for these regions. Because SU susceptibility is also affected by environmental management, including housing and nutrition, we sought to minimize environmental variability by sampling cows at dairies with similar nutrition and flooring, as the diets fed at the five dairies were similar and all dairies used a freestall flush barn system and rubber flooring in alleys.

Whereas previous published studies of noninfectious claw lesions have not used the high-density panel, our study with the 777K SNP panel allowed for higher resolution when defining the LD blocks. Furthermore, RF analysis and Bayesian regression methods were implemented to perform joint association testing of multiple top SNPs while working around the uneven sampling of controls. The two published GWAS for SU susceptibility found associated SNPs on different chromosomes than those identified in this study, specifically on BTA 8, 10, 11, 18, and 22 using a linear animal model (van der Spek et al., 2015b) and on BTA12 and 25 using a linear mixed model (Sánchez-Molano et al., 2019). Other GWAS for traits related to SU + WLD included digital cushion thickness (Sánchez-Molano et al., 2019; Stambuk et al., 2020), sole hemorrhage susceptibility (van der Spek et al., 2015b;

TABLE 5 | Candidate genes in linkage disequilibrium blocks defined by suggestive SNPs from the generalized linear mixed model and chunk-based association testing for sole ulcers (SU), white line disease (WLD), sole ulcers and/or white line disease (SU + WLD), and noninfectious claw lesions (NICL) and the tissues in which they were expressed.

Claw lesion	Gene symbol	Gene description	Functional relevance	RNA tissue specificity
SU	<i>DCAF12</i>	DDB1 (damage-specific binding protein) and CUL4 (cullin 4)-associated factor 12	Regulates apoptosis required for tissue homeostasis under stress conditions (Hwangbo et al., 2016)	Ubiquitous
	<i>APTX</i>	Aprataxin	Decreased bone mineral content (MGI) Increased total body fat amount (MGI)	Ubiquitous
	<i>AQP7</i>	Aquaporin 7	Abnormal white adipose tissue physiology (MGI) Increased fat cell size (MGI)	Adipose, cardiovascular, and bone marrow
	<i>B4GALT1</i>	Beta-1,4-galactosyltransferase 1	Decreased subcutaneous adipose tissue amount (MGI) Delayed wound healing (MGI) Hyperkeratosis (MGI) Skin lesions (MGI) Thin skin (MGI)	Mammary gland
	<i>ENHO</i>	Energy homeostasis associated	Increased body fat mass (MGI) Increased percent body fat/body weight (MGI)	Central nervous system
	<i>GALT</i>	Galactose-1-phosphate uridylyltransferase	Decreased subcutaneous adipose tissue amount (MGI) Delayed wound healing (MGI) Hyperkeratosis (MGI) Skin lesions (MGI) Thin skin (MGI)	Ubiquitous
	<i>GULO</i>	Gulonolactone (L-)-oxidase	Abnormal bone mineralization (MGI) Abnormal long bone epiphyseal plate morphology (MGI) Abnormal trabecular bone morphology (MGI) Decreased bone mineral density (MGI) Decreased compact bone thickness (MGI)	Liver
	<i>TMEM132B</i>	Transmembrane protein 132B	Cell adhesion molecule that connects the extracellular medium with the intracellular skeleton (Sanchez-Pulido and Ponting, 2018)	Central nervous system, testes
	<i>UBAP2</i>	Ubiquitin-associated protein 2	Abnormal adipose tissue amount (MGI)	Ubiquitous
	<i>DIP2C</i>	Disco-interacting protein 2 homolog C	Regulates DNA methylation and the epithelial-mesenchymal transition in human cell lines (Larsson et al., 2017)	Ubiquitous
WLD, SU + WLD, NICL	<i>PCNA</i>	Proliferating cell nuclear antigen	Mutations associated with skeletal dysplasia (Maddirevula et al., 2018) Abnormal adipose tissue development (MGI) Decreased percent body fat/body weight (MGI) Decreased white fat cell number	Ubiquitous

(Continued)

TABLE 5 | Continued

Claw lesion	Gene symbol	Gene description	Functional relevance	RNA tissue specificity
	RASSF2	Ras association (RalGDS/AF-6) domain family member 2	Abnormal bone mineralization (MGI)	White blood cells and immune tissues
			Abnormal trabecular bone morphology (MGI)	
			Decreased bone marrow cell number (MGI)	
			Decreased bone mass (MGI)	
			Decreased bone mineral density (MGI)	
			Decreased bone trabecula number (MGI)	
	WDR37	WD repeat domain 37	Decreased trabecular bone thickness (MGI)	Ubiquitous
			Decreased trabecular bone volume (MGI)	
			Increased bone mineral content (MGI)	

MGI, mammalian phenotype associated with gene from the Mammalian Genome Informatics batch query.

Sánchez-Molano et al., 2019), and laminitis susceptibility (Naderi et al., 2018), although the SNPs detected in these studies were also on different chromosomes from those from this study.

Because noninfectious claw lesions have a similar etiology, it has been postulated that pleiotropy may exist across the different noninfectious claw lesions and related traits. For instance, estimates of the genetic correlation between SU and WLD are significant, ranging from 0.41 to 0.60 depending on parity (van der Linde et al., 2010). However, past GWAS have not found associations on the same chromosomes among SU, WLD, digital cushion thickness, sole hemorrhage, or laminitis (van der Spek et al., 2015b; Naderi et al., 2018; Sánchez-Molano et al., 2019), or if SNPs from the same chromosome were detected, they were in different regions. Specifically, the only common chromosome among these three GWAS was BTA11: van der Spek et al. (2015b) found Hapmap38795-BTA-97039 for SU at BTA11:23302850, and Naderi et al. (2018) found BTB-00466773 for laminitis at BTA11:48309332 (the SNP positions were updated to the ARS-UCD1.2 map). The QTL identified on BTA13 may thus represent a portion of the common genetic contribution to the different types of noninfectious claw lesions.

CONCLUSION

Using logistic mixed model single-marker regression and CBAT, genomic regions associated with susceptibility were identified on BTA8 for SU and BTA13 for WLD, SU + WLD, and NICL. The associated regions on BTA8 and BTA13 contained candidate genes related to wound healing, skin lesions, bone growth and mineralization, adipose tissue, and keratin. The RF approach was unable to overcome the complexity of these lesion traits and reliably identify potential candidate QTL. Although these findings must be validated in larger populations in other geographical regions, the detection of a region associated with SU susceptibility that included a previously reported locus suggested that the study cohort was adequate to identify the regions of susceptibility for NICL. Further exploration of these regions through targeted sequencing or RNA-seq in claw tissues with and without noninfectious claw lesions may uncover variants in the genes or regulatory elements contributing to lameness. The multiplicity of associations detected in this and other studies demonstrated the complexity of the genetic architecture underlying noninfectious claw lesion susceptibility.

DATA AVAILABILITY STATEMENT

The microarray datasets generated for this study can be found in NCBI's Gene Expression Omnibus data repository (GEO series record GSE159157 and GSE165945).

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee. Written informed consent was

obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

AO, TF, AD, and EL conceptualized the research aims with supervision from AO and TF and prepared and edited the manuscript. AD and EL arranged the blood sample collection and processing, curated hoof trimming records, developed the methodology, and performed the computational analyses. TF was instrumental in developing code for the computational analyses. AO provided the funding, lab space, and computing resources for this research. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by a Western Sustainable Agriculture Research and Education Graduate Student grant (project number GW18-126), U.S. Department of Agriculture Cooperative State

Research, Education, and Extension Service Animal Health Funding (project number 2250-AH), a Department of Animal Science Kellogg Endowment, and Jastro Shields awards from the College of Agricultural and Environmental Sciences at the University of California, Davis.

ACKNOWLEDGMENTS

We thank the dairy producers and hoof trimmers for their participation and expertise. We gratefully acknowledge the infrastructure support of the Department of Animal Science, College of Agricultural and Environmental Sciences, and the California Agricultural Experiment Station of the University of California, Davis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.657375/full#supplementary-material>

REFERENCES

- Bicalho, R. C., Machado, V. S., and Caixeta, L. S. (2009). Lameness in dairy cattle: A debilitating disease or a disease of debilitated cattle? A cross-sectional study of lameness prevalence and thickness of the digital cushion. *J. Dairy Sci.* 92, 3175–3184. doi: 10.3168/jds.2008-1827
- Blowey, R. W., Ossent, P., Watson, C. L., Hedges, V., Green, L. E., and Packington, A. J. (2000). Possible distinction between sole ulcers and heel ulcers as a cause of bovine lameness. *Vet. Rec.* 147, 110–112. doi: 10.1136/vr.147.4.110
- Briec, M. S. O., Waters, C. D., Drinan, D. P., and Naish, K. A. (2018). A practical introduction to random forest for genetic association studies in ecology and evolution. *Mol. Ecol. Resour.* 18, 755–766.
- Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015
- Capion, N., Thamsborg, S. M., and Enevoldsen, C. (2008). Conformation of hind legs and lameness in danish holstein heifers. *J. Dairy Sci.* 91, 2089–2097. doi: 10.3168/jds.2006-457
- Cha, E., Hertl, J. A., Bar, D., Gröhn, Y. T., and Grohn, Y. T. (2010). The cost of different types of lameness in dairy cows calculated by dynamic programming. *Prev. Vet. Med. J.* 97, 1–8. doi: 10.1016/j.prevetmed.2010.07.011
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8
- Chapinal, N., Koeck, A., Sewalem, A., Kelton, D. F., Mason, S., Cramer, G., et al. (2013). Genetic parameters for hoof lesions and their relationship with feet and leg traits in Canadian Holstein cows. *J. Dairy Sci.* 96, 2596–2604. doi: 10.3168/jds.2012-6071
- Charfeddine, N., and Pérez-Cabal, M. A. (2017). Effect of claw disorders on milk production, fertility, and longevity, and their economic impact in Spanish Holstein cows. *J. Dairy Sci.* 100, 653–665. doi: 10.3168/jds.2016-11434
- Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* 98, 653–666. doi: 10.1016/j.ajhg.2016.02.012
- Christopoulos, D. (2017). Introducing unit invariant knee (UIK) as an objective choice for elbow point in multivariate data analysis techniques. *SSRN Electron. J.* doi: 10.2139/ssrn.3043076
- Christopoulos, D. T. (2016). On the efficient identification of an inflection point on the efficient identification of an inflection point. *Int. J. Math. Sci. Comput.* 6, 13–20.
- Cramer, G., Lissemore, K. D., Guard, C. L., Leslie, K. E., and Kelton, D. F. (2008). Herd- and cow-level prevalence of foot lesions in ontario dairy cattle. *J. Dairy Sci.* 91, 3888–3895. doi: 10.3168/jds.2008-1135
- Danscher, A. M., Toelboell, T. H., and Wattle, O. (2010). Biomechanics and histology of bovine claw suspensory tissue in early acute laminitis. *J. Dairy Sci.* 93, 53–62. doi: 10.3168/jds.2009-2038
- DeFraen, J. M., Socha, M. T., and Tomlinson, D. J. (2013). Analysis of foot health records from 17 confinement dairies. *J. Dairy Sci.* 96, 7329–7339. doi: 10.3168/jds.2012-6017
- Dolecheck, K. A., Overton, M. W., Mark, T. B., and Bewley, J. M. (2019). Use of a stochastic simulation model to estimate the cost per case of digital dermatitis, sole ulcer, and white line disease by parity group and incidence timing. *J. Dairy Sci.* 102, 715–730. doi: 10.3168/jds.2018-14901
- Eicher, S. D., Lay, D. C., Arthington, J. D., and Schutz, M. M. (2013). Effects of rubber flooring during the first 2 lactations on production, locomotion, hoof health, immune functions, and stress1. *J. Dairy Sci.* 96, 3639–3651. doi: 10.3168/jds.2012-6049
- Fang, L., Cai, W., Liu, S., Canela-Xandri, O., Gao, Y., Jiang, J., et al. (2020). Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* 30, 790–801. doi: 10.1101/gr.250704.119
- Fjeldaas, T., Sogstad, Å.M., and Østerås, O. (2011). Locomotion and claw disorders in Norwegian dairy cows housed in freestalls with slatted concrete, solid concrete, or solid rubber flooring in the alleys. *J. Dairy Sci.* 94, 1243–1255. doi: 10.3168/jds.2010-3173
- Functional Annotation of Animal Genomes (FAANG) (2019). *FAANG Mine*. Available online at: <http://128.206.116.18:8080/faangmine/begin.do> (accessed August 3, 2020)
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *J. R. Stat. Soc. Ser. A* 182, 389–402. doi: 10.1111/rssa.12378
- Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., et al. (2020). *R Package ARM: Data Analysis Using Regression and Multilevel/Hierarchical Models*. Available online at: <https://cran.r-project.org/package=arm> (accessed January 21, 2021).
- Goldstein, B. A., Hubbard, A. E., Cutler, A., and Barcellos, L. F. (2010). An application of random forests to a genome-wide association dataset:

- methodological considerations and new findings. *BMC Genet.* 11:49. doi: 10.1186/1471-2156-11-49
- Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). *Rstanarm: Bayesian Applied Regression Modeling Via Stan*. Available online at: <https://mc-stan.org/rstanarm/> (accessed January 21, 2021).
- Green, L. E., Borkert, J., Monti, G., and Tadich, N. (2010). Associations between lesion-specific lameness and the milk yield of 1,635 dairy cows from seven herds in the Xth region of Chile and implications for management of lame dairy cows worldwide. *Anim. Welf.* 19, 419–427.
- Green, L. E., Hedges, V. J., Schukken, Y. H., Blowey, R. W., and Packington, A. J. (2002). The impact of clinical lameness on the milk yield of dairy cows. *J. Dairy Sci.* 85, 2250–2256.
- Greenough, P. R. (2007). *Bovine Laminitis and Lameness*, 1st Edn, eds C. Bergsten, A. Brizzir, and C. K. W. Mülling (Canada: Elsevier Ltd), doi: 10.1016/B978-0-7020-2780-2.X5001-0
- Häggman, J., and Juga, J. (2013). Genetic parameters for hoof disorders and feet and leg conformation traits in Finnish Holstein cows. *J. Dairy Sci.* 96, 3319–3325. doi: 10.3168/jds.2012-6334
- Häggman, J., Juga, J., Sillanpää, M. J., and Thompson, R. (2013). Genetic parameters for claw health and feet and leg conformation traits in Finnish Ayrshire cows. *J. Anim. Breed. Genet.* 130, 89–97. doi: 10.1111/j.1439-0388.2012.01007.x
- Hernandez, J. A., Garbarino, E. J., Shearer, J. K., Risco, C. A., and Thatcher, W. W. (2005). Comparison of milk yield in dairy cows with different degrees of lameness. *J. Am. Vet. Med. Assoc.* 227, 1292–1296. doi: 10.2460/javma.2005.227.1292
- Hu, Z. L., Park, C. A., and Reecy, J. M. (2019). Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* 47, D701–D710. doi: 10.1093/nar/gky1084
- Hwangbo, D. S., Biteau, B., Rath, S., Kim, J., and Jasper, H. (2016). Control of apoptosis by Drosophila DCAF12. *Dev. Biol.* 413, 50–59. doi: 10.1016/j.ydbio.2016.03.003
- Koufariotis, L., Chen, Y. P. P., Bolormaa, S., and Hayes, B. J. (2014). Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. *BMC Genomics* 15:1–16. doi: 10.1186/1471-2164-15-436
- Koufariotis, L. T., Chen, Y. P. P., Stothard, P., and Hayes, B. J. (2018). Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. *BMC Genomics* 19:237. doi: 10.1186/s12864-018-4617-x
- Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05
- Lai, E., Danner, A. L., Famula, T. R., and Oberbauer, A. M. (2020). Genome-wide association studies reveal susceptibility loci for digital dermatitis in holstein cattle. *Animals* 10:2009. doi: 10.3390/ani10112009
- Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247. doi: 10.1038/ng1195-241
- Larsson, C., Ali, M. A., Pandzic, T., Lindroth, A. M., He, L., and Sjöblom, T. (2017). Loss of DIP2C in RKO cells stimulates changes in DNA methylation and epithelial-mesenchymal transition. *BMC Cancer* 17:487. doi: 10.1186/s12885-017-3472-5
- Lehner, S., Zerbin, I., Doll, K., Rehage, J., and Distl, O. (2018). A genome-wide association study for left-sided displacement of the abomasum using a high-density single nucleotide polymorphism array. *J. Dairy Sci.* 101, 1258–1266. doi: 10.3168/jds.2017-13216
- Li, M.-X., Yeung, J. M. Y., Cherny, S. S., and Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* 131, 747–756. doi: 10.1007/s00439-011-1118-2
- Lischer, C. J., Ossent, P., Räber, M., and Geyer, H. (2002). Suspensory structures and supporting tissues of the third phalanx of cows and their relevance to the development of typical sole ulcers (*Rusterholz ulcers*). *Vet. Rec.* 151, 694–698. doi: 10.1136/vr.151.23.694
- Listgarten, J., Lippert, C., Kang, E. Y., Xiang, J., Kadie, C. M., and Heckerman, D. (2013). A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 29, 1526–1533. doi: 10.1093/bioinformatics/btt177
- Maddirevula, S., Alsahli, S., Alhabeeb, L., Patel, N., Alzahrani, F., Shamseldin, H. E., et al. (2018). Expanding the phenome and variome of skeletal dysplasia. *Genet. Med.* 20, 1609–1616. doi: 10.1038/gim.2018.50
- Makanjuola, B. O., Miglior, F., Abdalla, E. A., Maltecca, C., Schenkel, F. S., and Baes, C. F. (2020). Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. *J. Dairy Sci.* 103, 5183–5199. doi: 10.3168/jds.2019-18013
- Malchiodi, F., Koeck, A., Chapinal, N., Sargolzaei, M., Fleming, A., Kelton, D. F., et al. (2015a). Genetic analyses of hoof lesions in canadian holsteins using an alternative contemporary group. *Int. Bull.* 49, 64–68.
- Malchiodi, F., Koeck, A., Christen, A. M., Schenkel, F. S., Kelton, D. F., and Miglior, F. (2015b). *Genetic Parameters and Genome Wide Association Study of Individual Hoof Lesions in Canadian Holsteins Using Different Contemporary Groups*. Guelph: Canadian Dairy Network.
- Moore, C. M., Jacobson, S. A., and Fingerlin, T. E. (2019). Power and sample size calculations for genetic association studies in the presence of genetic model misspecification. *Hum. Hered.* 84, 256–271. doi: 10.1159/000508558
- Mostert, P. F., van Middelaar, C. E., de Boer, I. J. M., and Bokkers, E. A. M. (2018). The impact of foot lesions in dairy cows on greenhouse gas emissions of milk production. *Agric. Syst.* 167, 206–212. doi: 10.1016/j.agry.2018.09.006
- Naderi, S., Bohlouli, M., Yin, T., König, S., and König, S. (2018). Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets. *Anim. Genet.* 49, 178–192. doi: 10.1111/age.12661
- National Research Council (NRC) (2001). *Nutrient Requirements of Dairy Cattle*, 7th Edn. Washington, D.C.: National Academies Press, doi: 10.17226/9825
- Newsome, R. F., Green, M. J., Bell, N. J., Bollard, N. J., Mason, C. S., Whay, H. R., et al. (2017a). A prospective cohort study of digital cushion and corium thickness. part 1: associations with body condition, lesion incidence, and proximity to calving. *J. Dairy Sci.* 100, 4745–4758. doi: 10.3168/jds.2016-12012
- Newsome, R. F., Green, M. J., Bell, N. J., Bollard, N. J., Mason, C. S., Whay, H. R., et al. (2017b). A prospective cohort study of digital cushion and corium thickness. part 2: does thinning of the digital cushion and corium lead to lameness and claw horn disruption lesions? *J. Dairy Sci.* 100, 4759–4771. doi: 10.3168/jds.2016-12013
- Newsome, R. F., Green, M. J., Bell, N. J., Chagunda, M. G. G., Mason, C. S., Rutland, C. S., et al. (2016). Linking bone development on the caudal aspect of the distal phalanx with lameness during life. *J. Dairy Sci.* 99, 4512–4525. doi: 10.3168/jds.2015-10202
- Oberbauer, A. M., Berry, S. L., Belanger, J. M., McGoldrick, R. M., Pinos-Rodriguez, J. M., and Famula, T. R. (2013). Determining the heritable component of dairy cattle foot lesions. *J. Dairy Sci.* 96, 605–613. doi: 10.3168/jds.2012-5485
- Pérez-Cabal, M. A., and Charfeddine, N. (2016). Short communication: association of foot and leg conformation and body weight with claw disorders in Spanish Holstein cows. *J. Dairy Sci.* 99, 9104–9108. doi: 10.3168/jds.2016-11331
- Purcell, S. M., and Chang, C. C. (2015). *Plink 1.9*. Available online at: <https://www.cog-genomics.org/plink2> (accessed December 4, 2020).
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Available online at: <https://www.r-project.org/> (accessed January 21, 2021).
- Richardson, I. W., Berry, D. P., Wiencko, H. L., Higgins, I. M., More, S. J., McClure, J., et al. (2016). A genome-wide association study for genetic susceptibility to *Mycobacterium bovis* infection in dairy cattle identifies a susceptibility QTL on chromosome 23. *Genet. Sel. Evol.* 48:19. doi: 10.1186/s12711-016-0197-x
- Ring, S. C., Twomey, A. J., Byrne, N., Kelleher, M. M., Pabiou, T., Doherty, M. L., et al. (2018). Genetic selection for hoof health traits and cow mobility scores can accelerate the rate of genetic gain in producer-scored lameness in dairy cows. *J. Dairy Sci.* 101, 10034–10047. doi: 10.3168/jds.2018-15009
- Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elisk, C. G., Tseng, E., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9, 1–9. doi: 10.1093/gigascience/giaa021
- Rusterholz, A. (1920). Das spezifisch-traumatische Klauensohlengeschwür des Rindes (the specific traumatic sole ulcer of claws in cattle). *Schweiz. Arch. Tierheilkd.* 62, 505–525.
- Sánchez-Molano, E., Bay, V., Smith, R. F., Oikonomou, G., and Banos, G. (2019). Quantitative trait loci mapping for lameness associated phenotypes in holstein-friesian dairy cattle. *Front. Genet.* 10:926. doi: 10.3389/fgene.2019.00926

- Sanchez-Pulido, L., and Ponting, C. P. (2018). TMEM132: an ancient architecture of cohesin and immunoglobulin domains define a new family of neural adhesion molecules. *Bioinformatics* 34, 721–724. doi: 10.1093/bioinformatics/btx689
- Shearer, J. K., Plummer, P., and Schleining, J. (2015). Perspectives on the treatment of claw lesions in cattle. *Vet. Med. Res. Reports* 6, 273–292. doi: 10.2147/vmrr.s62071
- Shearer, J. K., and van Amstel, S. R. (2001). Functional and corrective claw trimming. *Vet. Clin. North Am. - Food Anim. Pract.* 17, 53–72. doi: 10.1016/S0749-0720(15)30054-2
- Shearer, J. K., and van Amstel, S. R. (2017). Pathogenesis and treatment of sole ulcers and white line disease. *Vet. Clin. North Am. Food Anim. Pract.* 33, 283–300. doi: 10.1016/j.cvfa.2017.03.001
- Smith, C. L., and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 1, 390–399. doi: 10.1002/wsbm.44
- Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* 91, 1011–1021. doi: 10.1016/j.ajhg.2012.10.010
- Stambuk, C. R., McArt, J. A. A., Bicalho, R. C., Miles, A. M., and Huson, H. J. (2019). A longitudinal study of digital cushion thickness and its function as a predictor for compromised locomotion and hoof lesions in Holstein cows. *Transl. Anim. Sci.* 3, 74–83. doi: 10.1093/tas/txy107
- Stambuk, C. R., Staiger, E. A., Nazari-Ghadikolaei, A., Heins, B. J., and Huson, H. J. (2020). Phenotypic characterization and genome-wide association studies of digital cushion thickness in Holstein cows. *J. Dairy Sci.* 103, 3289–3303. doi: 10.3168/jds.2019-17409
- Thoenes, M. B., Pollitt, C. C., Van Eps, A. W., Milinovich, G. J., Trott, D. J., Wattle, O., et al. (2004). Acute bovine laminitis: a new induction model using alimentary oligofructose overload. *J. Dairy Sci.* 87, 2932–2940. doi: 10.3168/jds.S0022-0302(04)73424-4
- Tsuka, T., Ooshita, K., Sugiyama, A., Osaki, T., Okamoto, Y., Minami, S., et al. (2012). Quantitative evaluation of bone development of the distal phalanx of the cow hind limb using computed tomography. *J. Dairy Sci.* 95, 127–138. doi: 10.3168/jds.2011-4316
- Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *bioRxiv [preprint]*:5165. doi: 10.1101/005165
- Twomey, A. J., Berry, D. P., Evans, R. D., Doherty, M. L., Graham, D. A., and Purfield, D. C. (2019). Genome-wide association study of endo-parasite phenotypes using imputed whole-genome sequence data in dairy and beef cattle. *Genet. Sel. Evol.* 51, 1–17. doi: 10.1186/s12711-019-0457-7
- van der Linde, C., de Jong, G., Koenen, E. P. C., and Eding, H. (2010). Claw health index for Dutch dairy cattle based on claw trimming and conformation data. *J. Dairy Sci.* 93, 4883–4891. doi: 10.3168/jds.2010-3183
- van der Spek, D., van Arendonk, J. A. M., and Bovenhuis, H. (2015a). Genetic relationships between claw health traits of dairy cows in different parities, lactation stages, and herds with different claw disorder frequencies. *J. Dairy Sci.* 98, 6564–6571. doi: 10.3168/jds.2015-9561
- van der Spek, D., van Arendonk, J. A. M., and Bovenhuis, H. (2015b). Genome-wide association study for claw disorders and trimming status in dairy cattle. *J. Dairy Sci.* 98, 1286–1295. doi: 10.3168/jds.2014-8302
- van der Spek, D., van Arendonk, J. A. M., Vallée, A. A., and Bovenhuis, H. (2013). Genetic parameters for claw disorders and the effect of preselecting cows for trimming. *J. Dairy Sci.* 96, 6070–6078. doi: 10.3168/jds.2013-6833
- Van der Waaij, E. H., Holzhauer, M., Ellen, E., Kamphuis, C., and De Jong, G. (2005). Genetic parameters for claw disorders in Dutch dairy cattle and correlations with conformation traits. *J. Dairy Sci.* 88, 3672–3678.
- Vanegas, J., Overton, M., Berry, S. L., and Sischo, W. M. (2006). Effect of rubber flooring on claw health in lactating dairy cows housed in free-stall barns. *J. Dairy Sci.* 89, 4251–4258. doi: 10.3168/jds.S0022-0302(06)72471-7
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Andrew, Y., Bürkner, P.-C., et al. (2020). *loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. Available online at: <https://mc-stan.org/loo/> (accessed January 21, 2021).
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4
- Xia, J., Fan, H., Chang, T., Xu, L., Zhang, W., Song, Y., et al. (2017). Searching for new loci and candidate genes for economically important traits through gene-based association analysis of Simmental cattle. *Sci. Rep.* 7, 1–9. doi: 10.1038/srep42048
- Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Lai, Danner, Famula and Oberbauer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genomic Prediction of Average Daily Gain, Back-Fat Thickness, and Loin Muscle Depth Using Different Genomic Tools in Canadian Swine Populations

Siavash Salek Ardestani¹, Mohsen Jafarikia^{2,3}, Mehdi Sargolzaei^{4,5}, Brian Sullivan² and Younes Miar^{1*}

¹ Department of Animal Science and Aquaculture, Dalhousie University, Truro, NS, Canada, ² Canadian Centre for Swine Improvement, Ottawa, ON, Canada, ³ Centre for Genetic Improvement of Livestock (CGIL), Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada, ⁴ Department of Pathobiology, University of Guelph, Guelph, ON, Canada, ⁵ Select Sires Inc., Plain City, OH, United States

OPEN ACCESS

Edited by:

Abdoulaye Baniré Diallo,
Université du Québec à Montréal,
Canada

Reviewed by:

Allan Schinckel,
Purdue University, United States
Ismo Strandén,
Natural Resources Institute Finland
(Luke), Finland
Linyuan Shen,
Sichuan Agricultural University, China

*Correspondence:

Younes Miar
miar@dal.ca

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 07 February 2021

Accepted: 15 April 2021

Published: 03 June 2021

Citation:

Salek Ardestani S, Jafarikia M,
Sargolzaei M, Sullivan B and Miar Y
(2021) Genomic Prediction of Average
Daily Gain, Back-Fat Thickness,
and Loin Muscle Depth Using
Different Genomic Tools in Canadian
Swine Populations.
Front. Genet. 12:665344.
doi: 10.3389/fgene.2021.665344

Improvement of prediction accuracy of estimated breeding values (EBVs) can lead to increased profitability for swine breeding companies. This study was performed to compare the accuracy of different popular genomic prediction methods and traditional best linear unbiased prediction (BLUP) for future performance of back-fat thickness (BFT), average daily gain (ADG), and loin muscle depth (LMD) in Canadian Duroc, Landrace, and Yorkshire swine breeds. In this study, 17,019 pigs were genotyped using Illumina 60K and Affymetrix 50K panels. After quality control and imputation steps, a total of 41,304, 48,580, and 49,102 single-nucleotide polymorphisms remained for Duroc ($n = 6,649$), Landrace ($n = 5,362$), and Yorkshire ($n = 5,008$) breeds, respectively. The breeding values of animals in the validation groups ($n = 392-774$) were predicted before performance test using BLUP, BayesC, BayesC π , genomic BLUP (GBLUP), and single-step GBLUP (ssGBLUP) methods. The prediction accuracies were obtained using the correlation between the predicted breeding values and their deregressed EBVs (dEBVs) after performance test. The genomic prediction methods showed higher prediction accuracies than traditional BLUP for all scenarios. Although the accuracies of genomic prediction methods were not significantly ($P > 0.05$) different, ssGBLUP was the most accurate method for Duroc-ADG, Duroc-LMD, Landrace-BFT, Landrace-ADG, and Yorkshire-BFT scenarios, and BayesC π was the most accurate method for Duroc-BFT, Landrace-LMD, and Yorkshire-ADG scenarios. Furthermore, BayesC π method was the least biased method for Duroc-LMD, Landrace-BFT, Landrace-ADG, Yorkshire-BFT, and Yorkshire-ADG scenarios. Our findings can be beneficial for accelerating the genetic progress of BFT, ADG, and LMD in Canadian swine populations by selecting more accurate and unbiased genomic prediction methods.

Keywords: BayesC, genomic prediction, GBLUP, single-step GBLUP, swine

INTRODUCTION

In the pork industry, genetics has a pivotal role in improving the economically important traits, such as meat quality, growth, and reproductive performance (Badke et al., 2014). The genetic improvement has been spectacularly successful using traditional genetic improvement tools, such as best linear unbiased prediction (BLUP) method (Henderson, 1975); however, the genetic gain achieved is relatively slow for traits that are expensive and difficult to measure such as those measured postmortem (Dekkers et al., 2010; Wolc et al., 2011; Miar et al., 2015). On the other hand, evaluation methods based on genotypic information or genomic evaluation are being used increasingly in modern pig breeding industry (Badke et al., 2014) because of their accelerative role in genetic improvement specifically through augmentation of predictive ability.

In genomic evaluation, the effects of markers are estimated based on genomic and phenotypic data in the reference group, and then it is used to predict genomic breeding values (GEBVs) in the validation group (Hayes and Goddard, 2001). The accuracy of genomic prediction can be influenced by several factors such as the reference population size (Samore and Fontanesi, 2016), genetic architecture of the trait, marker density, and method assumptions used for prediction (Zhang et al., 2019). Several former studies have compared different methods on genomic prediction accuracy in livestock species such as cattle (Luan et al., 2009; Guo et al., 2017; Mehrban et al., 2017; Li et al., 2018; Mrode et al., 2018) and mink (Karimi et al., 2019). However, there are still limited numbers of literature in pig (Song et al., 2017; Jafarikia et al., 2018; Zhang et al., 2018). Some popular genomic evaluation approaches such as genomic BLUP (GBLUP) and Bayesian methods have been widely used in the recent studies. For example, Zhang et al. (2018) showed higher predictive ability of BayesB method over GBLUP for average daily feed intake in a small population ($n = 1,363$) of Duroc pigs. The single-step GBLUP (ssGBLUP) method was considerably more accurate than GBLUP and BayesR for growth traits in a larger population ($n = 2,084$) of Yorkshire breed (Song et al., 2017). The differences between these approaches are their assumptions, for example, about the distribution of marker effects (Hayes and Goddard, 2001). Therefore, determining the most accurate method for genomic evaluation of different swine breeds is an important step of selective breeding.

Several genomic prediction models of GBLUP (Badke et al., 2014; Song et al., 2017, 2019b, 2020; Jafarikia et al., 2018; Zhang et al., 2018), ssGBLUP (Song et al., 2017, 2019b; Thekkoot et al., 2018; Hong et al., 2019; Lopez et al., 2019; Zhou et al., 2019; Aliakbari et al., 2020), and BayesC (Esfandyari et al., 2016; Song et al., 2020) have frequently been used for prediction of various traits in the previous studies in swine. The main assumption of GBLUP method is based on the infinitesimal model (i.e., the genetic variation of the trait was explained by a large number of loci) (Karaman et al., 2016, 2018) that has been widely used in genomic evaluation, principally due to its ease of implementation, in developed countries'

breeding programs, such as Holstein cattle breeding programs in Canada¹. Similar to GBLUP method, the infinitesimal model is assumed in ssGBLUP method. The pivotal difference between ssGBLUP and GBLUP methods is applying a blended genetic relationship matrix (using genomic and pedigree relationship matrices) in the ssGBLUP method (Legarra et al., 2009). This method has been implemented in developing countries and breeding companies with small number of genotyped animals due to the improving role of blended relationship matrix in increasing the accuracy of predicted GEBVs (Mrode et al., 2018). This method can also allow detecting the conflicts of pedigree and regulating the relationship between genotyped and non-genotyped animals (Amaya Martínez et al., 2020). Additionally, breeding values for non-genotyped animals can be obtained simultaneously using ssGBLUP method, which is not the case in GBLUP analysis (Tsuruta et al., 2013; Misztal et al., 2014). The Bayesian genomic evaluation approaches (A, B, C, etc.) are nonlinear methods and are mostly applied using Markov chain Monte Carlo (MCMC) algorithm (Iheshiulor et al., 2017). Despite the linear genomic evaluation methods (e.g., ssGBLUP and GBLUP), some Bayesian methods (e.g., BayesC π) assume that the genetic variation is explained by a fewer number of loci. This characteristic can be helpful to improve the evaluation accuracies for traits where their genetic architecture violates the infinitesimal model assumption (Hayes and Goddard, 2001; VanRaden et al., 2008). However, GBLUP can be considerably faster than Bayesian approaches in terms of computational speed (Song et al., 2019b). In BayesC model (Kizilkaya et al., 2010), a common variance for single-nucleotide polymorphisms (SNPs) with nonzero effects is assumed instead of a locus-specific variance. Although assuming that the probability of SNPs with nonzero effects (π) is known, it might be problematic for some traits in BayesC model. Habier et al. (2011) developed the BayesC model through hypothesizing unknown π that can be estimated, and therefore its prior distribution becomes uniform (0, 1) (Habier et al., 2011).

The prediction ability of genomic evaluation methods, which is considerably affected by their assumptions, is an important factor for genetic improvement in swine breeding companies. Therefore, the main goal of our study was to compare the prediction accuracies of traditional BLUP with different popular genomic evaluation methods including GBLUP, ssGBLUP, BayesC, and BayesC π for average daily gain (ADG), back-fat thickness (BFT), and loin muscle depth (LMD) traits in Canadian swine populations.

MATERIALS AND METHODS

Ethics Statement

The hogs used in this study were cared for according to the Canadian Council on Animal Care (Olfert et al., 1993) guidelines.

¹<https://interbull.org>

Animals, Genotyping, Quality Control, and Imputation

In this study, we used genotypic and phenotypic data of BFT, ADG, and LMD in three swine breeds of Duroc, Landrace, and Yorkshire. The BFT and LMD were measured by ultrasonic machine (B mode), and ADG was calculated using Eq. 1:

$$\text{ADG} = \frac{\text{Final weight} - \text{birth weight}}{\text{Days}} \quad (1)$$

These phenotypes were collected from 2010 to 2019 and adjusted to the weight of 120 kg (Table 1). The average numbers of phenotyped boars and gilts per litter for Duroc, Landrace, and Yorkshire breeds are reported in **Supplementary Table 1**. In this study, phenotypic and genotypic information was collected from distributed swine breeding companies across Canada, which participated in the Canadian Swine Improvement Program coordinated by the Canadian Centre for Swine Improvement².

Animals in the reference ($n = 4,615$ – $5,875$) and validation (the performance tested animals after January 2019, $n = 392$ – 774) groups were genotyped with Illumina 60K (Illumina Inc., San Diego, CA, United States) or Affymetrix 50K (Affymetrix, Santa Clara, CA, United States) panels (Table 1). Quality control was performed by removing SNPs with minor allele frequency <0.05 , call rate <0.9 , and Hardy–Weinberg $P < 0.0001$. After quality control steps, the remaining SNPs (52,025 for all breeds) were used for imputation of the missing genotypes with FImpute 2.2 software (Sargolzaei et al., 2014). After the imputation step, 1,341 SNPs on sex chromosomes were discarded.

Statistical and Genetic Analyses

Traditional Estimated Breeding Value

Model 1 was used for estimation of breeding values of each animal using the AIREMLF90 1.61 software (Misztal et al., 2002):

$$\text{Model 1. } y_c = I\mu + Xb + Za + Wu + e,$$

where y is the vector of phenotypic data, μ is the overall mean, X is the incidence matrix relating fixed effects of herd–year–season–sex to phenotypes, b is the vector of fixed effects, Z is the incidence matrix relating phenotypes to additive genetic effects, a is the vector of additive genetic effects, W is the incidence matrix relating phenotypes to random common litter effects, u is the vector of random common litter effects, and e is the vector of random residual effects. It was assumed that $a \sim N(0, A\sigma_a^2)$, $u \sim N(0, I\sigma_u^2)$, and $e \sim N(0, I\sigma_e^2)$, where A is the pedigree-based relationship matrix, σ_a^2 is the variance of additive genetic effects, σ_u^2 is the variance of common litter effects, I is the identity matrix, and σ_e^2 is the residual variance. The estimated variance components (before and after performance tests) using AIREMLF90 implemented in BLUP model (Model 1) are reported in **Supplementary Table 2**. The estimated breeding

values (EBVs) of parents were used to calculate parent average EBV (PA) for each animal using the following equation:

$$\text{PA} = \frac{\text{EBV (sire)} + \text{EBV (dam)}}{2} \quad (2)$$

GBLUP

After calculating the deregressed EBVs (dEBVs) as pseudophenotypes according to the approach proposed by Garrick et al. (2009), the GBLUP method was performed using Model 2 implemented in SNP1101 software 1.0 (Sargolzaei, 2014).

$$\text{Model 2. } y_c = 1\mu + Zg + e$$

In Model 2, y_c is the vector of dEBVs (reference population) as pseudophenotypes, μ is the overall mean effect, Z is the incidence matrix relating phenotypes (dEBVs) to GEBVs, g is the vector of GEBVs, and e is the vector of random residual effects. It was assumed that $g \sim N(0, G\sigma_g^2)$ and $e \sim N(0, W\sigma_e^2)$, where G is the genomic relationship matrix, σ_g^2 is the genomic variance, W is a diagonal matrix of residual weights, and σ_e^2 is the residual variance. The residual weights ($w_i = \frac{1-r_i^2}{r_i^2}$) were calculated based on the reliability of dEBVs (r_i^2) as described by Garrick et al. (2009).

The genomic relationship matrix was constructed as VanRaden (2008) described:

$$G = \frac{ZZ'}{2 \sum_{j=1}^i p_j(1-p_j)}, \quad (3)$$

where Z is the allele frequency adjusted genotype matrix with $0-2p_j$ (for AA genotype), $1-2p_j$ (for AB genotype), and $2-2p_j$ (for BB genotype) elements, and dimension of the number of individuals by the number of markers. p_j is the minor allele frequency for j -th marker.

The estimated variance components obtained from the “aireml” procedure (in SNP1101 software) implemented in GBLUP model (Model 2) are reported in **Supplementary Table 2**. The genomic relationship matrix visualization was performed using a custom-made script in python.

Single-Step GBLUP

The ssGBLUP analysis (Legarra et al., 2009; Christensen and Lund, 2010) was performed using AIREMLF90 1.61 software (Misztal et al., 2014). Model 3 was used for single-step genomic evaluation of each animal:

$$\text{Model 3. } y = 1\mu + Xb + Zg + Wu + e$$

where y , μ , X , b , W , u , and e were explained in Model 1, Z is the incidence matrix relating phenotypes to GEBVs, and g is the vector of GEBVs. It was assumed that the variance of genomic effects (σ_g^2), variance of common litter effects (σ_u^2), and residual variance (σ_e^2) are governed by the normal distribution ($g \sim N(0, H\sigma_g^2)$, $u \sim N(0, I\sigma_u^2)$, and $e \sim N(0, I\sigma_e^2)$, respectively). In ssGBLUP model, the H matrix was used, which was a

²<https://www.ccsi.ca/>

TABLE 1 | Number of animals with phenotypes and genotypes in reference and validation groups for back-fat thickness at 120 kg (BFT), average daily gain from birth to 120 kg (ADG), and loin muscle depth at 120 kg (LMD) in three breeds of Duroc, Landrace, and Yorkshire (validation groups were phenotyped after January 2019).

Trait	Breed	No. of phenotypes			No. of genotypes					
		Boar	Gilt/sow	Total	Boar	Reference Gilt/sow	Total	Boar	Validation Gilt/sow	Total
BFT (mm)	Duroc	24,461	25,588	50,049	3,690	2,184	5,874	504	270	774
	Landrace	32,304	50,259	82,563	2,519	2,371	4,890	467	4	471
	Yorkshire	33,421	60,916	94,337	2,160	2,455	4,615	386	7	393
ADG (g/day)	Duroc	24,470	25,590	50,060	3,691	2,184	5,875	504	270	774
	Landrace	32,311	50,288	82,599	2,519	2,372	4,891	467	4	471
	Yorkshire	33,428	60,947	94,375	2,161	2,455	4,616	385	7	392
LMD (mm)	Duroc	24,461	25,588	50,049	3,690	2,184	5,874	504	270	774
	Landrace	32,304	50,259	82,563	2,519	2,371	4,890	467	4	471
	Yorkshire	33,421	60,916	94,337	2,160	2,455	4,615	386	7	393

combination of relationship matrices based on marker genotypes (G) and pedigree information (A). In the inverse of H matrix, A_{22} is the pedigree-based relationship matrix for genotyped animals, and τ and ω are the scaling factors, which both were set equal to one as the default option in AIREMLF90 1.61 software (Misztal et al., 2014). The blending factors of G (α) and A (β) matrices in the inverse of H matrix were set equal to 0.95 and 0.05, respectively, which were defined as $H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha G + \beta A_{22})^{-1} - \omega A_{22}^{-1} \end{bmatrix}$ to avoid singularity problems and improve predictions (VanRaden, 2008; Lourenco et al., 2014). The estimated variance components from AIREMLF90 implemented in ssGBLUP model (Model 3) are reported in **Supplementary Table 2**.

Bayesian Approaches

In BayesC framework, the genomic evaluation was performed by implementing Model 4 and MCMC process in GS3 2.0 software (Legarra et al., 2011a) with the following criteria: NITER (number of iterations) = 100,000, BURNIN (beginning of the MCMC run) = 20,000, and THIN (thin interval) = 100. The applied prior of variance components were similar to the estimated variance components in BLUP model before performance test (**Supplementary Table 2**).

$$\text{Model 4. } y_C = 1\mu \sum_{i=1}^n Z_i \alpha_i \delta_i + e$$

In Model 4, y_C is the vector of pseudophenotypes as defined in Model 2, μ is the overall mean, n is the number of SNPs, Z_i is the vector of genotype covariates, α_i is the allele substitution effect for SNP $_i$, δ_i is an indicator for whether the SNP $_i$ has effect (1) or not (0), and e is the vector of random residual effects. The residuals were weighed based on the reliabilities of dEBVs as defined by Legarra et al. in GS3 software (Legarra et al., 2011a). For BayesC π , the π prior as the probability level of an SNP having no effect was set equal to 0.99.

Validation and Model Comparison

The accuracies of genomic predictions and PAs were calculated as the correlation between breeding values (GEBVs or PAs) of the validation group and their dEBVs after performance test. The standard errors of prediction accuracies were calculated using Eq. 4:

$$\text{Standard error} = \frac{1 - \text{accuracy}^2}{\sqrt{\text{number of individuals} - 1}} \quad (4)$$

The regression coefficients of dEBVs (after performance test in January 2019) on predicted breeding values (before performance test in January 2019) were calculated to evaluate the bias of predictions (**Figure 1**). The regression coefficients and their standard errors were calculated using “lm” and “summary” functions in R 4.0.2 software (R Core Team, 2013).

The accuracy improvements were calculated using Eq. 5:

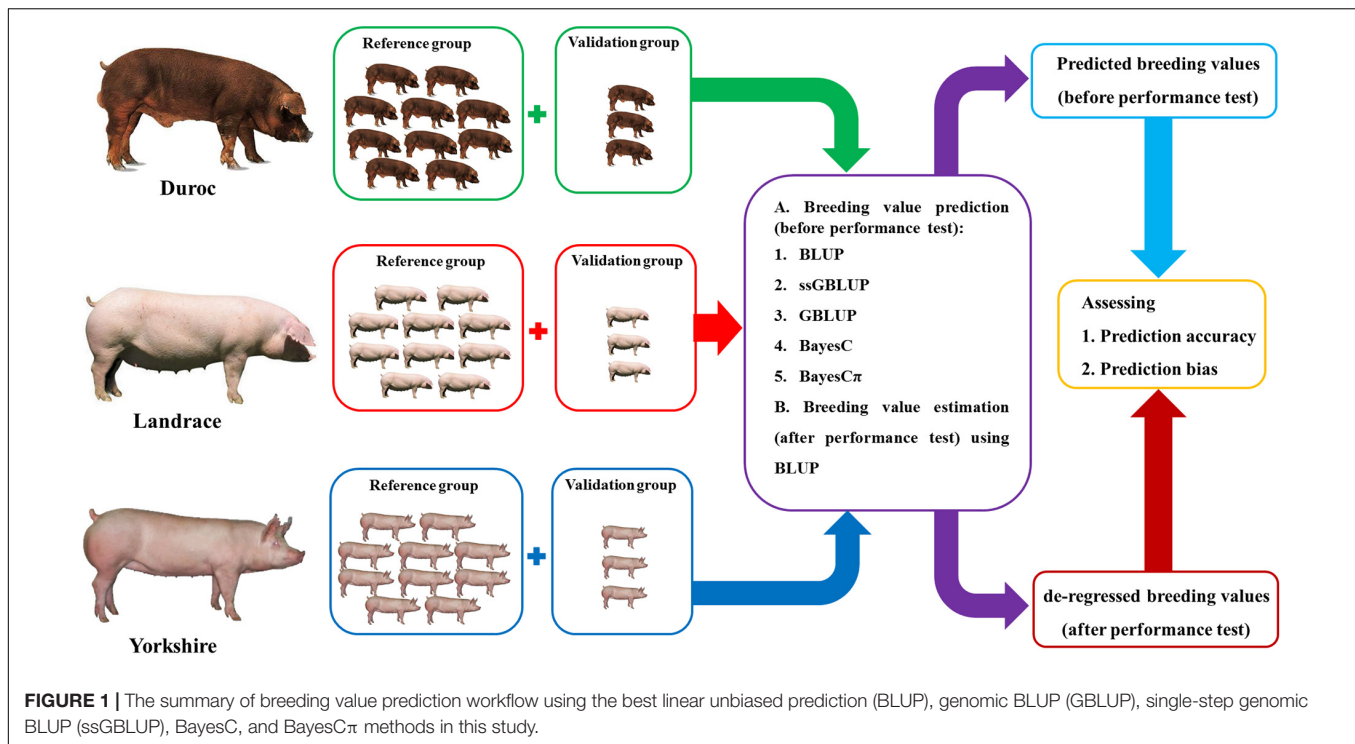
$$\text{Improvement of accuracy} = \left(\frac{\text{accuracy of GEBV} - \text{accuracy of PA}}{\text{accuracy of PA}} \right) \times 100. \quad (5)$$

Two criteria of accuracy and regression coefficient of dEBV on GEBV were used to compare the predictive ability of different genomic prediction methods. Additionally, a fixed model equation (Model 5) was employed to detect the significant differences of the prediction accuracies obtained from different methods.

$$\text{Model 5. } y_{ij} = \mu + M_i + B_j + e_{ij},$$

where y is the accuracy of prediction for the trait, μ is the overall mean, M is the fixed effect of prediction method i , B is the fixed effect of breed j , and e is the random residual effect for i -th method and j -th breed. The computational times of different scenarios with the same number of threads ($n = 80$) and memories (202 GB) were calculated using the default option of Slurm 20.02.3 software³ in Compute Canada server

³<https://slurm.schedmd.com>



(Niagara cluster⁴). In this study, the computing time included the construction of genomic relationship matrix and its inverse, variance components estimation, and solving models.

RESULTS AND DISCUSSION

Descriptive Statistics

The standard deviation, minimum and maximum values, and the coefficient of variation for each trait and breed are given in **Table 2**. The minimum ($n = 50,060$) and maximum ($n = 94,375$) numbers of phenotyped animals were observed for Duroc-ADG and Yorkshire-ADG scenarios. The Landrace-BFT (14.27 mm), Duroc-ADG (736.72 g/day), and Duroc-LMD (72.61 mm) had the highest averages. The coefficients of variation ranged from 18.78 (Duroc) to 23.07 (Landrace), from 8.34 (Landrace) to 8.65 (Yorkshire), and from 8.25 (Duroc) to 8.83 (Yorkshire) for BFT, ADG, and LMD, respectively.

Validation of Genomic Evaluations

After quality control, removing SNPs on sex chromosomes, and imputation, the total numbers of 41,304, 48,580, and 49,102 SNPs remained for genomic evaluation of Duroc, Landrace, and Yorkshire breeds, respectively. Finding an accurate and unbiased genomic prediction method can be a lucrative strategy for genetic improvement of key traits in livestock species (Mrode et al., 2018). The predictive ability of genomic methods depends on various factors such as method hypothesis (Momen et al., 2018). However, limited numbers of literature are available in swine

(Esfandyari et al., 2016; Song et al., 2017, 2019a; Zhang et al., 2018). Several genomic evaluation studies in livestock species such as cattle (Cardoso et al., 2014; Neves et al., 2014; Brown et al., 2016; Júnior et al., 2016; Silva et al., 2016; Costa et al., 2019) investigated and compared different genomic prediction methods; for example, Silva et al. (2016) showed the superior prediction accuracy of ssGBLUP over BayesC π and GBLUP methods for residual feed intake (RFI) and feed conversion ratio (FCR) in Nelore cattle. Therefore, comparing different genomic prediction methods is important to detect more accurate methods for genomic evaluation of key traits in the swine breeding industry.

The average computational times for GBLUP, ssGBLUP, BayesC, and BayesC π were 00:00:33, 00:18:47, 15:16:37, and 15:40:02 h, respectively. The computational times of GBLUP, ssGBLUP, BayesC, and BayesC π ranged from 00:00:27 h (Yorkshire-BFT and Landrace-BFT) to 00:00:43 h (Duroc-LMD), 00:09:54 h (Duroc-LMD) to 00:31:43 h (Landrace-BFT), 14:21:19 h (Yorkshire-ADG) to 16:21:31 h (Duroc-ADG), and 14:16:28 h (Yorkshire-BFT) to 16:22:40 h (Duroc-LMD), respectively (**Supplementary Table 3**). The correlation between dEBVs (after performance test) and predicted GEBVs (before performance test) was considered as the accuracy of genomic prediction (**Table 3**). Moreover, the correlation between PAs (before performance test) and dEBVs (after performance test) was calculated as the accuracy of PAs (**Table 3**).

The accuracies ranged from 13.9% (Duroc-PA) to 52.7% (Landrace-ssGBLUP) for BFT, from 5.7% (Duroc-PA) to 34.5% (Landrace-ssGBLUP) for ADG, and from 3.7% (Duroc-PA) to 25.1% (Landrace-BayesC π) for LMD. The accuracy improvements over PA ranged from 35.6% (Yorkshire-BayesC)

⁴<https://docs.scinet.utoronto.ca>

TABLE 2 | Descriptive statistics of back-fat thickness at 120 kg (BFT), average daily gain from birth to 120 kg (ADG), and loin muscle depth at 120 kg (LMD) for Yorkshire, Landrace, and Duroc breeds.

Breed	Trait	n	Mean	SD	Min	Max	CV (%)
BFT (mm)	Duroc	50,049	12.5	2.55	4.9	32.8	18.78
	Landrace	82,563	14.27	3.29	5.1	39.4	23.07
	Yorkshire	94,337	14.26	3.04	5.8	39.2	21.30
ADG (g/day)	Duroc	50,060	736.72	62.56	354.4	1,072.1	8.49
	Landrace	82,599	711.74	59.34	396	1,077.5	8.34
	Yorkshire	94,375	706.97	61.16	353.9	1,061.6	8.65
LMD (mm)	Duroc	50,049	72.61	5.99	41.3	101.1	8.25
	Landrace	82,563	68.39	5.95	38.7	96.4	8.71
	Yorkshire	94,337	69.51	6.14	39.6	103.7	8.83

n, number of animals per trait; *SD*, standard deviation; *Min*, minimum value; *Max*, maximum value; *CV*, coefficient of variation.

to 204.6% (Duroc-BayesC π) for BFT, from 62.1% (Landrace-BayesC) to 314.3% (Duroc-ssGBLUP) for ADG, and from 6.4% (Landrace-BayesC) to 284.5% (Yorkshire-GBLUP) for LMD (Table 3). The prediction accuracies of genomic methods were not significantly ($P > 0.05$) different from each other based on Tukey test implemented in Model 5 (Supplementary Table 4). However, the prediction accuracies obtained from PA were significantly ($P < 0.05$) lower than those obtained from genomic methods except for BayesC in LMD trait (Supplementary Table 4). Moreover, the prediction accuracies obtained from genomic methods were significantly ($P < 0.05$) different among breeds except for Yorkshire-Duroc (for BFT) and Yorkshire-Landrace (for ADG and LMD).

Back-Fat Thickness

The results demonstrated that ssGBLUP was the most accurate method for genomic evaluation of BFT in Landrace (52.7%) and Yorkshire (44.7%) breeds (Figure 2). The ssGBLUP method had the same prediction accuracy (44.1%) for BFT in Chinese Yorkshire population as our prediction (Zhou et al., 2019). However, the accuracy of ssGBLUP method for Yorkshire-BFT scenario was considerably higher than Thekkoot et al.'s (2018) result (30%), which might be due to their smaller number of genotyped animals in the reference group ($n = 2,489$). Higher prediction accuracies obtained from ssGBLUP for BFT in Yorkshire and Landrace breeds compared to Duroc breed might be due to the similarity of these breeds as maternal lines as well as using the larger population size (pedigree and phenotype data) of Yorkshire-BFT ($n = 94,337$) and Landrace-BFT ($n = 82,563$) in comparison with Duroc-BFT ($n = 50,049$). For BFT, the prediction accuracy superiority of ssGBLUP method over the other genomic prediction methods such as GBLUP was confirmed by previous studies in American Yorkshire breed (Song et al., 2017, 2019a). The regression coefficients of dEBVs on predicted GEBVs were estimated and applied as indices of prediction bias of the genomic evaluation methods (Table 4).

The regression coefficients higher and lower than 1 indicate overestimation and underestimation, respectively. The ssGBLUP method showed higher prediction accuracies in Landrace-BFT and Yorkshire-BFT scenarios, but BayesC π was the least biased method in these scenarios (1.06 for Landrace-BFT and 1.12 for

Yorkshire-BFT). In Duroc-BFT scenario, BayesC π (42.4%) and GBLUP (1.11) were the most accurate and least biased methods, respectively. The higher prediction accuracy of BayesC π for Duroc-BFT scenario might be due to a low number of major-effect SNPs underlying genetic variation of BFT in Duroc breed (Zhang et al., 2020). However, Tukey test showed that the genomic prediction accuracies for BFT were not significantly ($P > 0.05$) different across breeds. In a previous genomic prediction study on a small population size of Canadian Duroc breed ($n = 1,363$), Zhang et al. (2018) showed the superiority of BayesRC approach for BFT. The BayesRC is a new method based on BayesR that combines prior biological datasets through explaining variant classes presumably to be enriched for causal polymorphisms (MacLeod et al., 2016). Compared to our results, they (Zhang et al., 2018) showed higher prediction accuracy (62 vs. 42.4%) for Duroc-BFT using whole-genome sequence data that might be due to using sequence data and a different method in their prediction.

Average Daily Gain

Our results indicated that ssGBLUP was the most accurate method (34.5%) for prediction of ADG in Landrace breed (Figure 3); however, BayesC π was the least biased (1) method for ADG genomic evaluation (Table 4). Our prediction accuracy of ssGBLUP for Landrace-ADG (34.5%) was slightly higher than the results of Hong et al. (2019) (31–33%), who used different blending strategies of pedigree and genomic relationship matrices in ssGBLUP method. For Yorkshire-ADG, they obtained 21–22% accuracy of prediction using ssGBLUP method (Hong et al., 2019). In this study, BayesC π was the most accurate (29.5%), and least biased (1.06) method for Yorkshire-ADG scenario. The prediction accuracy obtained from ssGBLUP method in Duroc-ADG scenario (23.7%) was similar to the study by Jiao et al. (2014) (24.10%) and slightly lower than that by Zhang et al. (2018) (25%). Jiao et al. (2014) used the Bayes A model in a small population size ($n = 1,022$) of Duroc breed in the reference group. The implementation of genomic multitrait models using ADG, FCR, and RFI traits might be beneficial for improvement of prediction accuracy for ADG in Duroc breed through implementing information from genetically correlated traits (Guo et al., 2014). Additionally, the prediction accuracy is

TABLE 3 | The prediction accuracies (%), their standard errors and their improvement (%) over parent average EBV (PA) for back-fat thickness (BFT), average daily gain (ADG), and loin muscle depth (LMD) traits in Duroc, Landrace, and Yorkshire breeds.

Trait	Breed	PA	GBLUP		ssGBLUP		BayesC		BayesC π	
		Accuracy	Accuracy	Improvement	Accuracy	Improvement	Accuracy	Improvement	Accuracy	Improvement
BFT	Duroc	13.9(3.5)	39.1(3.1)	181.4	39.2(3)	182.1	35.0(3.2)	151.4	42.4(3)	204.6
	Landrace	28.5(4.2)	49.2(3.5)	72.4(3.5)	52.7(3.3)	84.6	52.6(3.3)	84.4	50.8(3.4)	78.1
	Yorkshire	30.4(4.6)	42.4(4.1)	39.8(4.1)	44.7(4)	47.3	41.1(4.2)	35.6	41.2(4.2)	35.6
ADG	Duroc	5.7(3.6)	20.9(3.4)	266.4	23.7(3.4)	314.3	21.5(3.4)	276.7	19.3(3.5)	238
	Landrace	16.4(4.5)	33.0(4.1)	100.8	34.5(4.1)	110.0	26.6(4.3)	62.1	32.8(4.1)	99.7
	Yorkshire	12.0(5)	28.8(4.6)	140.7	26.2(4.6)	119.5	27.5(4.7)	129.8	29.5(4.6)	147
LMD	Duroc	3.7(3.6)	12.0(3.6)	225.6	12.6(3.5)	240.8	11.2(3.6)	202.6	10.3(3.6)	178.4
	Landrace	17.2(4.5)	22.5(4.4)	31.2	22.8(4.4)	33.1	18.3(4.5)	6.4	25.1(4.3)	46.2
	Yorkshire	5.6(5)	21.6(4.8)	284.5	21.3(4.8)	277.9	17.7(4.9)	213.6	20.7(4.8)	267.4

GBLUP, genomic BLUP; ssGBLUP, single-step genomic BLUP.

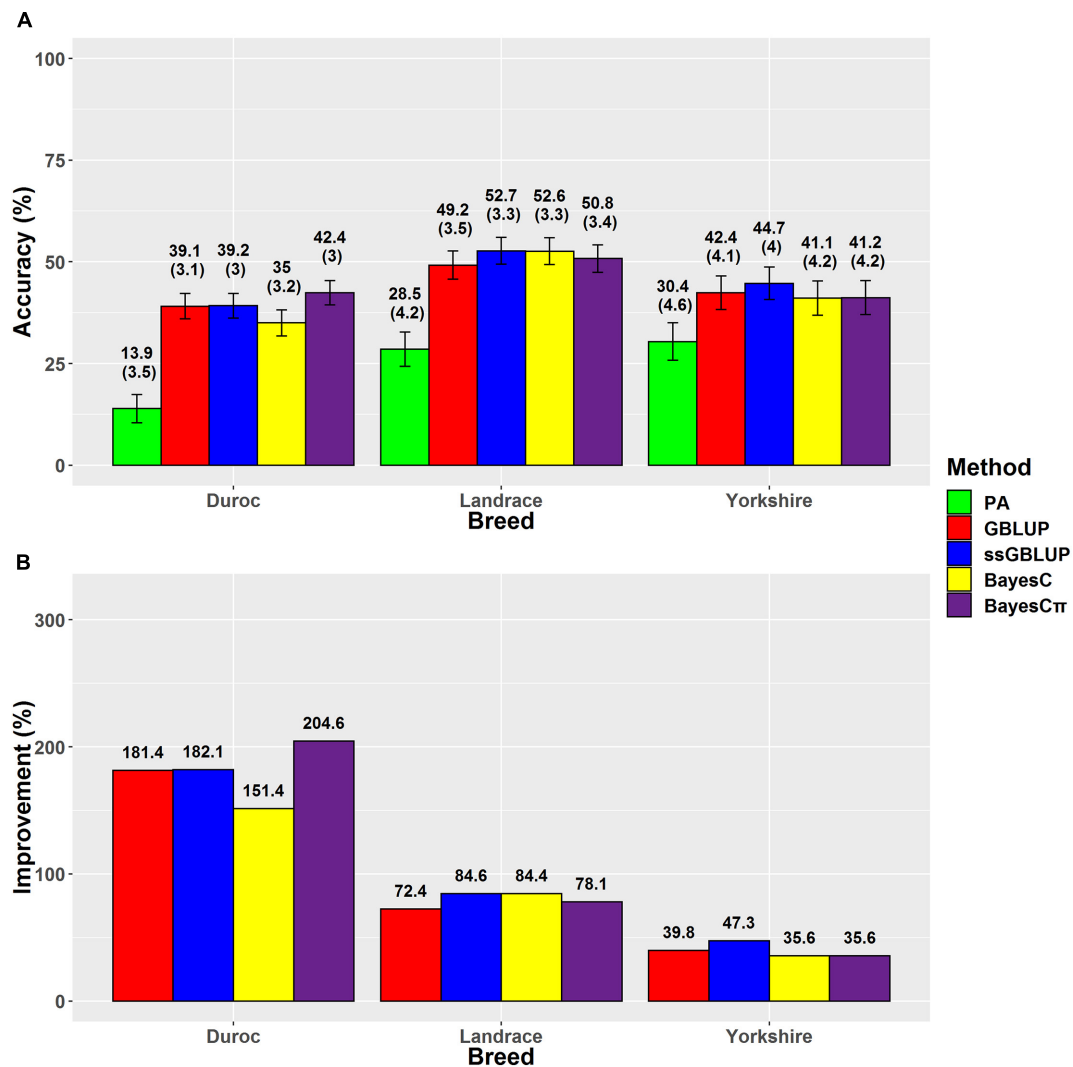
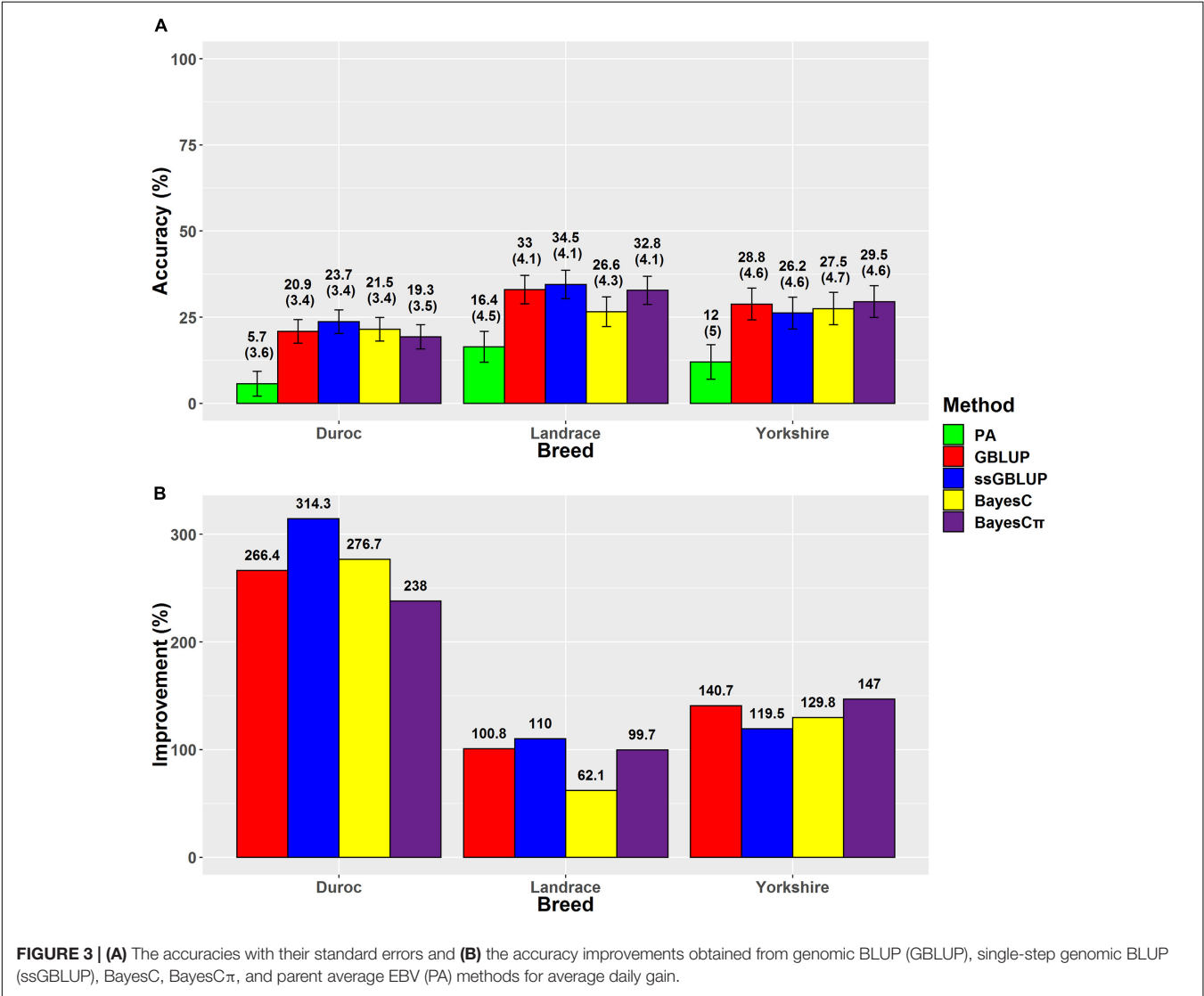
**FIGURE 2 | (A)** The accuracies with their standard errors and **(B)** accuracy improvements obtained from genomic BLUP (GBLUP), single-step genomic BLUP (ssGBLUP), BayesC, BayesC π , and parent average EBV (PA) methods for back-fat thickness.

TABLE 4 | Regression coefficient and their standard errors of deregressed EBV (dEBV) on predicted breeding values obtained from genomic BLUP (GBLUP), single-step genomic BLUP (ssGBLUP), BayesC, BayesC π , and parent average EBV (PA) for back-fat thickness (BFT), average daily gain (ADG), and loin muscle depth (LMD) in Duroc, Landrace, and Yorkshire breeds.

Trait	Breed	PA	GBLUP	ssGBLUP	BayesC	BayesC π
BFT	Duroc	0.90 (0.23)	1.11 (0.09)	1.60 (0.13)	0.58 (0.05)	1.12 (0.08)
	Landrace	0.76 (0.11)	1.15 (0.09)	1.14 (0.08)	0.62 (0.04)	1.05 (0.08)
	Yorkshire	1.16 (0.18)	1.22 (0.13)	1.31 (0.13)	0.55 (0.06)	1.11 (0.12)
ADG	Duroc	0.98 (0.62)	1.15 (0.19)	2.36 (0.34)	1.23 (0.20)	1.27 (0.23)
	Landrace	1.00 (0.28)	1.17 (0.15)	1.27 (0.16)	0.37 (0.06)	1.0 (0.13)
	Yorkshire	0.85 (0.36)	1.30 (0.22)	1.23 (0.23)	0.46 (0.80)	1.05 (0.17)
LMD	Duroc	0.58 (0.56)	1.21 (0.36)	1.48 (0.42)	0.40 (0.12)	0.83 (0.20)
	Landrace	0.75 (0.20)	1.01 (0.20)	0.90 (0.17)	0.32 (0.08)	0.90 (0.16)
	Yorkshire	0.34 (0.31)	1.15 (0.26)	0.97 (0.22)	0.27 (0.07)	0.82 (0.19)



considerably related to the genomic heritability, which is affected by the number of markers (Goddard, 2009). By comparing the genomic prediction abilities using different SNP densities (whole-genome sequence, 650K and 80K SNP panels) and prediction methods, Zhang et al. (2018) revealed that the density of SNPs could affect the prediction accuracy in a small population size of Canadian Duroc breed ($n = 1,363$) (Zhang et al., 2018). Although, except for BayesRC (25%) method, the GBLUP

(12%), and BayesB (12%) methods showed lower prediction accuracies using the whole-genome sequence data in the study by Zhang et al. (2018) compared to our result (23.7%) for Duroc-ADG-ssGBLUP, these results can confirm the importance of population size, method assumptions, and SNP densities in genomic prediction analysis.

Loin Muscle Depth

The highest accuracies for LMD in Duroc, Landrace, and Yorkshire breeds were derived from ssGBLUP (12.6%), BayesC π (25.1%), and GBLUP (21.6%) methods, respectively (**Figure 4**). However, the regression coefficients followed a different trend for LMD scenarios. For LMD scenarios, the least biased methods were BayesC π in Duroc (0.83), GBLUP in Landrace 1.01, and ssGBLUP (0.97) in Yorkshire breed. In this study, the highest genomic prediction accuracies for Yorkshire-LMD (21.6%) and Landrace-LMD (25.1%) scenarios were lower than the results by Jafarikia et al. (2018) for Yorkshire

(38%) and Landrace (38%) that might be due to their larger reference group size of Canadian Yorkshire ($n = 8,756$) and Landrace ($n = 6,754$) pigs (Jafarikia et al., 2018), although, our genomic prediction accuracy (25.1%) for Landrace-LMD using the BayesC π method was higher than the result of another study on Landrace (17.9–18.8%) pigs using different genomic relationship matrices (Sevillano et al., 2017). In Duroc-LMD scenario (GBLUP = 12%, ssGBLUP = 12.6%, BayesC = 11.2, and BayesC π = 10.3%), the genomic prediction accuracies were lower than Landrace-LMD (GBLUP = 22.5%, ssGBLUP = 22.8%, BayesC = 18.3%, and BayesC π = 25.1%) and Yorkshire-LMD (GBLUP = 21.6%, ssGBLUP = 21.3%, BayesC = 17.7%, and BayesC π = 20.7%) that were similar to most of the scenarios for ADG and BFT traits. A small portion of these lower accuracies in Duroc scenarios might be due to slightly lower genomic relationships between validation and reference groups in Duroc (**Supplementary Figure 1**) compared to Landrace (**Supplementary Figure 2**) and Yorkshire

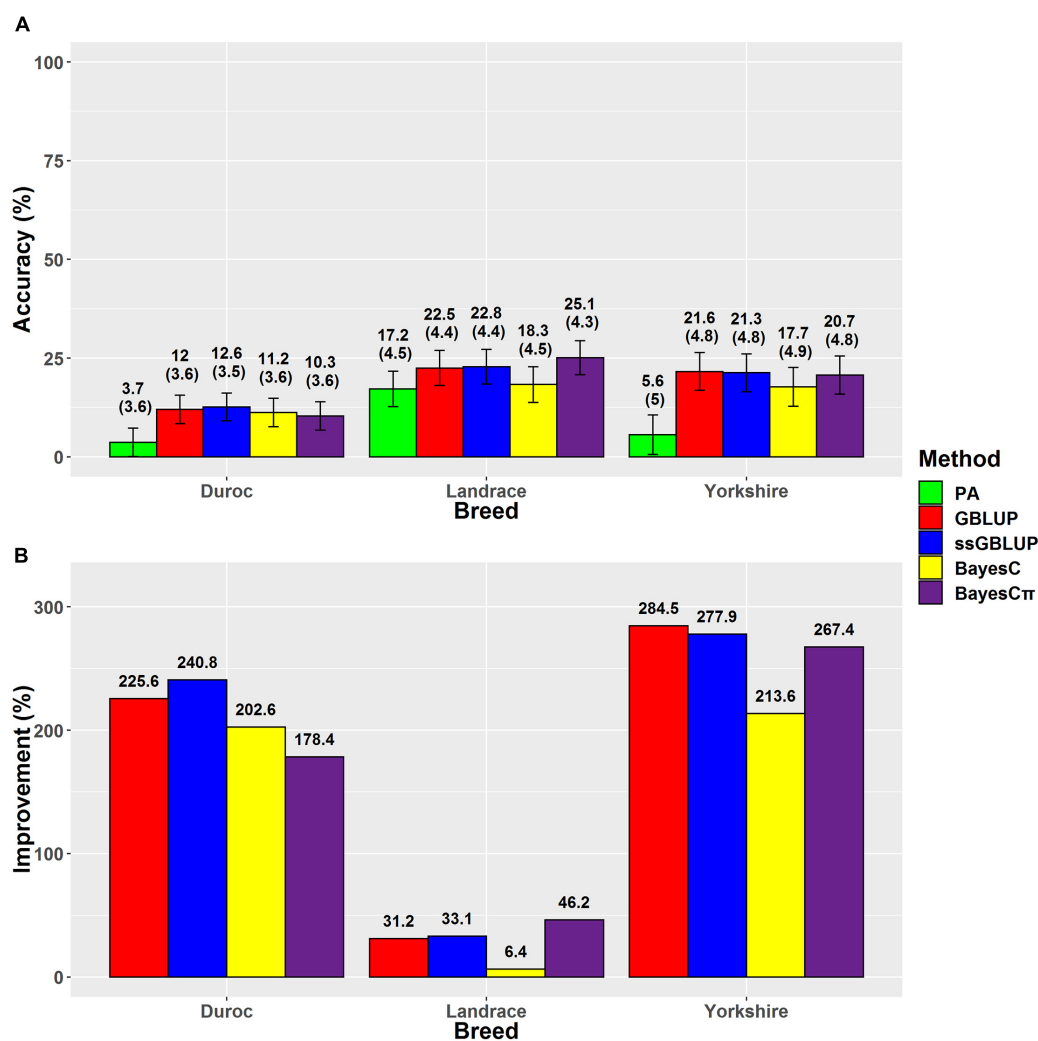


FIGURE 4 | (A) The accuracies with their standard errors and **(B)** the accuracy improvements obtained from genomic BLUP (GBLUP), single-step genomic BLUP (ssGBLUP), BayesC, BayesC π , and parent average EBV (PA) methods for loin muscle depth.

(**Supplementary Figure 3**). Our results also showed the same patterns for genomic relationships among animals within the validation sets. However, the patterns were reverse in the pedigree-based relationships. The negative effects of weak genomic relationships between validation and reference groups on predictive ability were reported previously (Hayes et al., 2009a; Song et al., 2017).

The prediction accuracy is also affected by response variables through correlating predicted breeding values and response variables (adjusted phenotype, EBV, or dEBV) after performance test. The reliability of dEBV as a response variable is highly dependent on the population size of phenotyped animals (Mrode et al., 2018). In Duroc, the main reason for these low accuracies can be referred to the low reliabilities of dEBVs in the validation group derived from the smaller population size of phenotyped Duroc (50,049–50,060) compared to Yorkshire (94,337–94,375) and Landrace (82,563–82,599) (**Table 1**), whereas the reference population size of Duroc (5,874–5,875) was higher than Yorkshire (4,615–4,616) and Landrace (4,890–4,891). Additionally, the size of validation group in Duroc (774) was higher than Yorkshire (471) and Landrace (392–393), which could be an important factor in calculating the genomic prediction accuracy. The validation group size and selection of response variable are important factors on prediction accuracy, which have not been highlighted in the previous genomic prediction studies in pig (Badke et al., 2014; Song et al., 2017, 2019a; Thekkoot et al., 2018; Zhang et al., 2018; Lopez et al., 2019; Aliakbari et al., 2020).

It is evident that ssGBLUP was the most accurate method in five scenarios of Duroc-ADG, Duroc-LMD, Landrace-BFT, Landrace-ADG, and Yorkshire-BFT, and BayesC π was the most accurate method in three scenarios of Duroc-BFT, Landrace-LMD, and Yorkshire-ADG. However, it should be noted that the accuracies obtained from different genomic prediction methods were not significantly ($P > 0.05$) different in all scenarios. High prediction accuracies obtained from ssGBLUP method in aforementioned scenarios might be due to the blended pedigree and genotypic data used for prediction of GEBVs (Silva et al., 2016; Mrode et al., 2018). The assumption of ssGBLUP method is based on the infinitesimal model of polygenic control of the trait (Karaman et al., 2016, 2018). Although we obtained the higher prediction accuracies for ssGBLUP compared to other methods in five scenarios, our regression coefficients showed considerably underestimated GEBVs for Duroc-BFT (1.60), Duroc-ADG (2.36), and Duroc-LMD (1.48) using ssGBLUP in comparison with the other methods. A reason for these underestimated values in Duroc-ssGBLUP scenarios might be due to the used scaling factors ($\tau = 1$ and $\omega = 1$) to combine G^{-1} and A_{22}^{-1} matrices. Alvarenga et al. (2020) indicated that optimization of scaling factors might be helpful for obtaining more unbiased values using ssGBLUP method (Alvarenga et al., 2020). Another possible reason for these underestimated values in Duroc-ssGBLUP scenarios can be due to the preferential treatment to select elite pigs for genotyping in breeding companies. The effect of preferential treatment on level bias was highlighted by former studies in dairy cattle breeding industry (Wiggans et al., 2011; Nordbø et al., 2019). However, it may need more

investigation for future studies on swine. In contrast to the GBLUP prediction methods, some genomic prediction methods based on Bayesian approaches such as BayesC π assume that the genetic variation of a trait is explained by a small number of SNPs (Habier et al., 2007; Hayes et al., 2009b; de los Campos et al., 2013). Therefore, the superiority of BayesC π in some scenarios might be due to the effective role of SNPs with major effects for Duroc-BFT, Yorkshire-ADG, and Landrace-LMD. Moreover, BayesC π was the least biased method for five scenarios including Duroc-LMD, Landrace-BFT, Landrace-ADG, Yorkshire-BFT, and Yorkshire-ADG. Based on the superiority of BayesC π in predictive ability for three scenarios and unbiasedness for five scenarios, it could be concluded that BayesC π could provide a more dynamic and realistic assumption (Legarra et al., 2011b) for genetic architecture of aforementioned traits, although its long computational time might be a nonpersuasive factor for implementation to the pig industry.

CONCLUSION

The accuracies of traditional BLUP, GBLUP, BayesC, ssGBLUP, and BayesC π methods in a moderate genotyped size of Canadian swine populations were evaluated to compare their predictive abilities. In most scenarios, ssGBLUP and BayesC π methods demonstrated the highest prediction accuracies and unbiasedness, respectively, although there were no significant differences ($P > 0.05$) among prediction accuracies obtained from these genomic methods in each scenario. These results can be beneficial for implementing the suggested genomic prediction methods for improvement of BFT, LMD, and ADG in swine breeding companies.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data are not publicly available due to privacy. Requests to access the datasets should be directed to YM.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because the animals in this study were raised in commercial farms and only data was provided to this study without any changes to the routine production and thus no animal ethics were needed. Animals were also cared for according to the Canadian Council on Animal Care guidelines.

AUTHOR CONTRIBUTIONS

YM, BS, and MJ conceived and designed this experiment. MJ prepared the imputed genotype files. SS wrote the first draft of the manuscript and analyzed the data under supervision of YM, BS, MS, and MJ. SS done visualizations. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Canadian swine breeding companies of AlphaGene and Alliance Genetics Canada that provided genotypic and phenotypic data for this study. We would also like to thank Laurence Maignel from the Canadian Centre for Swine Improvement for providing the pedigree and performance records.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.665344/full#supplementary-material>

Supplementary Figure 1 | Heatmap plot of (A) scaled genomic and (B) pedigree relationship matrices for Duroc group. The means of genomic relationships among reference-validation and validation-validation groups were -0.0047 (range = -0.3065 to 0.6399) and 0.0348 (range = -0.2140 to 0.8164), respectively. The means of pedigree relationships among reference-validation and validation-validation groups were 0.3934 (range = 0.0877 to 0.9339) and 0.4676 (range = 0.4229 to 0.9219), respectively.

Supplementary Figure 2 | Heatmap plot of (A) scaled genomic and (B) pedigree relationship matrices for Landrace group. The means of genomic relationships

among reference-validation and validation-validation groups were -0.0037 (range = -0.2064 to 0.6296) and 0.0364 (range = -0.1740 to 0.6711), respectively. The means of pedigree relationships among reference-validation and validation-validation groups were 0.3026 (range = 0.0738 to 0.8396) and 0.3298 (range = 0.1944 to 0.8369), respectively.

Supplementary Figure 3 | Heatmap plot of (A) scaled genomic and (B) pedigree relationship matrices for Yorkshire group. The means of genomic relationships among reference-validation and validation-validation groups were -0.0041 (range = -0.2601 to 0.6921) and 0.0468 (range = -0.1707 to 0.7612), respectively. The means of pedigree relationships among reference-validation and validation-validation groups were 0.2100 (range = 0.0353 to 0.7457) and 0.2553 (range = 0.1160 to 0.7445), respectively.

Supplementary Table 1 | The average number of phenotyped boars and gilts per litter for Duroc, Landrace, and Yorkshire groups.

Supplementary Table 2 | Estimated variance components (standard errors) to use in best linear unbiased prediction (BLUP), single-step genomic BLUP (ssGBLUP), and genomic BLUP (GBLUP).

Supplementary Table 3 | Computational time (h) of genomic BLUP (GBLUP), single-step genomic BLUP (ssGBLUP), BayesC, and BayesC π methods for back-fat thickness (BFT), average daily gain (ADG), and loin muscle depth (LMD) for Duroc, Landrace, and Yorkshire breeds.

Supplementary Table 4 | The results of Tukey tests using Model 5 ($y_{ij} = \mu + M_i + B_j + e_{ij}$). y is the accuracy of prediction for the trait, μ is the overall mean, M is the fixed effect of genomic evaluation method, B is the fixed effect of breed, and e is the random residual effect.

REFERENCES

- Aliakbari, A., Delpuech, E., Labrune, Y., Riquet, J., and Gilbert, H. (2020). The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. *Genet. Sel. Evol.* 52:57.
- Alvarenga, A. B., Veroneze, R., Oliveira, H. R., Marques, D. B., Lopes, P. S., Silva, F. F., et al. (2020). Comparing alternative single-step GBLUP approaches and training population designs for genomic evaluation of crossbred animals. *Front. Genet.* 11:263. doi: 10.3389/fgene.2020.00263
- Amaya Martínez, A., Martínez Sarmiento, R., and Cerón-Muñoz, M. (2020). Genetic evaluations in cattle using the single-step genomic best linear unbiased predictor. *Cienc. Tecnol. Agropecuaria* 21, 19–31.
- Badke, Y. M., Bates, R. O., Ernst, C. W., Fix, J., and Steibel, J. P. (2014). Accuracy of estimation of genomic breeding values in pigs using low density genotypes and imputation. *G3* 4, 623–631. doi: 10.1534/g3.114.010504
- Brown, A., Ojango, J., Gibson, J., Coffey, M., Okeyo, M., and Mrode, R. (2016). Genomic selection in a crossbred cattle population using data from the dairy genetics East Africa project. *J. Dairy Sci.* 99, 7308–7312. doi: 10.3168/jds.2016-11083
- Cardoso, F., Sollero, B., Gulas-Gomes, C., Comin, H., Roso, V., Higa, R., et al. (2014). Accuracy of Genomic Prediction for Tick Resistance in Braford and Hereford Cattle Embrapa Pecuaría Sul-Artigo em anais de congresso (ALICE), Vol. 10, Vancouver, BC: World Congress On Genetics Applied To Livestock Production, 2014.
- Christensen, O. F., and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2.
- Costa, R. B., Irano, N., Diaz, I. D. P. S., Takada, L., da Costa Hermisdorff, I., Carvalheiro, R., et al. (2019). Prediction of genomic breeding values for reproductive traits in nellore heifers. *Theriogenology* 125, 12–17. doi: 10.1016/j.theriogenology.2018.10.014
- de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608.t002
- Dekkers, J. C., Mathur, P. K., and Knol, E. F. (2010). “Genetic of the Pig Improvement,” in *The Genetics of the Pig*, eds A. Ruvinski and M. F. Rothschild 390. (Wallingford, UK: CAB International). doi: 10.1079/9781845937560.0390
- Esfandiyari, H., Bijma, P., Henryon, M., Christensen, O. F., and Sørensen, A. C. (2016). Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. *Genet. Sel. Evol.* 48, 1–9.
- Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* 41:55.
- Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0
- Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15:30. doi: 10.1186/1471-2156-15-30
- Guo, P., Zhu, B., Xu, L., Niu, H., Wang, Z., Guan, L., et al. (2017). Genomic prediction with parallel computing for slaughter traits in Chinese Simmental beef cattle using high-density genotypes. *PLoS One* 12:e0179885. doi: 10.1371/journal.pone.0179885
- Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190
- Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186
- Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819
- Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009a). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41, 1–9.
- Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/s0016672308009981
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430
- Hong, J. K., Kim, Y. S., Cho, K. H., Lee, D. H., Min, Y. J., and Cho, E. S. (2019). Application of single-step genomic evaluation using social genetic effect model for growth in pig. *Asian-Australas. J. Anim. Sci.* 32, 1836–1843. doi: 10.5713/ajas.19.0182

- Iheshiulor, O. O., Woolliams, J. A., Svendsen, M., Solberg, T., and Meuwissen, T. H. (2017). Simultaneous fitting of genomic-BLUP and Bayes-C components in a genomic prediction model. *Genet. Sel. Evol.* 49:63.
- Jafarikia, M., Sullivan, B., McSween, P., and Maignel, L. (2018). Validation of genomic evaluation on some economically important traits of Canadian purebred pigs. *J. Anim. Sci.* 96, 113–113. doi: 10.1093/jas/sky404.248
- Jiao, S., Maltecca, C., Gray, K., and Cassady, J. (2014). Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: I. Genetic parameter estimation and accuracy of genomic prediction. *J. Anim. Sci.* 92, 2377–2386. doi: 10.2527/jas.2013-7338
- Júnior, G. A. F., Rosa, G. J., Valente, B. D., Carvalheiro, R., Baldi, F., Garcia, D. A., et al. (2016). Genomic prediction of breeding values for carcass traits in Nelore cattle. *Genet. Sel. Evol.* 48:7.
- Karaman, E., Cheng, H., Firat, M. Z., Garrick, D. J., and Fernando, R. L. (2016). An upper bound for accuracy of prediction using GBLUP. *PLoS One* 11:e0161054. doi: 10.1371/journal.pone.0161054
- Karaman, E., Lund, M. S., Anche, M. T., Janss, L., and Su, G. (2018). Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. *G3* 8, 3549–3558. doi: 10.1534/g3.118.200673
- Karimi, K., Sargolzaei, M., Plastow, G. S., Wang, Z., and Miari, Y. (2019). Opportunities for genomic selection in American mink: a simulation study. *PLoS One* 14:e0213873. doi: 10.1371/journal.pone.0213873
- Kizilkaya, K., Fernando, R., and Garrick, D. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88, 544–551. doi: 10.2527/jas.2009-2064
- Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92, 4656–4663. doi: 10.3168/jds.2009-2061
- Legarra, A., Ricard, A., and Filangi, O. (2011a). *GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCπ)*. Paris, France: INRA.
- Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011b). Improved Lasso for genomic selection. *Genet. Res.* 93, 77–87. doi: 10.1017/s0016672310000534
- Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 9:237. doi: 10.3389/fgene.2018.00237
- Lopez, B. I., Seo, K. S., Viterbo, V., and Song, C. W. (2019). Estimation of genetic parameters and accuracy of genomic prediction for production traits in Duroc pigs. *Czech J. Anim. Sci.* 64, 160–165. doi: 10.17221/150/2018-cjas
- Lourenco, D., Misztal, I., Tsuruta, S., Aguilar, I., Ezra, E., Ron, M., et al. (2014). Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *J. Dairy Sci.* 97, 1742–1752. doi: 10.3168/jds.2013-6916
- Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183, 1119–1126. doi: 10.1534/genetics.109.107391
- MacLeod, I., Bowman, P., Vander Jagt, C., Haile-Mariam, M., Kemper, K., Chamberlain, A., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6
- Mehrban, H., Lee, D. H., Moradi, M. H., IlCho, C., Naserkheil, M., and Ibáñez-Escriche, N. (2017). Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. *Genet. Sel. Evol.* 49:1.
- Miari, Y., Plastow, G., and Wang, Z. (2015). Genomic selection, a new era for pork quality Improvement. *Springer Sci. Rev.* 3, 27–37. doi: 10.1007/s40362-015-0029-3
- Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., and Vitezica, Z. (2014). *Manual for BLUPF90 Family of Programs*. Athens: University of Georgia.
- Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. (2002). “BLUPF90 and related programs (BGF90),” in *Proceedings of the 7th world Congress on Genetics Applied to Livestock Production*, Montpellier, 743–744.
- Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., et al. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* 8:12309.
- Mrode, R., Ojango, J. M., Okeyo, A., and Mwacharo, J. M. (2018). Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. *Front. Genet.* 9:694. doi: 10.3389/fgene.2018.00694
- Neves, H. H., Carvalheiro, R., O'Brien, A. M. P., Utsunomiya, Y. T., Do Carmo, A. S., Schenkel, F. S., et al. (2014). Accuracy of genomic predictions in Bos indicus (Nelore) cattle. *Genet. Sel. Evol.* 46:17. doi: 10.1186/1297-9686-46-17
- Nordbø, Ø., Gjuvsland, A. B., Eikje, L. S., and Meuwissen, T. (2019). Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. *Genet. Sel. Evol.* 51, 1–8.
- Olfert, E. D., Cross, B. M., and McWilliam, A. A. (1993). *Guide to the Care and Use of Experimental Animals*. Ottawa, ONT: Canadian Council on Animal Care.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing
- Samore, A. B., and Fontanesi, L. (2016). Genomic selection in pigs: state of the art and perspectives. *Ital. J. Anim. Sci.* 15, 211–232. doi: 10.1080/1828051x.2016.1172034
- Sargolzaei, M. (2014). *SNP1101 User's guide, Version*.
- Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15:478. doi: 10.1186/1471-2164-15-478
- Sevillano, C. A., Vandenplas, J., Bastiaansen, J. W., Bergsma, R., and Calus, M. P. (2017). Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. *Genet. Sel. Evol.* 49:75.
- Silva, R., Fragomeni, B., Lourenco, D., Magalhães, A., Irano, N., Carvalheiro, R., et al. (2016). Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. *J. Anim. Sci.* 94, 3613–3623. doi: 10.2527/jas.2016-0401
- Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019a). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. *Genet. Sel. Evol.* 51:58.
- Song, H., Zhang, J., Jiang, Y., Gao, H., Tang, S., Mi, S., et al. (2017). Genomic prediction for growth and reproduction traits in pig using an admixed reference population. *J. Anim. Sci.* 95, 3415–3424. doi: 10.2527/jas.2017.1656
- Song, H., Zhang, J., Zhang, Q., and Ding, X. (2019b). Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig. *Front. Genet.* 9:730. doi: 10.3389/fgene.2018.00730
- Song, H., Zhang, Q., and Ding, X. (2020). The superiority of multi-trait models with genotype-by-environment interactions in a limited number of environments for genomic prediction in pigs. *J. Anim. Sci. Biotechnol.* 11, 1–13.
- Thekkoot, D., Kemp, R., Boddicker, N., Mwansa, P., and Plastow, J. D. (2018). “Single step genomic prediction accuracies for growth and reproductive traits in Yorkshire pigs with genotypes from two different SNP panels,” in *Proceedings of the World Congress on Genetics Applied to Livestock Production*, Oakville, MB, 374.
- Tsuruta, S., Misztal, I., and Lawlor, T. (2013). Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *J. Dairy Sci.* 96, 3332–3335. doi: 10.3168/jds.2012-6272
- VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., and Schenkel, F. (2008). Reliability of genomic predictions for North American dairy bulls. *J. Dairy Sci.* 91, 305.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Wiggans, G., Cooper, T., VanRaden, P., and Cole, J. (2011). Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. *J. Dairy Sci.* 94, 6188–6193. doi: 10.3168/jds.2011-4481
- Wolc, A., Arango, J., Settari, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., et al. (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet. Sel. Evol.* 43:23.

- Zhang, C., Kemp, R. A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., et al. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet. Sel. Evol.* 50:14.
- Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* 10:189. doi: 10.3389/fgene.2019.00189
- Zhang, Z., Zhang, Z., Oyelami, F., Sun, H., Xu, Z., Ma, P., et al. (2020). Identification of genes related to intramuscular fat independent of backfat thickness in Duroc pigs using single-step genome-wide association. *Anim. Genet.* 52, 108–113. doi: 10.1111/age.13012
- Zhou, L., Wang, Y., Yu, J., Yang, H., Kang, H., Zhang, S., et al. (2019). Improving genomic prediction for two Yorkshire populations with a limited size using the single-step method. *Anim. Genet.* 50, 391–394. doi: 10.1111/age.12806

Conflict of Interest: MS is employed at Select Sires Inc. This organization did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript and only provided financial support in the form of MS's salary.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Salek Ardestani, Jafarikia, Sargolzaei, Sullivan and Miar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Omics Multi-Layers Networks Provide Novel Mechanistic and Functional Insights Into Fat Storage and Lipid Metabolism in Poultry

Farzad Ghafouri^{1†}, Abolfazl Bahrami^{1,2*†}, Mostafa Sadeghi^{1*†}, Seyed Reza Miraei-Ashtiani¹, Maryam Bakherad^{3*}, Herman W. Barkema⁴ and Samantha Larose⁵

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Guirong Sun,
Henan Agricultural University, China
Paolo Zambonelli,
University of Bologna, Italy

*Correspondence:

Abolfazl Bahrami
A.Bahrami@ut.ac.ir
Mostafa Sadeghi
Sadeghimos@ut.ac.ir
Maryam Bakherad
Bakherad@khu.ac.ir

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 26 December 2020

Accepted: 04 June 2021

Published: 07 July 2021

Citation:

Ghafouri F, Bahrami A,
Sadeghi M, Miraei-Ashtiani SR,
Bakherad M, Barkema HW and
Larose S (2021) Omics Multi-Layers
Networks Provide Novel Mechanistic
and Functional Insights Into Fat
Storage and Lipid Metabolism in
Poultry. *Front. Genet.* 12:646297.
doi: 10.3389/fgene.2021.646297

¹ Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran, ² Biomedical Center for Systems Biology Science Munich, Ludwig-Maximilians-University, Munich, Germany, ³ Department of Cell and Molecular Biology, Faculty of Biological Sciences, Kharazmi University, Tehran, Iran, ⁴ Department of Production Animal Health, University of Calgary, Calgary, AB, Canada, ⁵ One Health at UCalgary, University of Calgary, Calgary, AB, Canada

Fatty acid metabolism in poultry has a major impact on production and disease resistance traits. According to the high rate of interactions between lipid metabolism and its regulating properties, a holistic approach is necessary. To study omics multilayers of adipose tissue and identification of genes and miRNAs involved in fat metabolism, storage and endocrine signaling pathways in two groups of broiler chickens with high and low abdominal fat, as well as high-throughput techniques, were used. The gene–miRNA interacting bipartite and metabolic-signaling networks were reconstructed using their interactions. In the analysis of microarray and RNA-Seq data, 1,835 genes were detected by comparing the identified genes with significant expression differences ($p_{\text{adjust}} < 0.01$, fold change ≥ 2 and ≤ -2). Then, by comparing between different data sets, 34 genes and 19 miRNAs were detected as common and main nodes. A literature mining approach was used, and seven genes were identified and added to the common gene set. Module finding revealed three important and functional modules, which were involved in the peroxisome proliferator-activated receptor (PPAR) signaling pathway, biosynthesis of unsaturated fatty acids, Alzheimer's disease metabolic pathway, adipocytokine, insulin, PI3K–Akt, mTOR, and AMPK signaling pathway. This approach revealed a new insight to better understand the biological processes associated with adipose tissue.

Keywords: lipid metabolism, transcriptome, systems biology, interactive bipartite network, omics multilayer

INTRODUCTION

Total carcass fat of broilers varies depending on sex, poultry age, nutrition, and genetic factors (about 12%) (Sakomura et al., 2005). The predominant fats stored in a broiler carcass include two kinds of subcutaneous fat and ventricular fat (approximately 18 to 22% of carcass fat) stored in the ventricular area (Crespo and Esteve-Garcia, 2001). For humans, as the foremost consumer

of poultry meat, over-fat storage in skeletal muscle is associated with metabolic diseases such as type 2 diabetes and cardiovascular disease and subsequently will lead to the risk of a heart attack. Fat production in poultry is a high-inheritance polygenic trait regulated by various behavioral, environmental, and hormonal factors (Le Mignon et al., 2009).

Many studies have identified genes related to storage lipids in broilers (Lagarrigue et al., 2006; Pinto et al., 2010; Nones et al., 2012). On the other hand, the integration of high-throughput genomic DNA and RNA sequencing leads to the identification of genomic regions that control traits at the whole genome scale (Cesar et al., 2018). Some studies of two poultry groups, obese [high fat (HF)] and lean [low fat (LF)], indicated genes associated with lipogenic pathways (Ji et al., 2012, 2014). By comparing expressed genes, numerous identified genes were related to endocrine, hemostatic, lipolytic, and lipid transduction (Resnyk et al., 2013).

In addition to identifying genes and pathways associated with lipid metabolism, a holistic approach for gene expression should be examined. MicroRNAs are regulatory molecules with a length of 19–25 nucleotides (Bartel, 2004). Mature microRNAs lead to decomposition or inhibit translation by complete or partial coupling to target mRNAs (usually paired with the 3'UTR region) (Iorio et al., 2011). We have witnessed the emergence of various areas in biology. One of these areas is the application of bioinformatics and systems biology and integrated multi-omics data. In major systems biology, researchers have attempted to identify the cellular system, formulate cell behaviors, and then design a cell model by combining genomic, transcriptomic, proteomic, and metabolomic layers (Cole et al., 2013). In this regard, interactive bi-partite networks of gene–miRNA are used in several studies to discover functional modules (Huang et al., 2006; Bahrami et al., 2017a,b).

However, identification of upstream and downstream genes, reconstruction of networks, bipartite interaction network of gene–miRNA, and metabolic-signaling networks involved in metabolism and adipose storage (particularly abdominal fat using high-throughput data in broilers) have not been reported. Fat storage in broilers is considered to be an important economic trait concerning high growth rate. Based on previous studies of fat metabolism in the body and signaling pathways related to fat storage and transmission in laboratory species, it was assumed that the two broiler groups of high-abdominal fat and low-abdominal fat have gene expression differences in metabolism and fat storage.

Accordingly, this study aims to use an integration of RNA-Seq and microarray data approach to identify and classify candidate genes and miRNAs involved in lipolysis and lipogenesis. In addition to the comprehensive survey of lipid metabolism, this study will focus on (1) reconstruction of the interactive bi-partite network of gene–miRNA (bi-partite networks are a particular class of complex networks, whose nodes are divided into two sets of genes and miRNA), (2) identification of functionally relevant modules (each of a set of genes or independent genes that can be used to construct a more complex structure), and (3) reconstruction of the metabolic-signaling network associated with the process of metabolism and fat storage in broilers.

MATERIALS AND METHODS

Figure 1 and **Supplementary Table 1** show the simple overall workflow for analyzing and finding functionally relevant modules with HF and LF storage in the broilers.

Poultry's Tissue Preparation

The 18 chickens used in this study were divergently selected based on the amount of carcass fat percentage at 42 days of age (slaughtering time). Chickens were bred and raised at the animal farm of Tehran University, Iran. Nine chickens were in the HF group (>27% fat storage) and nine chickens in the LF group (<10% fat storage). Each group was divided into three subgroups with three chickens in each group. To eliminate other environmental effects and sampling error, abdominal adipose tissue samples of three chickens in each subgroup were pooled. Therefore, we had three samples for HF and three samples for LF chickens, separately. In this regard, both groups were placed together and raised in floor pens (4.4 m × 3.9 m). Abdominal adipose tissue samples were immediately pooled (before RNA extraction), snap-frozen in liquid nitrogen, and stored at −80°C until further processing for RNA analysis.

RNA Extraction

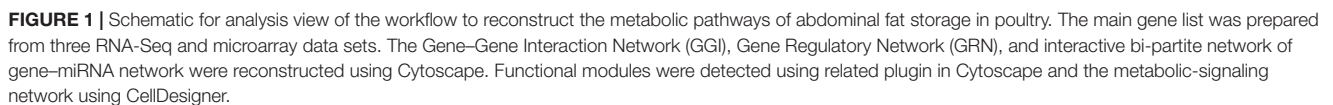
Abdominal fat aliquots from six chickens (three HF and three LF per age at 42 weeks) were homogenized, and total cellular RNA was extracted using guanidine thiocyanate and CsCl gradient purification, followed by DNase I treatment. The quality of RNA was determined with an RNA 6000 Nano Assay kit and the Model 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States). All samples used for RNA analyses had an RNA integrity number (RIN) greater than 9.0.

MiRNA-Seq Library Preparation and Sequencing

About 1 µg of total RNA from each sample was used to construct a small RNA library using the TruSeq Small RNA Sample Preparation kit (Illumina, San Diego, CA, United States). The kit was used according to the manufacturer's instructions, which included ligating adapters to 3' and 5' end of the RNA molecules, reversely transcribing and amplifying libraries, purifying cDNA, and checking and normalizing libraries. All libraries were sequenced at Génome Québec (Montréal, Canada) using the HiSeq 2000 system (Illumina, San Diego, CA, United States) to generate 50-bp single reads.

Data Mining

In the biological system and the reconstruction of biological networks, namely, gene regulation, interactions, protein–protein interaction (PPI), and metabolic networks, the first step is to collect and evaluate the available data. In this regard, data from this study were obtained by investigating and reviewing related articles and collecting microarray and RNA-Seq data from different databases, by searching the Gene Expression Omnibus



¹www.ncbi.nlm.nih.gov/geo
²www.Ebi.ac.Uk/arrayexpress

Microarray data were pre-processed in software R, using package Lumi (Du et al., 2008) and Affy (Gautier et al., 2004). The processed data were then evaluated using packages Limma (Ritchie et al., 2015), GEOquary (Davis and Meltzer, 2007), and Biobase (Huber et al., 2015) (versions and parameters were used for analysis of microarray and RNA-Seq presented in

TABLE 1 | GEO accession numbers for RNA-Seq and microarray data sets.

No.	No. sample(s)	GSE	Platforms	Data type	Contributor(s)
1	16	GSE49121	GPL16133 (Illumina HiSeq 2000)	RNA-Seq	Resnyk et al., 2017
2	24	GSE42980	GPL16133 (Illumina HiSeq 2000)	RNA-Seq	Resnyk et al., 2015
3	24	GSE37585	GPL1731 (DEL-MAR 14K Integrated Systems)	Microarray	Resnyk et al., 2013
4	24	GSE8812	GPL1731 (DEL-MAR 14K Integrated Systems)	Microarray	Resnyk et al., 2015
5	24	GSE45825	GPL1731 (DEL-MAR 14K Integrated Systems)	Microarray	Resnyk et al., 2017
6	8	GSE10052	GPL1731 (DEL-MAR 14K Integrated Systems)	Microarray	Byerly et al., 2010
7	28	GSE3867	GPL3265 (Chicken cDNA DDMET 1700 array version 1.0)	Microarray	Bourneuf et al., 2006

GEO, *gene expression omnibus*.

Supplementary Table 1). Among the number of identified genes, the genes that were common in terms of five accession numbers (related to microarray data sets) were identified; and the gene list was considered as gene set 1 (**Supplementary Table 2**).

RNA-Seq Data and Statistical Analyses

Various programs were used to analyze the RNA-Seq data related to the accession numbers. First, FastQC quality control software (Andrews, 2010) was used to control the quality of existing data. Sequences were trimmed for quality using Trimmomatic software (Bolger et al., 2014). Boxplot graphing of pre- and post-trimming reads confirmed the absence of outlier samples based on read count. After trimming, reads were mapped to the chicken genome assembly GRCg6a³ using Tophat (version 1.3.3) (Kim et al., 2013), followed by assembly and quantitation using CuffDiff software (v2.2.1.6) (Trapnell et al., 2010). The fragments per kilobase of exon per million fragments mapped (FPKM) threshold for detection of a gene was set at FPKM > 0.5. The resulting gtf files differential expression was assessed using Cuffdiff. The two-sided *p*-value was corrected using the false discovery rate (FDR), which accounts for multiple testing procedures. Genes with an FDR-adjusted *p*-value ($p \leq 0.05$) and fold change ≥ 2 or ≤ -2 were considered to be differentially expressed (DE) transcripts. The genes that were common in terms of accession numbers (related to RNA-Seq data sets) were identified and considered as gene set 2 (**Supplementary Table 3**).

Functional Gene Set Annotation and Enrichment

Gene ontology (GO) analysis, canonical pathway, and network identification were performed using Database for Annotation, Visualization and Integrated Discovery (DAVID)⁴,

Bioinformatics Resources 6.8 with (Huang et al., 2009) the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, g: profiler⁵ (Raudvere et al., 2019), GeneCards⁶, and PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Mi et al., 2012).

Main Gene List

Genes with significant differences related to microarray and RNA-Seq data were examined and listed as gene set 1 and 2, respectively. Finally, genes that were common in these two gene sets were chosen as the main gene list.

Identification of miRNAs and Target Genes

Accession number GSE122224, which is related to miRNA in chicken and associated with lipid metabolism, was analyzed. The potentially targeted genes were predicted using miRWalk 3.0 (Sticht et al., 2018). The platform integrates information from different miRNA-target databases, including validated information and prediction data sets: MiRWalk (Dweep et al., 2014), miRDB⁷, miRMap (Vejnar and Zdobnov, 2012), miRNAmap (Hsu et al., 2008), miRanda⁸, miRBridge (Tsang et al., 2010), PICTAR2⁹, Targetscan (Grimson et al., 2007), PITA¹⁰, and RNA22 (Loher and Rigoutsos, 2012). The target genes that were predicted by at least five mentioned tools were chosen and submitted to DAVID, KEGG (the potential KEGG), Reactome pathways, and PANTHER databases for the enrichment target genes of each miRNA.

⁵<https://biit.cs.ut.ee/gprofiler/gost>

⁶<https://www.genecards.org/>

⁷<http://mirdb.org/>

⁸<https://mirnablog.com/microrna-target-prediction-tools/>

⁹<https://pictar.mdc-berlin.de/>

¹⁰https://tools4mirs.org/software/target_prediction/pita/

³http://ftp.ensembl.org/pub/release-102/fasta/gallus_gallus/dna/

⁴<https://david.ncifcrf.gov/>

Reconstruction of Omics Multilayered Networks

The miRNA–gene network was reconstructed based on the candidate genes, and the molecular interactions were documented in related papers and online interaction databases. PPI data were abstracted from the Biomolecular Interaction Network Database (BIND¹¹), Database of Interacting Proteins (DIP¹²), Biological General Repository for Interaction Datasets (BioGRID¹³), and Protein–Protein Interactions Database (MIPS¹⁴). In addition, pathway data were obtained from searches in pathway databases, such as STRING¹⁵ (Szklarczyk et al., 2018) and GeneMania databases¹⁶ (Warde-Farley et al., 2010). Each gene and miRNA was entered into the database, and resulting interactions were imported to the networks using Cytoscape 3.7.2 (National Institute of General Medical Sciences, Bethesda Softworks, Rockville, MD, United States) (Shannon et al., 2003). Genes and miRNAs in generated networks are represented as nodes, and the interactions between these nodes as edges. The metabolic-signaling pathways involved in the lipid metabolism and storage were reconstructed by different databases and Cell Designer version 4.4.2 (Funahashi et al., 2008).

Modules and Hub Node Detection

For finding sub-graphs and hub nodes (nodes with a high connectedness coefficient), MCODE, one of the Cytoscape plugins, was used. MCODE finds clusters (highly interconnected regions) in a network. Clusters mean different things in different types of networks. For instance, clusters in a PPI network are often protein complexes and parts of pathways, while clusters in a protein similarity network represent protein families (Bader and Hogue, 2003). MCODE effectively finds densely connected regions of a molecular interaction network, many of which correspond to known molecular complexes, based solely on connectivity data. Given that this approach to analyzing protein interaction networks performs well using minimal qualitative information implies that large amounts of available knowledge are buried in large protein interaction networks. More accurate data mining algorithms and systems models could be constructed to understand and predict interactions, complexes, and pathways by taking into account more existing biological knowledge. Structured molecular interaction data resources such as BIND will be vital in creating these resources (Bader et al., 2003).

RESULTS

Differentially expressed genes were defined as those having a significant adjusted *p*-value (<0.01), fold change (≥ 2 , ≤ -2), and FDR (≤ 0.05). Statistical analysis of the time-course microarray studies provided 1,451 significant genes from five data sets: the

first data set (GSE37585: 612 DE genes), the second data set (GSE8812: 107 DE genes), the third data set (GSE45825: 582 DE genes), the fourth data set (GSE10052: 104 DE genes), and the fifth data set (GSE3867: 46 DE genes). In the data analysis of RNA-Seq, 1,867 genes were identified; and then 314 and 70 genes were detected after considering the threshold ($p_{\text{adjust}} < 0.01$ and fold change > 2) of expression change in accession numbers GSE49121 and GSE42980, respectively.

Identification of miRNAs

Overall, 34 miRNAs were identified in data analysis of microRNAs differential expression, of which 19 upregulated miRNA and 15 downregulated genes were detected by considering the threshold ($\text{LogFC} < -2$, $\text{LogFC} > 2$, and $p_{\text{adjust}} < 0.01$) for DE in the deposited accession number (GSE122224) (Table 2).

Identification of Common Genes Available in Gene Sets 1 and 2

Thirty-four genes were common in two gene sets 1 and 2 relating to microarray and RNA-Seq data sets, respectively. In this regard, 16 and 18 genes were associated with lipogenesis and lipolysis processes, respectively (Table 2). *THBS1* and *INSIG2* genes in the gene set were associated with the lipogenesis process; and *COLEC12*, *HMGCR*, *APP*, and *IRS1* genes were associated with the lipolysis process, which was closely suppressed by miRNAs (Table 2).

Main Gene List

Literature related to lipid metabolism was also reviewed to increase study accuracy and seven genes. If the genes did not exist in the list of evaluated data sets, they were selected and added to the gene list. The selected seven genes included *BACE1*, *BACE2*, *PSEN1*, *PSEN2*, *PERP*, *SIK1*, and *LOC421682* genes. The list of genes in Table 2 (41 genes) was named as the main gene list or reference genes (Supplementary Table 4).

Gene–Gene Interaction Network, Gene Ontology Terms, and Pathways

Figure 2 shows the network of the reconstruction of gene–gene interactions (gene–gene interaction (epistasis) is the effect of one gene on a disease or traits modified by another gene or several other genes), GO (describes our knowledge of the biological domain with respect to three aspects: Molecular Function, Cellular Component, and Biological Process), terms, and pathways. In this network, *APP*, *SREBF1*, *HMGCR*, *FADS2*, *SCD*, *ACAT1*, *FASN*, *HADHB*, and *EHHADH* genes had the highest interaction (connectedness) with other genes in the network.

Reconstruction of the Interactive Gene–miRNA Bipartite Network

The network contains 49 nodes (including 32 genes and 17 miRNAs) and 95 edges. The reconstructed network with.cys format was stored for further analyses (Figure 3).

¹¹<http://binddb.org>

¹²<https://www.uniprot.org/database/DB-0016>

¹³<https://thebiogrid.org/>

¹⁴<http://mips.helmholtz-muenchen.de/proj/ppi/>

¹⁵<https://string-db.org/>

¹⁶<https://genemania.org/>

TABLE 2 | Genes and miRNAs, annotation, and genes involved in lipogenesis and lipolysis.

Lipogenesis					Lipolysis				
Gene	Gene expression		MiRNA expression		Gene	Gene expression		MiRNA expression	
	Downregulation	Upregulation	Downregulation	Upregulation		Downregulation	Upregulation	Downregulation	Upregulation
<i>THBS1</i>	*		—	gga-miR-6554-5p gga-miR-6667-5p gga-miR-6562-3p	<i>COLEC12</i>	*		—	gga-miR-6554-5p gga-miR-6554-3p gga-miR-6667-5p gga-miR-3532-5p gga-miR-466
<i>ANXA7</i>	*		—	gga-miR-466	<i>RGS19</i>	*		—	—
<i>TMEM258</i>		*	—	—	<i>HTR7L</i>		*	—	—
<i>DHCR7</i>		*	—	—	<i>G6PC</i>		*	—	—
<i>FADS2</i>		*	—	—	<i>HMGCR</i>		*	gga-miR-1710	—
<i>FASN</i>		*	—	—	<i>ACAT1</i>	*		—	—
<i>INSIG2</i>		*	gga-miR-7444-5p	—	<i>ADH1C</i>	*		—	—
<i>LCAT</i>		*	—	—	<i>APP</i>	*		—	gga-miR-6554-5p
<i>MVD</i>		*	—	—	<i>EHHADH</i>	*		—	—
<i>SCD</i>		*	—	—	<i>GAMT</i>	*		—	—
<i>SREBF1</i>		*	—	—	<i>HADHB</i>	*		—	—
<i>APOA1</i>	*		—	—	<i>HSD17B4</i>	*		—	—
<i>BCO2</i>	*		—	—	<i>HSD17B6</i>	*		—	—
<i>CYP27A1</i>	*		—	—	<i>IRS1</i>	*		—	gga-miR-6554-5p gga-miR-6562-3p gga-miR-466
<i>CYP2E1</i>	*		—	—	<i>PHYH</i>	*		—	—
<i>SLC2A2</i>	*		—	—	<i>SOD3</i>	*		—	—
					<i>TP53</i>	*		—	—
					<i>UCP3</i>	*		—	—

Upregulated and downregulated abdominal fat in genetically fat compared with lean chickens.

*Gene or miRNA was up-/downregulated in the biological processes.

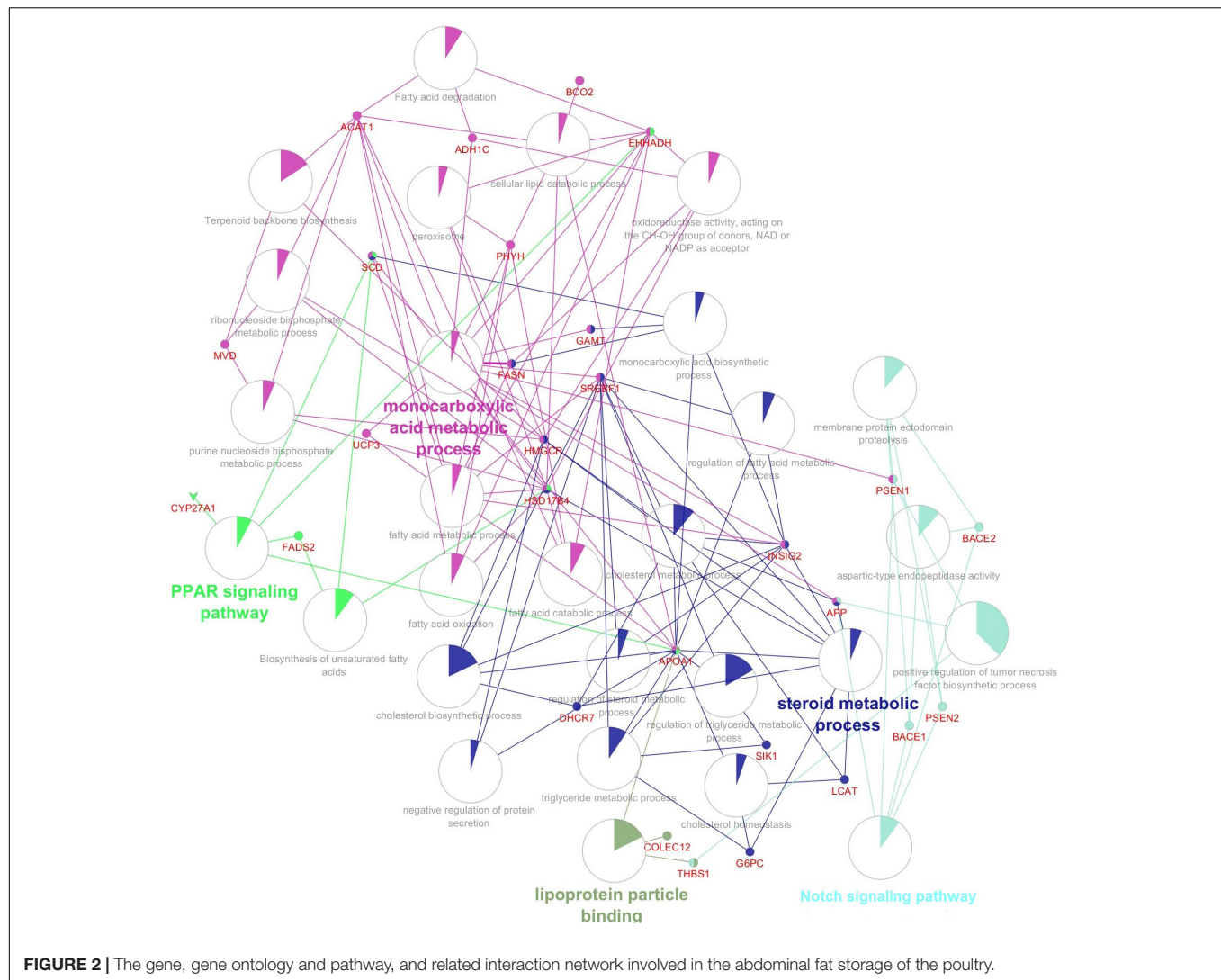


FIGURE 2 | The gene, gene ontology and pathway, and related interaction network involved in the abdominal fat storage of the poultry.

Important and Functional Network Modules or Sub-Networks

According to interactive gene-miRNA bipartite network and sub-networks or module finding analysis, three modules were identified. These modules contained 31 genes and seven miRNA as presented in **Table 3**. The table also presents important signaling pathways and cellular processes (metabolic pathways) (**Figure 2**). Module 1 contains 22 nodes (20 genes and two miRNAs) and 47 edges (**Figure 4**). Module 2 contains 10 nodes (five genes and five miRNAs) and 16 edges (**Figure 5**); and Module 3 includes six nodes (six genes) and six edges (**Figure 6**).

Reconstruction of the Metabolic-Signaling Network

Based on pathway analysis, the crucial pathways were identified and reconstructed. For this purpose, the gene lists were first input into DAVID and STRING to identify biological processes, the involvement of cellular components, molecular functions, and KEGG pathways that were significantly different between

two lines (to identify metabolic pathways and signaling). Different genes express identified Pathways such as Notch signaling pathways relating to Alzheimer's disease, peroxisome proliferator-activated receptor (PPAR), adipocytokine, insulin, PI3K-Akt, mTOR, and AMPK signaling pathways. Finally, resources were reviewed for each of the identified paths, using different databases and Cell Designer software version 4.4.2; the reconstruction is illustrated in **Figure 7**.

DISCUSSION

The prioritization of abdominal fat tissue in broiler chickens to identify genes involved in metabolism and fat storage is due to the fact that it can be as a proxy model in other species and individuals of a species due to its specific metabolic characteristics (Resnyk et al., 2015). The present study integrated different data sets in distinguished conditions to identify the most important genes involved in lipid metabolism. As a result, we detected a total of 34 common genes that played roles in

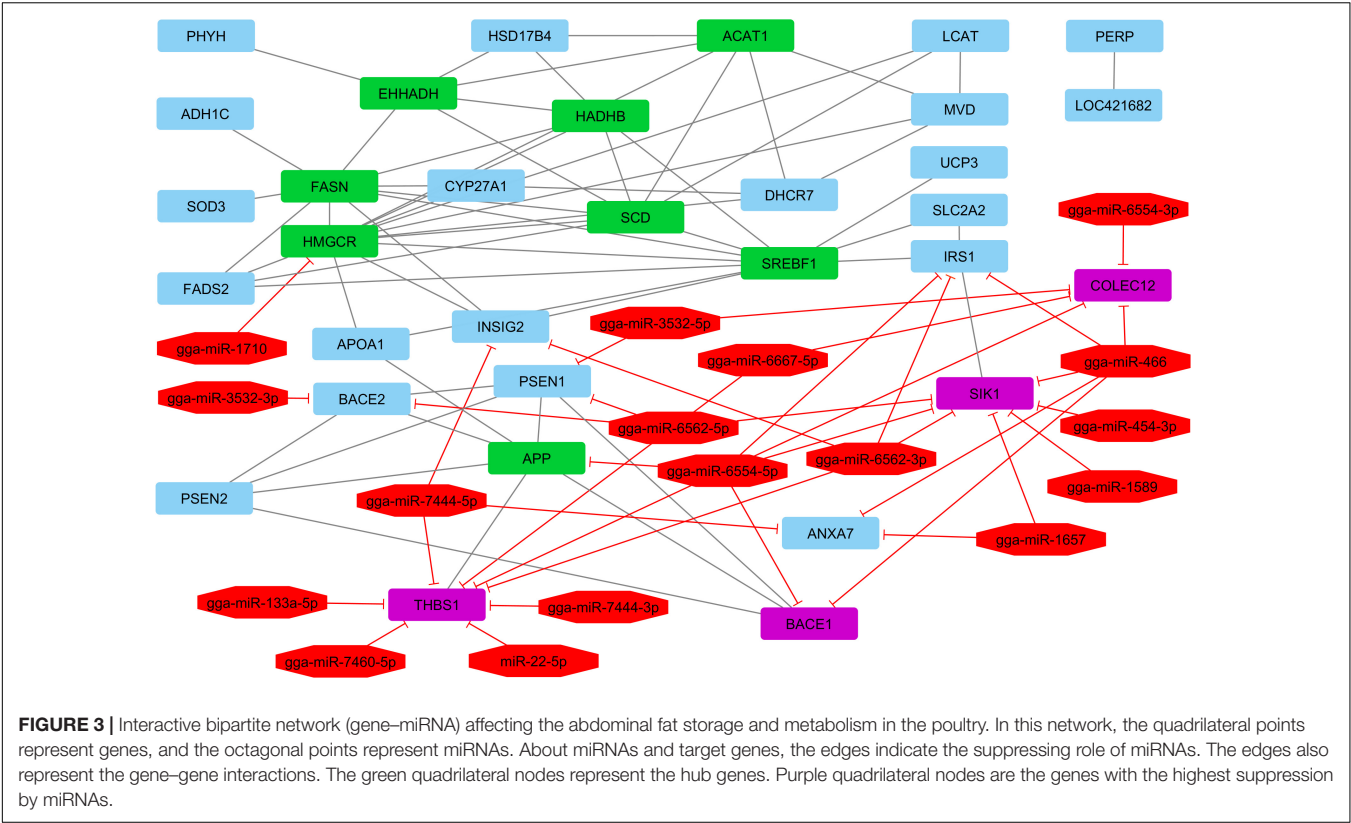
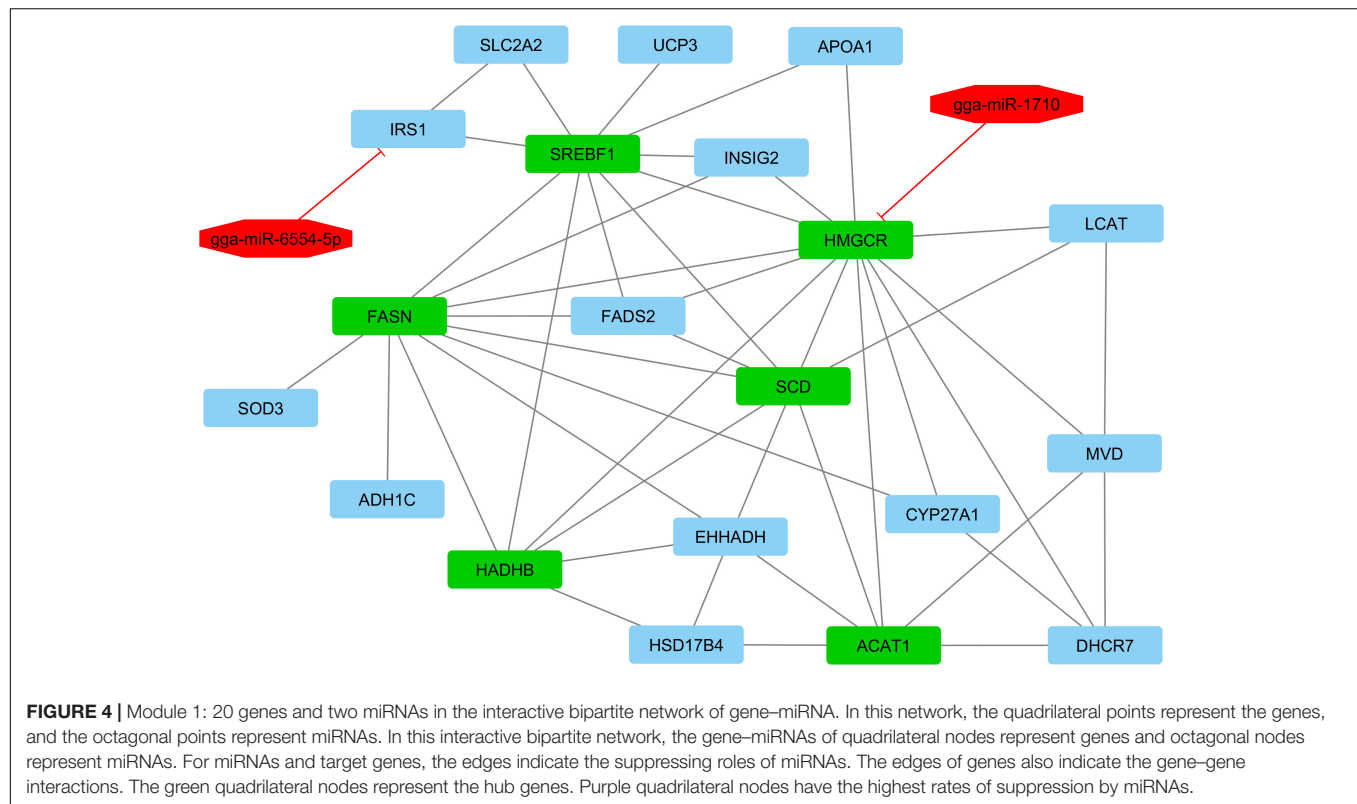


TABLE 3 | Main, modules, genes, miRNAs, signaling pathway, and phenotypic explanations in the integrated gene-miRNA bipartite network involved in fat metabolism and deposition.

Module	Genes		miRNAs		Signaling pathway	Explanation
	Downregulated	Upregulated	Downregulated	Upregulated		
Main	COLEC12, ANXA7, SOD3, SIK1, UCP3, ADH1C, SLC2A2, IRS1, BACE2, PERP, LOC421682, PHYH, CYP27A1, HADHB, THBS1, APOA1, BACE1, ACAT1, HSD17B4, EHHADH, APP, PSEN2, PSEN1	LCAT, INSIG2, FADS2, SCD, DHCR7, FASN, SREBF1, MVD, HMGCR	gga-miR-454-3p, gga-miR-7460-5p, gga-miR-133a-5p, gga-miR-1710, gga-miR-1589, gga-miR-22-5p, gga-miR-7444-3p, gga-miR-1657, gga-miR-7444-5p	gga-miR-6562-3p, gga-miR-3532-5p, gga-miR-6667-5p, gga-miR-6554-3p, gga-miR-6562-5p, gga-miR-3532-3p, gga-miR-6554-5p, gga-miR-466	PPAR/AMPK	Fatty acid metabolism Fatty acid degradation Terpenoid backbone biosynthesis Biosynthesis of unsaturated fatty acids Metabolic pathways Alzheimer disease
1	UCP3, SLC2A2, APOA1, IRS1, SOD3, ADH1C, HADHB, EHHADH, CYP27A1, HSD17B4, ACAT1	SREBF1, INSIG2, HMGCR, LCAT, FADS2, FASN, SCD, MVD, DHCR7	gga-miR-1710	gga-miR-6554-5p	PPAR/AMPK	Fatty acid metabolism Fatty acid degradation Terpenoid backbone biosynthesis Biosynthesis of unsaturated fatty acids Metabolic pathways Valine, leucine and isoleucine degradation Cholesterol metabolism Primary bile acid biosynthesis
2	BACE2, PSEN2, PSEN1, APP, BACE1	—	—	gga-miR-3532-3p, gga-miR-3532-5p, gga-miR-466, gga-miR-6554-5p, gga-miR-6562-5p	Notch signaling pathway	Alzheimer disease
3	CYP27A1, ACAT1, HSD17B4	DHCR7, MVD, LCAT	—	—	—	Primary bile acid biosynthesis Terpenoid backbone biosynthesis Metabolic pathways Cholesterol metabolism Fatty acid metabolism



the main process of synthesis route control, metabolism and fat storage, and signaling pathways of endocrine glands activated by adipokines, AMPK, and PPAR.

The lower expression of a large number of genes associated with the lipolysis indicated a reduction in decomposition of fats and then an increase in the anabolism and fat storage in broiler chickens, especially in abdominal fat tissue. On the contrary, the higher expression of a large number of genes in the gene set associated with the lipogenesis confirms the increase in metabolism and abdominal fat storage.

In most similar studies published on different species, it has been concluded that multi-omics data sets or omics multilayered networks provide a valuable resource for comparative analyses with other experimental data sets. Also, applications for data integration and analysis can be demonstrated and provide novel functional insights (Yao et al., 2015; Suravajhala et al., 2016; Hasin et al., 2017; Arora et al., 2018; Backman et al., 2019; Dao et al., 2019; Corral-Jara et al., 2020; Lee et al., 2020). In one study, an attempt has been made to investigate the effects of a transgenic supplement in mice using a molecular systems biology approach and a combination of statistical tools using high-throughput techniques. They concluded that the integration of omics data provides better molecular insight into the relationships between biological variables. Thus, such approaches can be effective in detecting mechanical, molecular, and biochemical interactions (Zhang et al., 2019).

Chickens with greater abdominal fat had hyperplasia and hypertrophy of fat cells at younger ages compared with chickens with lower abdominal fat. *SREBF1*, *SREBF2*, *SCD*, and *FASN*

were among the most important genes that play major roles in fat storage and metabolism (Resnyk et al., 2013). *THBS1*, *ANXA7*, *APOA1*, *BCO2*, *CYP27A1*, *CYP2E1*, and *SLC2A2* genes were downregulated, whereas *TMEM258*, *DHCR7*, *FADS2*, *FASN*, *INSIG2*, *LCAT*, *MVD*, *SCD*, and *SREBF1* genes were upregulated in the lipogenesis process. Additionally, *COLEC12*, *RGS19*, *ACAT1*, *ADH1C*, *APP*, *EHHADH*, *GAMT*, *HADHB*, *HSD17B4*, *HSD17B6*, *IRS1*, *PHYH*, *SOD3*, *TP53*, and *UCP3* were downregulated, whereas *HTR7L*, *G6PC*, and *HMGCR* genes were upregulated in the lipolysis process. Briefly, hub genes in this study were *APP*, *SREBF1*, *HMGCR*, *FADS2*, *SCD*, *ACAT1*, *FASN*, *HADHB*, and *EHHADH* (Figure 2).

The *APP* gene was downregulated in the lipolysis process because the *APP* gene is a cell surface receptor and an extra-membrane precursor protein that is decomposed by enzymes to form a number of peptides. Some of these peptides are secreted and can be bound to an acetyl transferase complex, *APBB1/TIP60*, to strengthen the transcription activities, while other proteins create amyloid plaques in brains of patients with Alzheimer's disease (Almkvist et al., 2019). It enhances the transcription through binding to *APBB1/KAT5* and inhibits Notch signals through interaction with *Numb*.

Sterol regulatory element-binding transcription factor 1 (*SREBF1*) gene was upregulated in the lipogenesis process because the *SREBF1* is a protein-encoding gene. Fatty liver disease is a *SREBF1* gene-related disease; and the mTOR signaling pathway is a pathway associated with *SREBF1*. Annotation of this gene includes the DNA and chromatin binding transcription

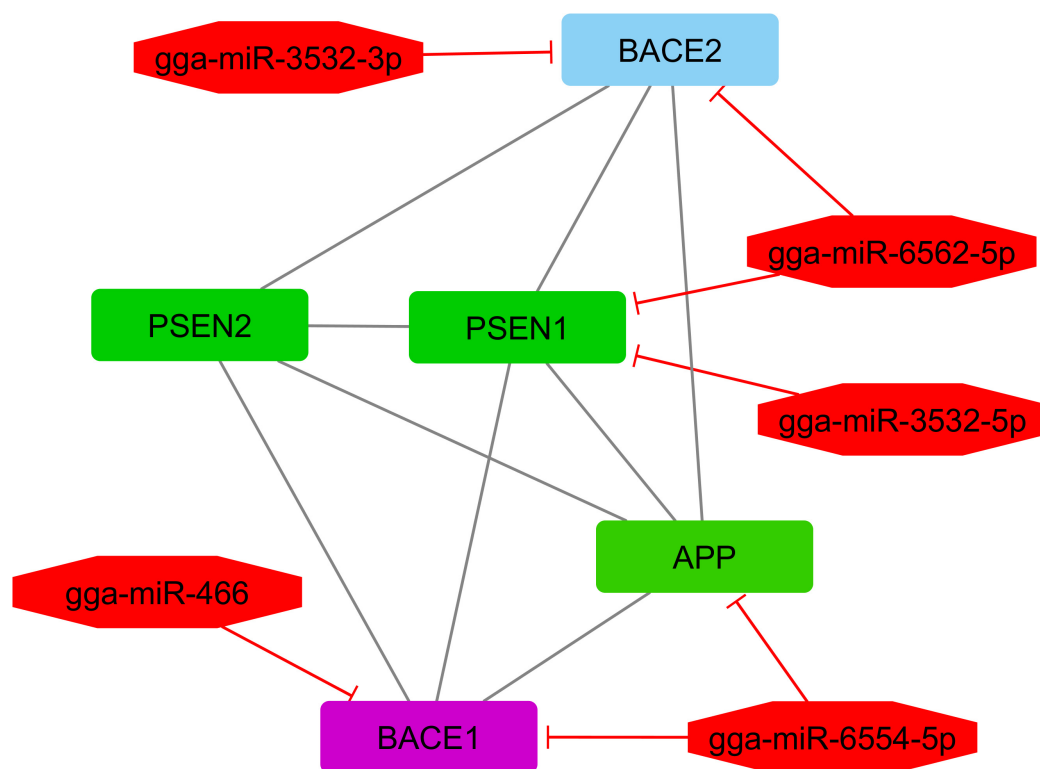


FIGURE 5 | Module 2: five genes and five miRNAs in the interactive bipartite network of gene-miRNA. In this network, the quadrilateral points represent genes; and the octagonal points represent miRNAs. In this interactive bipartite network of gene-miRNA, quadrilateral nodes represent genes; and octagonal nodes represent the miRNAs. For miRNAs and target genes, the edges indicate the suppressing roles of miRNAs. The edges of genes also indicate the gene-gene interactions. The green quadrilateral nodes represent the genes with the highest gene-gene interactions with other genes in the network (or hub genes). Purple quadrilateral nodes indicate the genes with the highest suppression by miRNAs.

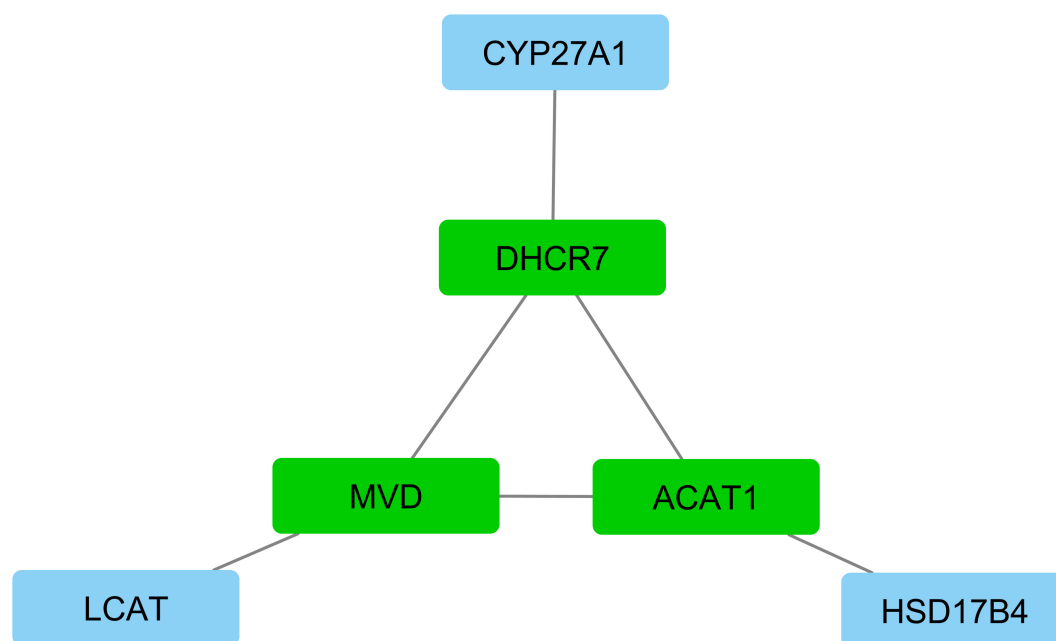


FIGURE 6 | Module 3: six genes in interactive bipartite network of gene-miRNA. In this network, the quadrilateral nodes represent genes; and edges indicate the gene-gene interaction effects. Green quadrilateral nodes represent the hub genes in the network. Blue nodes represent other genes in the network.

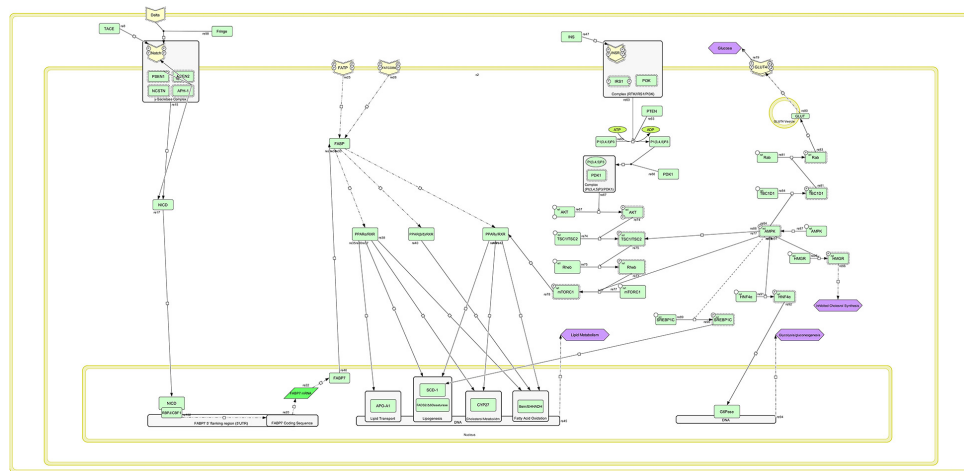


FIGURE 7 | Schematic of the regenerated metabolic-signaling network associated with fat metabolism and storage using CellDesigner.

factor activity, and it regulates the rate of transcription of the *LDL receptor* gene, fatty acid, and the cholesterol synthesis pathway to a lesser extent (Stachowiak et al., 2013).

HMGCR or 3-hydroxy-3-methylglutaryl coenzyme A reductase was downregulated in the lipolysis process because the *HMGCR* gene is a protein-encoding gene; the Terpenoid backbone biosynthesis pathway is a pathway associated with this gene (Wang et al., 2018).

Fatty acid desaturase 2 gene was upregulated in the lipogenesis process because the *FADS2* gene is a protein-encoding gene with pathways such as fatty acid beta-oxidation (peroxisome) and alpha-linolenic acid metabolism. This gene is a part of the lipid metabolic pathway that catalyzes the biosynthesis of unsaturated fatty acids from unsaturated fatty acids of linoleic acid (18:2n-6) and linolenic acid (18:3n-3) (Chen et al., 2019).

The *Stearoyl-coenzyme a desaturase (SCD)* gene was upregulated in the lipogenesis process, as this gene encodes the enzyme that is involved in the biosynthesis of fatty acids, so that it is first responsible for the synthesis of oleic acid. The produced protein belongs to the desaturase fatty acid family (Calvo et al., 2019).

Acetyl-Coenzyme A acetyltransferase 1 (ACAT1) was downregulated in the lipolysis process. This gene is a protein-encoding gene that is involved in metabolic pathways of ketone body metabolism and the Terpenoid backbone biosynthesis. The gene plays a key role in the ketone body metabolism (Chanyshv et al., 2018).

The *FASN (Fatty acid synthase)* gene was upregulated in the lipogenesis process because this gene is a protein-encoding gene with pathways such as the metabolism of water-soluble vitamins and cofactors, as well as the enzymatic complex pathway of AMPK. Therefore, upregulation of this gene is necessary for lipid biosynthesis (Raza et al., 2018).

The *Hydroxyacyl-CoA Dehydrogenase Trifunctional Multienzyme Complex Subunit Beta* gene was downregulated in the lipolysis process. The *HADHB* gene is a protein-encoding gene with pathways such as beta-oxidation of mitochondrial

fatty acids and biosynthesis of glycerophospholipids (Diebold et al., 2019).

Enoyl-CoA Hydratase And 3-Hydroxyacyl CoA Dehydrogenase genes were downregulated in the lipolysis process. The *EHHADH* gene is a protein-encoding gene with pathways such as PPAR alpha pathway and propanoate metabolism. The gene annotation includes the binding of signaling receptors and oxidoreductase activity (Assmann et al., 2016). Given the ontology expression and functions of important and main genes in the network of genes interactions, it can be stated that these genes are the main genes in the metabolism and fat storage as well as the signaling pathways of endocrine glands, especially AMPK and PPAR signaling pathways.

In **Figure 3**, green quadrilateral nodes represent the genes with the highest interaction in the network and are the main candidates in lipid metabolism and storage. These nodes play roles in the list of desired genes (reference), metabolic, and signaling pathways. The genes with the highest repression levels include *THBS1*, *SIK1*, *COLEC12*, and *BACE1*, respectively.

A combined biological system approach is used to detect metabolic and signaling pathways associated with the interactive bipartite network of gene-miRNA in the process of fat storage and metabolism of broiler chicken. Fat stored in the skeletal muscles plays a role in important metabolic processes such as immune function, food consumption, hormone sensitivity, and relevant signaling pathways (Jung and Choi, 2014).

In module 1, *gga-miR-1710* suppressed the *HMGCR* gene and *gga-miR-1710* was downregulated. Its target gene represents the increased expression in higher abdominal fat tissue compared with lower abdominal fat tissue. The gene is classified into a set of genes associated with the lipolysis process. Reducing the expression of *gga-miR-1710* and increased gene expression of *HMGCR* leads to the lipolysis process, thereby reducing abdominal fat. *HMG-CoA reductase* protein-encoding gene is the cholesterol synthesis-limiting enzyme that regulates the product of catalyzed reactions by reductase through a negative feedback mechanism caused by sterols and non-sterol metabolites derived

from mevalonate. The enzyme in mammalian cells is usually suppressed by cholesterol derived from the construction and destruction of low-density lipoprotein (LDL) through the LDL receptor (Wang et al., 2018).

The *SCD* gene indicates a higher expression in larger abdominal fat tissue compared with the lower abdominal fat tissue. The *SCD* gene is put into the set of genes associated with the lipogenesis process. Therefore, increasing the *SCD* gene expression raises the amount of fat storage in the body, especially in abdominal part. *SCD* gene (*Stearoyl-coenzyme A desaturase*) is a protein-encoding gene with pathways including adipogenesis and angiopoietin, such as the protein 8 regulatory pathway. It also plays an important role in lipid biosynthesis and regulating the expression of genes in the mitochondrial fatty acid oxidation and lipogenesis cycle (Aali et al., 2016).

gga-miR-6554-5p suppresses the *IRS1* gene. This miRNA has higher expression; and its target gene shows a lower expression in greater abdominal fat tissue compared with the lower abdominal fat tissue. *IRS1* gene is among the set of genes associated with the lipolysis process. Therefore, increasing expression of *gga-miR-6554-5p* miRNA decreases the *IRS1* gene expression, thereby reducing the amount of fat catabolism and increasing the abdominal fat storage and anabolism. *IRS1* gene encodes a protein that is phosphorylated by insulin receptor tyrosine kinase. Mutations in the gene are associated with type 2 diabetes and insulin resistance (Song et al., 2019).

The *SREBF1* gene shows the higher expression in greater abdominal fat tissue compared with lower abdominal fat tissue. The gene is among the set of genes associated with the lipogenesis process. The higher *SREBF1* gene expression increases the abdominal fat storage and anabolism. *SREBF1* genes encode the Helix-Loop-Helix-Leucine Zipper (bHLH-Zip) that binds the sterol-1 regulator. It is also found in the promoter for low-density lipoprotein receptor gene and other genes in the sterol biosynthesis (Stachowiak et al., 2013).

In this module, the *HMGCR* gene is suppressed by miRNAs. The gene is associated with the lipolysis process. Therefore, its suppression can prevent the fat tissue catabolism and lead to the higher fat storage and anabolism in abdominal fat tissue of broiler chickens. In this module, there are six genes, namely, *HMGCR*, *SREBF1*, *SCD*, *FASN*, *HADHB*, and *ACAT1* with certain color (green), and have the highest interaction with other genes involved in the module. The enzyme that is encoded by the *FASN* gene is a multi-functional protein. Its main function is the canalization of the synthesis of Palmitate from Acetyl-CoA and Malonyl-CoA in the presence of NADPH to long-chain saturated fatty acids. The *ACAT1* gene encodes a topical mitochondrial enzyme that catalyzes the reversible form of Acetoacetyl CoA from two acetyl CoA molecules. Further, the *HADHB* gene is responsible for encoding the beta subunit of mitochondrial function protein and catalyzes the final three stages of the mitochondrial beta-oxidation process of long-chain fatty acids (Diebold et al., 2019).

The gene set of this module, as presented in **Table 3**, encodes signaling pathways AMPK and PPAR as well as metabolic pathways of fatty acid synthase, unsaturated fatty acid synthase, and cholesterol metabolism pathways. Therefore, it can be

concluded that the module and genes involved in the process can be functional modules associated with abdominal fat metabolism and storage in broiler chickens.

The receptor increases the insulin-mediated glucose uptake and improves the blood lipid profile by regulating lipid metabolism, glucose, and free fatty acid oxidation. Target genes of PPARs are related to several proteins that are necessary for absorption, intercellular transfer, and beta-oxidation of fatty acids. They include fatty acid transport proteins, the Fatty Acid Translocase enzyme, and the synthase enzyme involved in the production of acetyl CoA (for long-chain fatty acids) and Carnitine palmitoyltransferase I (Brown and Plutzky, 2007). PPARs play roles in the regulation of the gene transcription process (P2) of fat cells, so that the lean and fat-free meat can be produced by manipulation of the differentiation of fat tissue cells and their fat content through these receptors.

The cellular response to insulin includes the regulation of blood sugar levels by increasing the glucose uptake in muscles and fat tissues in a way that energy is reserved in fat tissue, liver, and muscle increase by stimulating lipogenesis, glycogen synthesis, and protein synthesis. Insulin signaling pathways decrease glucose production by the liver and the total inhibition of energy stored through lipolysis, glycogenolysis, and breakdown of proteins. This pathway also acts as a growth factor and stimulates cell growth, differentiation, and survival (Boucher et al., 2014). The insulin signaling pathway is an important biochemical pathway that regulates some basic biological functions such as glucose and lipid metabolism, synthesis of proteins, cell proliferation and differentiation, and apoptosis (Di Camillo et al., 2016).

The signaling pathway of phosphatidylinositol (PI3K)/protein kinase B (Akt) is involved in the regulation of many physiological cell processes by activating effective cross-downstream molecules that play important roles in the cellular cycle, growth, and proliferation (Shi et al., 2019).

The Mammalian Target of Rapamycin (mTOR) signaling pathway has both internal and external signals and acts as a main regulator of cellular metabolism, growth, proliferation, and survival. Exploration carried out over the past decade indicates that the mTOR signaling pathway is activated in various cellular processes such as tumor formation and angiogenesis, insulin resistance, lipid metabolism, and lymphocyte T activation and is regulated in human diseases such as cancer and type 2 diabetes (Laplanche and Sabatini, 2009).

In module 2, *APP* gene plays the main role. The gene is suppressed by *gga-miR-6554-5p*. *gga-miR-6554-5p* represents the upregulation; and its target gene represents the downregulation in greater abdominal fat tissue compared with the lower abdominal fat tissue. The *APP* gene is a set of genes associated with the lipolysis process. Therefore, its repression by miRNAs in humans is necessary. In poultry, its lower expression is equivalent to a decrease in abdominal fat; and a decrease in body fat is equivalent to an increase in proliferation performance and other functional traits. Increased body weight or obesity caused by increased body fat storage is characterized by excessive accumulation of fat in the body and increased levels of adipokines and inflammatory cytokines. This indicates an increased risk of

Alzheimer's disease, type 2 diabetes, and cardiovascular diseases. It has been recently found that the gene expression level of *APP* increases as brain tissue fat and fat storage tissues increase in the body (Puig et al., 2017).

gga-miR-6554-5p and *gga-miR-466* miRNAs suppress *BACE1* gene. These two miRNAs represent the upregulation, and their target genes indicate the downregulation in greater abdominal fat tissue compared with lower abdominal fat tissue. *BACE1* gene encodes an enzyme that cuts the amyloid precursor protein (APP) and produces amyloid beta peptides that cause amyloid plaque in the brains of patients with Alzheimer's disease (Faghihi et al., 2008; Ghafouri et al., 2018).

gga-miR-6562-5p and *gga-miR-3532-5p* suppress the *PSEN1* gene. These two miRNAs indicate the upregulation; and their target gene indicates the downregulation in the greater abdominal fat tissue compared with the lower abdominal fat tissue. *PSEN1* encodes a protein that is called Presenilin 1. Presenilins are APP regulators according to their effects on gamma secretase as APP-decomposing enzymes (Ramakrishnan et al., 2017).

PSEN2 gene, which has about 67% of similarity to *PSEN1* gene, was identified after *PSEN1* gene. *PSEN2* gene indicated a lower expression in greater abdominal fat tissue compared with the lower abdominal fat tissue. *PSEN2* gene is a protein-encoding gene with associated diseases such as Alzheimer's disease and heart muscle diseases. It encodes the intermediate signaling Presenilin and Wnt/Hedgehog/Notch pathways (Muchnik et al., 2015).

gga-miR-3532-3p suppresses *BACE2* gene. The miRNAs indicate the upregulation; and their target gene, *BACE2*, indicates the low expression in greater abdominal fat tissue compared with lower abdominal fat tissue. *BACE2* gene encodes a full membrane glycoprotein that is known as an aspartic protease (Yu and Jia, 2009).

In module 3, the *MVD* gene indicated a higher expression in the greater abdominal fat tissue compared with the lower abdominal fat tissue. The *MVD* gene is a set of genes associated with the lipogenesis process. This gene encodes a mevalonate diphosphate decarboxylase (MVD) enzyme. Its related pathways include the protein metabolism and synthesis of available substrates in the biosynthesis of N-glycans. The *DHCR7* gene is another important gene of this module, indicating the higher expression in the greater abdominal fat tissue compared with the lower abdominal fat tissue. *DHCR7* or 7-dehydrocholesterol reductase is a protein-encoding gene that plays a role in eliminating an enzyme that creates a double bond of C (7–8) in loop B of sterol and catalyzes the conversion of 7-dehydrocholesterol to cholesterol. Cholesterol I biosynthesis and vitamin D metabolism are its associated pathways. The *TM7SF2* gene is an important paralog of this gene [see text footnote 6; 64].

Another important gene in this module, *ACAT1* gene, indicates low expression in greater abdominal fat tissue compared with the lower abdominal fat tissue. This gene catalyzes Acetoacetyl CoA using two acetyl coenzyme A molecules (Chanyshv et al., 2018).

Given the roles of the three main genes involved in the structure of this module as well as using the online database, this

module encodes metabolic pathways of cholesterol metabolism and the metabolism of fatty acids.

In the Notch signaling pathway, the Notch receptor is phosphorylated and activates the *NICD* gene in collaboration with the *PSEN1* gene as a γ -secretase complex. Inside the cell nucleus, this gene encodes the sequence of the *FABP7* gene and triggers the construction of *FABP7* mRNA by cooperation with RBPJ/CBF1 complex. *FABP* gene is activated by two phosphorylated receptors, called *FATP* and *FATCDB6*, in the cell membrane. Thereafter, three signaling complexes, *PPAR α -RXR*, *PPAR β -RXR*, and *PPAR γ -RXR*, are activated. These signaling pathways encode genes related to the fat storage and metabolism in the cell nucleus. These complexes in the nucleus are related to lipid transport, lipogenesis, cholesterol metabolism, and fatty acid oxidation, leading to the process of lipid metabolism by transcription and translation of the genes. In the signaling path of PPAR, *PPAR γ /RXR* complex is associated with the insulin-related signaling pathway through the phosphorylated mTORC1 gene in the mTOR pathway. The phosphorylation of this gene results in activation of *PPAR γ /RXR* complex. The AMPK signaling pathway is also associated with the *mTORC1* gene and has an inhibitory effect, in a way that the AMPK pathway prevents the phosphorylation of the *mTORC1* gene, so that *PPAR γ /RXR* complex is not activated; and the lipid metabolism process (e.g., lipogenesis, cholesterol, and oxidation metabolisms) is not performed. Two signaling pathways, PPAR (the main pathway of lipid metabolism) and AMPK (the main pathway of cellular energy exchanges), are important in this metabolic-signaling network. These two signaling pathways control each other by the *mTORC1* gene in the mTOR signaling path, so that increasing or decreasing the intracellular energy levels of the AMPK signaling pathway with an inhibitory or activating effect on the *mTORC1* gene can cause anabolism or catabolism of lipids in cells (Figure 7).

According to the ontology and functions of genes, which encode two signaling pathways, AMPK and PPAR, these two pathways are the main pathways of cellular energy exchange and lipid metabolism, respectively.

Peroxisome proliferator-activated receptors are transcription factors belonging to the nuclear receptor superfamily, and they are activated by long-chain unsaturated fatty acids with several double bonds, eicosanoids, and lipid-lowering agents such as fibrates. Among the unsaturated fatty acids with double bonds, eicosapentaenoic acid (EPA) and docosahexaenoic acid have been widely studied because of their ability to activate PPARs. The expression profile of *PPAR α* in different organs of poultry is largely similar to that of mammals, in such a way that it expresses similar functions of *PPAR α* in poultry and mammals. PPARs are nuclear hormone receptors that are activated by fatty acids and their derivatives. Each of them is encoded in a separate gene and bind fatty acids and eicosanoids. Ligand property of PPAR–RXR heterodimers for fatty acids causes the binding of these heterodimers to “Specific Receptor Elements” in the promoter region of several genes and changes the transcription of downstream genes involved in immune processes, lipid metabolism, and cholesterol metabolism (Zoete et al., 2007).

AMP-activated protein kinase (AMPK) is a serine/threonine kinase that has a high protective system. The AMPK system acts as a cellular energy sensor. When AMPK is activated, it simultaneously inhibits the energy consumption in biosynthetic pathways, such as protein, fatty acids, and glycogen synthesis, and activates the catabolic pathways (breakdown) of ATP production, including fatty acid oxidation and glycolysis (Miyamoto et al., 2012). The reduced regulation of liver AMPK activity plays a pathophysiological role in lipid metabolic disorders. However, the signaling pathway of AMPK for regulation of cellular energy balance is essential for the lipid metabolism, so that the pathway activates the catabolism of fat in the shortage of energy in the cell to provide the necessary rate of ATP. Therefore, the AMPK is a main regulator of cell metabolism and metabolism organ in eukaryotes, and it is activated by lowering the intra-cellular ATP level. AMPK plays an important role in the growth regulation and re-planning of cell metabolism (Mihaylova and Shaw, 2011).

CONCLUSION

The combination of omics data for obtaining and identifying genes with differences in gene expression led to the successful identification of 41 genes in the main process of metabolism (anabolism and catabolism), fat storage, signaling pathways of endocrine glands, and the cell membrane in abdominal fat tissue for two groups of broiler chickens with higher and lower abdominal fat storage. The same identified genes were involved in the signaling pathways of endocrine glands; AMPK and PPAR are associated with lipid metabolism and energy catabolism and could be considered as the genes that were similar in different species. The present study identified important common genes relating to lipid metabolism and metabolic and signaling pathways, and detected mechanisms associated with lipid transfer by different cell membranes and tissues by an explanation of relevant genes. Furthermore, the gene–gene and gene–miRNA interactions were also examined by investigating the biological system and reconstruction of various regulatory and interactive networks that can affect the regulation of fat metabolism and storage in poultry. They also facilitate better understanding biology of metabolism and fat storage and the discovery of potential molecular markers in poultry industry programs to increase animal protein production efficiency and reduce abdominal fat storage.

DATA AVAILABILITY STATEMENT

All sequencing data have been submitted to the National Center for Biotechnology Information (NCBI) Gene Expression

Omnibus (GEO). Each dataset contains microRNA expression raw data files (fastq format), processed data files (raw counts of sequencing reads), and a metadata spreadsheet referring to the information about the overall study and individual samples. All data can be used without restrictions. Related accession number is GSE122224. As well as reconstructed metabolic signaling pathways (in.xml format) have been submitted to the BioModels database in the European Bioinformatics Institute (EMBL-EBI). Related biomodel accession number is MODEL2010270002 (<https://www.ebi.ac.uk/biomodels/>).

ETHICS STATEMENT

The animal study was reviewed and approved by The Committee on the Care and Use of Chicken Breeding Station of Tehran University farm has approved the experiments.

AUTHOR CONTRIBUTIONS

AB, SM-A, and MS contributed to conceptualization. AB contributed to methodology. FG and AB contributed to formal analysis. FG, SM-A, MS, and AB contributed to investigation. FG, AB, and MB contributed to data curation. FG, MB, and AB contributed to writing – original draft preparation. MB, HB, and SL contributed to writing – review and editing. MS, AB, and SM-A contributed to supervision. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

This work was financially supported by the University of Tehran, Iran. We thank all the teams who worked on the experiments and who provided technical assistance in the laboratory during this study. We also thank the reviewers whose critical comments helped in improving the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.646297/full#supplementary-material>

REFERENCES

- Aali, M., Moradi-Shahrbabak, H., Moradi-Shahrbabak, M., Sadeghi, M., and Kohram, H. (2016). Polymorphism in the SCD gene is associated with meat quality and fatty acid composition in Iranian fat-and thin-tailed sheep breeds. *Livestock Sci.* 188, 81–90. doi: 10.1016/j.livsci.2016.04.003
- Almkvist, O., Rodriguez-Vieitez, E., Thordardottir, S., Nordberg, A., Viitanen, M., Lannfelt, L., et al. (2019). Longitudinal cognitive decline in autosomal-dominant Alzheimer's disease varies with mutations in APP and PSEN1 genes. *Neurobiol. Aging* 82, 40–47. doi: 10.1016/j.neurobiolaging.2019.06.010
- Andrews, S. (2010). *FastQC: a Quality Control Tool for High Throughput Sequence Data*. Available Online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

- Arora, N., Pienkos, P. T., Pruthi, V., Poluri, K. M., and Guarnieri, M. T. (2018). Leveraging algal omics to reveal potential targets for augmenting TAG accumulation. *Biotechnol. Adv.* 36, 1274–1292. doi: 10.1016/j.biotechadv.2018.04.005
- Assmann, N., Dettmer, K., Simbuerger, J. M., Broeker, C., Nuernberger, N., Renner, K., et al. (2016). Renal fanconi syndrome is caused by a mistargeting-based mitochondriopathy. *Cell Rep.* 15, 1423–1429. doi: 10.1016/j.celrep.2016.04.037
- Backman, M., Flenkenthaler, F., Blutke, A., Dahlhoff, M., Ländström, E., Renner, S., et al. (2019). Multi-omics insights into functional alterations of the liver in insulin-deficient diabetes mellitus. *Mol. Metab.* 26, 30–44. doi: 10.1016/j.molmet.2019.05.011
- Bader, G. D., Betel, D., and Hogue, C. W. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.* 31, 248–250. doi: 10.1093/nar/gkg056
- Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2.
- Bahrami, A., Miraie-Ashtiani, S. R., Sadeghi, M., and Najafi, A. (2017a). miRNA-mRNA network involved in folliculogenesis interactome: systems biology approach. *Reproduction* 154, 51–65. doi: 10.1530/REP-17-0049
- Bahrami, A., Miraie-Ashtiani, S. R., Sadeghi, M., Najafi, A., and Ranjbar, R. (2017b). Dynamic modeling of folliculogenesis signaling pathways in the presence of miRNAs expression. *J. Ovar. Res.* 10, 76–85. doi: 10.1186/s13048-017-0371-y
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi: 10.1016/S0092-8674(04)00045-5
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Boucher, J., Kleinriders, A., and Kahn, C. R. (2014). Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harb. Perspect. Biol.* 6:a009191. doi: 10.1101/cshperspect.a009191
- Bourneuf, E., Héroult, F., Chicault, C., Carré, W., Assaf, S., Monnier, A., et al. (2006). Microarray analysis of differential gene expression in the liver of lean and fat chickens. *Gene* 372, 162–170. doi: 10.1016/j.gene.2005.12.028
- Brown, J. D., and Plutzky, J. (2007). Peroxisome proliferator-activated receptors as transcriptional nodal points and therapeutic targets. *Circulation* 115, 518–533. doi: 10.1161/circulationaha.104.475673
- Byerly, M. S., Simon, J., Cogburn, L. A., Le Bihan-Duval, E., Duclos, M. J., Aggrey, S. E., et al. (2010). Transcriptional profiling of hypothalamus during development of adiposity in genetically selected fat and lean chickens. *Physiol. Genom.* 42, 157–167. doi: 10.1152/physiolgenomics.00029.2010
- Calvo, J. H., González-Calvo, L., Dervishi, E., Blanco, M., Iguácel, L. P., Sarto, P., et al. (2019). A functional variant in the stearoyl-CoA desaturase (SCD) gene promoter affects gene expression in ovine muscle. *Livestock Sci.* 219, 62–70. doi: 10.1016/j.livsci.2018.11.015
- Cesar, A. S., Regitano, L. C., Reecy, J. M., Poleti, M. D., Oliveira, P. S., de Oliveira, G., et al. (2018). Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. *BMC Genomics* 19:499–519. doi: 10.1186/s12864-018-4871-y
- Chanyshv, M. D., Razumova, Y. V., Ovchinnikov, V. Y., and Gulyaeva, L. F. (2018). MiR-21 regulates the ACAT1 gene in MCF-7 cells. *Life Sci.* 209, 173–178. doi: 10.1016/j.lfs.2018.08.010
- Chen, X., Wu, Y., Zhang, Z., Zheng, X., Wang, Y., Yu, M., et al. (2019). Effects of the rs3834458 single nucleotide polymorphism in FADS2 on levels of n-3 long-chain polyunsaturated fatty acids: a meta-analysis. *Prostaglandins Leukot. Essent. Fat. Acids* 150, 1–6. doi: 10.1016/j.plefa.2019.08.005
- Cole, J. B., Lewis, R. M., Maltecca, C., Newman, S., Olson, K. M., and Tait, R. G. (2013). Systems biology in animal breeding: identifying relationships among markers, genes, and phenotypes. *Breed. Genet. Sym.* 91, 521–522. doi: 10.2527/jas2012-6166
- Corral-Jara, K. F., Cantini, L., Poupin, N., Ye, T., Rigaudière, J. P., De Saint Vincent, S., et al. (2020). An integrated analysis of mirna and gene expression changes in response to an obesogenic diet to explore the impact of transgenerational supplementation with omega 3 fatty acids. *Nutrients* 12:3864. doi: 10.3390/nu12123864
- Crespo, N., and Esteve-García, E. (2001). Dietary fatty acid profile modifies abdominal fat deposition in broiler chickens. *Poult. Sci.* 80, 71–78. doi: 10.1093/ps/80.1.71
- Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 14, 1846–1847. doi: 10.1093/bioinformatics/btm254
- Dao, M. C., Sokolovska, N., Brazeilles, R., Affeldt, S., Pelloux, V., Prifti, E., et al. (2019). A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity. *Front. Physiol.* 9:1958.
- Diebold, I., Schön, U., Horvath, R., Schwartz, O., Holinski-Feder, E., Köbel, H., et al. (2019). HADHA and HADHB gene associated phenotypes-Identification of rare variants in a patient cohort by next generation sequencing. *Mol. Cell. Prob.* 44, 14–20. doi: 10.1016/j.mcp.2019.01.003
- Di Camillo, B., Carlon, A., Eduati, F., and Toffolo, G. M. (2016). A rule-based model of insulin signalling pathway. *BMC Syst. Biol.* 10:38–51.
- Du, P., Kibbe, W. A., and Lin, S. M. (2008). 'lumi: a pipeline for processing Illumina microarray'. *Bioinformatics* 24, 1547–1548. doi: 10.1093/bioinformatics/btn224
- Dweep, H., Gretz, N., and Sticht, C. (2014). miRWalk database for miRNA-target interactions. *Methods Mol. Biol.* 1182, 289–305. doi: 10.1007/978-1-4939-1062-5_25
- Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., et al. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β -secretase. *Nat. Med.* 14:723. doi: 10.1038/nm1784
- Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE* 96, 1254–1265. doi: 10.1109/JPROC.2008.925458
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405
- Ghafouri, F., Sadeghi, M., and Bahrami, A. (2018). Long non-coding RNAs (LncRNAs): roles, functions, and mechanisms. *Genet. Eng. Biosaf. J.* 7, 226–235.
- Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engle, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* 27, 91–105. doi: 10.1016/j.molcel.2007.06.017
- Hasin, Y., Seldin, M., and Lusi, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83.
- Hsu, S. D., Chu, C. H., Tsou, A. P., Chen, S. J., Chen, H. C., Hsu, P. W., et al. (2008). miRNAmap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res.* 36, D165–D169. doi: 10.1093/nar/gkm1012
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Prot.* 4, 44–57. doi: 10.1038/nprot.2008.211
- Huang, J. C., Morris, Q. D., and Frey, B. J. (2006). "Detecting microRNA targets by linking sequence, microRNA and gene expression data," in *Annual International Conference on Research in Computational Molecular Biology*, eds A. Apostolico, C. Guerra, S. Istrail, P. A. Pevzner, and M. Waterman (Berlin: Springer), 114–129. doi: 10.1007/11732990_11
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Meth.* 12:115. doi: 10.1038/nmeth.3252
- Iorio, M. V., Casalini, P., Piovani, C., Braccioli, L., and Tagliabue, E. (2011). Breast cancer and microRNAs: therapeutic impact. *Breast* 20, S63–S70. doi: 10.1016/S0960-9776(11)70297-1
- Ji, B., Ernest, B., Gooding, J. R., Das, S., Saxton, A. M., Simon, J., et al. (2012). Transcriptomic and metabolomic profiling of chicken adipose tissue in response to insulin neutralization and fasting. *BMC Genomics* 13:441. doi: 10.1186/1471-2164/13/441
- Ji, B., Middleton, J. L., Ernest, B., Saxton, A. M., Lamont, S. J., Campagna, S. R., et al. (2014). Molecular and metabolic profiles suggest that increased lipid catabolism in adipose tissue contributes to leanness in domestic chickens. *Physiol. Genom.* 46, 315–327. doi: 10.1152/physiolgenomics.00163.2013
- Jung, U. J., and Choi, M. S. (2014). Obesity and its metabolic complications: the role of adipokines and the relationship between obesity, inflammation, insulin resistance, dyslipidemia and nonalcoholic fatty liver disease. *Int. J. Mol. Sci.* 15, 6184–6223. doi: 10.3390/ijms15046184
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptsomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.
- Lagarrigue, S., Pitel, F., Carré, W., Abasht, B., Le Roy, P., Neau, A., et al. (2006). Mapping quantitative trait loci affecting fatness and breast muscle weight in

- meat-type chicken lines divergently selected on abdominal fatness. *Genet. Sel. Evol.* 38, 85–97. doi: 10.1051/gse:2005028
- Laplanche, M., and Sabatini, D. M. (2009). mTOR signaling at a glance. *J. Cell Sci.* 122, 3589–3594. doi: 10.1242/jcs.051011
- Le Mignon, G., Pitel, F., Gilbert, H., Le Bihan-Duval, E., Vignoles, F., Demeure, O., et al. (2009). A comprehensive analysis of QTL for abdominal fat and breast muscle weights on chicken chromosome 5 using a multivariate approach. *Anim. Genet.* 40, 157–164. doi: 10.1111/j.1365-2052.2008.01817.x
- Lee, B., Zhang, S., Poleksic, A., and Xie, L. (2020). Heterogeneous multi-layered network model for omics data integration and analysis. *Front. Genet.* 10:1381.
- Loher, P., and Rigoutsos, I. (2012). Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28, 3322–3323. doi: 10.1093/bioinformatics/bts615
- Mi, H., Muruganujan, A., and Thomas, P. D. (2012). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* 41, D377–D386. doi: 10.1093/nar/gks1118
- Mihaylova, M. M., and Shaw, R. J. (2011). The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. *Nat. Cell Biol.* 13, 1016–1023. doi: 10.1038/ncb2329
- Miyamoto, L., Ebihara, K., Kusakabe, T., Aotani, D., Yamamoto-Kataoka, S., Sakai, T., et al. (2012). Leptin activates hepatic 5'-AMP-activated protein kinase through sympathetic nervous system and α 1-adrenergic receptor: a potential mechanism for improvement of fatty liver in lipodystrophy by leptin. *J. Biol. Chem.* 287, 40441–40447. doi: 10.1074/jbc.M112.384545
- Muchnik, C., Olivari, N., Dalmasso, M. C., Azurmendi, P. J., Liberczuk, C., Morelli, L., et al. (2015). Identification of PSEN2 mutation p. N141I in argentine pedigrees with early-onset familial Alzheimer's disease. *Neurobiol. Aging* 36, 2674–2677. doi: 10.1016/j.neurobiolaging.2015.06.011
- Nones, K., Ledur, M. C., Zanella, E. L., Klein, C., Pinto, L. F. B., Moura, A. S. A. M. T., et al. (2012). Quantitative trait loci associated with chemical composition of the chicken carcass. *Anim. Genet.* 43, 570–576. doi: 10.1111/j.1365-2052.2012.02321.x
- Pinto, L. F. B., Packer, I. U., Ledur, M. C., Moura, A. S. A. M. T., Nones, K., and Coutinho, L. L. (2010). Mapping quantitative trait loci in *Gallus gallus* using principal components. *Revista Brasileira Zootecnia* 39, 2434–2441. doi: 10.1590/S1516-35982010001100016
- Puig, K. L., Brose, S. A., Zhou, X., Sens, M. A., Combs, G. F., Jensen, M. D., et al. (2017). Amyloid precursor protein modulates macrophage phenotype and diet-dependent weight gain. *Sci. Rep.* 7:43725. doi: 10.1038/srep43725
- Ramakrishnan, V., Husain, R. A., and Ahmed, S. S. (2017). PSEN1 gene polymorphisms in caucasian Alzheimer's disease: a meta-analysis. *Clin. Chim. Acta* 473, 65–70. doi: 10.1016/j.cca.2017.08.016
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucl. Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369
- Raza, S. H. A., Gui, L., Khan, R., Schreurs, N. M., Xiaoyu, W., Wu, S., et al. (2018). Association between FASN gene polymorphisms ultrasound carcass traits and intramuscular fat in Qinchuan cattle. *Gene* 645, 55–59. doi: 10.1016/j.gene.2017.12.034
- Resnyk, C. W., Carré, W., Wang, X., Porter, T. E., Simon, J., Le Bihan-Duval, E., et al. (2013). Transcriptional analysis of abdominal fat in genetically fat and lean chickens reveals adipokines, lipogenic genes and a link between hemostasis and leanness. *BMC Genomics* 14:557–583. doi: 10.1186/1471-2164/14/557
- Resnyk, C. W., Chen, C., Huang, H., Wu, C. H., Simon, J., Le Bihan-Duval, E., et al. (2015). RNA-Seq analysis of abdominal fat in genetically fat and lean chickens highlights a divergence in expression of genes controlling adiposity, hemostasis, and lipid metabolism. *PLoS One* 10:10.
- Resnyk, C. W., Carré, W., Wang, X., Porter, T. E., Simon, J., Le Bihan-Duval, E., et al. (2017). Transcriptional analysis of abdominal fat in chickens divergently selected on bodyweight at two ages reveals novel mechanisms controlling adiposity: validating visceral adipose tissue as a dynamic endocrine and metabolic organ. *BMC Genomics* 18:626–657. doi: 10.1186/s12864-017-4035-5
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Sakomura, N. K., Longo, F. A., Oviedo-Rondon, E. O., Boa-Viagem, C., and Ferraudo, A. (2005). Modeling energy utilization and growth parameter description for broiler chickens. *Poul. Sci.* 84, 1363–1369. doi: 10.1093/ps/84.9.1363
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Shi, X., Wang, J., Lei, Y., Cong, C., Tan, D., and Zhou, X. (2019). Research progress on the PI3K/AKT signaling pathway in gynecological cancer. *Mol. Med. Rep.* 19, 4529–4535.
- Song, Q., Song, J., Li, C., Liu, Z., Wang, Y., Qi, L., et al. (2019). Physical activity attenuates the association between the IRS1 genotype and childhood obesity in Chinese children. *Nutr. Metab. Cardiovasc. Dis.* 29, 793–801. doi: 10.1016/j.numecd.2019.05.058
- Stachowiak, M., Nowacka-Woszek, J., Szydlowski, M., and Switonski, M. (2013). The ACACA and SREBF1 genes are promising markers for pig carcass and performance traits, but not for fatty acid content in the longissimus dorsi muscle and adipose tissue. *Meat Sci.* 95, 64–71. doi: 10.1016/j.meatsci.2013.04.021
- Sticht, C., De La Torre, C., Parveen, A., and Gretz, N. (2018). miRWalk: An online resource for prediction of microRNA binding sites. *PLoS One* 13:10. doi: 10.1371/journal.pone.0206239
- Suravajhala, P., Kogelman, L. J., and Kadarmideen, H. N. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genet. Sel. Evol.* 48:38.
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucl. Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131
- Tsang, J. S., Ebert, M. S., and van Oudenaarden, A. (2010). Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell* 38, 140–153. doi: 10.1016/j.molcel.2010.03.007
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.* 28, 511–515. doi: 10.1038/nbt.1621
- Vejanar, C. E., and Zdobnov, E. M. (2012). MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.* 40, 11673–11683. doi: 10.1093/nar/gks901
- Wang, Y. Z., Yang, L., and Li, C. F. (2018). Protective effect of atorvastatin mediated by HMGCR gene on diabetic rats with atherosclerosis: An in vivo and in vitro study. *Biomed. Pharma.* 104, 240–251.
- Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucl. Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537
- Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., et al. (2015). Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Sci. Rep.* 5:17201.
- Yu, Y., and Jia, J. (2009). Lack of association between the polymorphisms of β -site APP-cleaving enzyme 2 (BACE2) 5'-flanking region and sporadic Alzheimer's disease. *Brain Res.* 1257, 10–15. doi: 10.1016/j.brainres.2008.12.024
- Zhang, C., Wang, Y., Zhang, C. L., and Wu, H. R. (2019). Prioritization of candidate metabolites for postmenopausal osteoporosis using multi-omics composite network. *Exp. Ther. Med.* 17, 3155–3161.
- Zoete, V., Grosdidier, A., and Michielin, O. (2007). Peroxisome proliferator-activated receptor structures: ligand specificity, molecular switch and interactions with regulators. *Biochim. Biophys. Acta (BBA) Mol. Cell Biol. Lipids* 1771, 915–925. doi: 10.1016/j.bbalip.2007.01.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Ghafari, Bahrami, Sadeghi, Miraei-Ashtiani, Bakherad, Barkema and Larose. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome-Wide Scanning for Signatures of Selection Revealed the Putative Genomic Regions and Candidate Genes Controlling Milk Composition and Coat Color Traits in Sahiwal Cattle

Satish Kumar Illa¹, Sabyasachi Mukherjee¹, Sapna Nath² and Anupama Mukherjee^{1*}

¹ Division of Animal Genetics and Breeding, Indian Council of Agricultural Research-National Dairy Research Institute, Karnal, India, ² Artificial Breeding Research Center, Indian Council of Agricultural Research-National Dairy Research Institute, Karnal, India

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

Guillermo Giovambattista,
CONICET Institute of Veterinary
Genetics (IGEVET), Argentina
Qianjun Zhao,
Chinese Academy of Agricultural
Sciences (CAAS), China

*Correspondence:

Anupama Mukherjee
writetoanupama@gmail.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 23 April 2021

Accepted: 14 June 2021

Published: 09 July 2021

Citation:

Illia SK, Mukherjee S, Nath S and
Mukherjee A (2021) Genome-Wide
Scanning for Signatures of Selection
Revealed the Putative Genomic
Regions and Candidate Genes
Controlling Milk Composition
and Coat Color Traits in Sahiwal
Cattle. *Front. Genet.* 12:699422.
doi: 10.3389/fgene.2021.699422

Background: In the evolutionary time scale, selection shapes the genetic variation and alters the architecture of genome in the organisms. Selection leaves detectable signatures at the genomic coordinates that provide clues about the protein-coding regions. Sahiwal is a valuable indicine cattle adapted to tropical environments with desirable milk attributes. Insights into the genomic regions under putative selection may reveal the molecular mechanisms affecting the quantitative and other important traits. To understand this, the present investigation was undertaken to explore signatures of selection in the genome of Sahiwal cattle using a medium-density genotyping INDUS chip.

Result: De-correlated composite of multiple selection signals (DCMS), which combines five different univariate statistics, was computed in the dataset to detect the signatures of selection in the Sahiwal genome. Gene annotations, Quantitative Trait Loci (QTL) enrichment, and functional analyses were carried out for the identification of significant genomic regions. A total of 117 genes were identified, which affect a number of important economic traits. The QTL enrichment analysis highlighted 14 significant [False Discovery Rate (FDR)-corrected p -value ≤ 0.05] regions on chromosomes BTA 1, 3, 6, 11, 20, and 21. The top three enriched QTLs were found on BTA 6, 20, and 23, which are associated with exterior, health, milk production, and reproduction traits. The present study on selection signatures revealed some key genes related with coat color (*PDGFRA*, *KIT*, and *KDR*), facial pigmentation (*LEF*), milk fat percent (*MAP3K1*, *HADH*, *CYP2U1*, and *SGMS2*), sperm membrane integrity (*OSTC*), lactation persistency (*MRPS30*, *NNT*, *CCL28*, *HMGCS1*, *NIM1K*, *ZNF131*, and *CCDC152*), milk yield (*GHR* and *ZNF469*), reproduction (*NKX2-1* and *DENND1A*), and bovine tuberculosis susceptibility (*RNF144B* and *PAPSS1*). Further analysis of candidate gene prioritization identified four hub genes, viz., *KIT*, *KDR*, *MAP3K1*, and *LEF*, which play a role in coat color, facial pigmentation, and milk fat percentage in cattle. Gene enrichment

analysis revealed significant Gene ontology (GO) terms related to breed-specific coat color and milk fat percent.

Conclusion: The key candidate genes and putative genomic regions associated with economic traits were identified in Sahiwal using single nucleotide polymorphism data and the DCMS method. It revealed selection for milk production, coat color, and adaptability to tropical climate. The knowledge about signatures of selection and candidate genes affecting phenotypes have provided a background information that can be further utilized to understand the underlying mechanism involved in these traits in Sahiwal cattle.

Keywords: indigenous cattle, SNP genotyping, selection signatures, gene identification, DCMS, gene

INTRODUCTION

Sahiwal is a well-known breed belonging to the humped zebu cattle group (*Bos indicus*), having its origin in the northwestern region of the Indian subcontinent (Muhuyi et al., 1999). Apart from their high milk-producing ability, these cattle breed also possesses unique adaptability to hot and humid climate prevailing in their native tract and their known resistance to tropical disease, ticks, and parasites (Singh et al., 2005). The milk of this cattle yields high fat (4.6–5.2%) and solid non-fat (SNF: 8.9–9.3%) (Joshi et al., 2001). Sahiwal cattle are utilized in crossbreeding programs around the world because of their higher milk production qualities and endurance to harsh environments (Maule, 1990). Such crossbreeding work has affected the numbers and distribution of indigenous purebred cattle to some extent in recent years (DAHDF, 2018/2019). In view of this, the Government of India established purebred cattle nucleus herds under All India Coordinated Research Project (AICRP). The major objectives of this AICRP are to conserve the valuable Sahiwal germplasm and to continue the genetic improvement program mostly for production traits.

It is now accepted broadly that the modern Zebu cattle were domesticated in the Indus valley approximately 6,800 years ago from Aurochs (*Bos primigenius nomadicus*) and was later introduced to different parts of tropical regions (Ajmone-Marsan et al., 2010). Domestication of livestock benefitted humankind in terms of milk, meat, and draft power. Selective breeding and genetic isolation help in formation of numerous cattle breeds and facilitates in maintaining the diversity of genome resources and retain the characteristics of adaptation to local environments (Felix, 1995). Artificial selection increases the beneficial alleles related to economic traits and aids in improving the production parameters (Diamond, 2002; Flori et al., 2009). The process of domestication and breed formation in mammals levied a constant source of selection pressure on a divergent variety of traits in all the domesticated species and left detectable impressions at individual genomes. These genomic regions provide straightforward clues about the functional variants concerning the traits (Andersson and Georges, 2004; Oleksyk et al., 2010). The advent of the cost-effective genotyping allowed more individuals to be genotyped with dense single nucleotide polymorphism (SNP) array and thereby facilitating the precise

identification of the genomic regions of livestock with better resolution and accuracy (Jensen et al., 2016). These developments also helped in the mapping of signatures under selection at the genome level in *Homo sapiens* and other species of animals (Mathieson et al., 2015; Moon et al., 2015). Selection signatures are specific variations at the DNA level that arise due to changes in the genomes of both selected and neutral loci of a species that has undergone selection over the years (Kreitman, 2000). Several statistical models were developed to determine the signatures of selection in recent years. Variants under selection pressure generate the typical genomic patterns such as (i) change in the allele frequency spectrum (either low or high frequencies); (ii) greater number of homozygous genotypes; (iii) long haplotypes are most common; and/or (iv) intense differentiation of local population. Several studies were conducted utilizing more than one statistic, viz., Integrated Haplotype Score (iHS), Cross-Population Extended Haplotype Homozygosity (XP-EHH), and Fixation index (F_{ST}) to detect the signatures of selection that exploit the advantage of complementarity of methods, intending to improve the statistical power (Weir and Cockerham, 1984; Voight et al., 2006). A new method was proposed in which different p -values are combined to give a composite of signals (CMS) for the first time (Grossman et al., 2010). Few studies suggested the use of combining different selection signals such as meta-SS and Composite Selection Signals (Utsunomiya et al., 2013; Randhawa et al., 2016). However, many of these earlier studies did not consider the covariance among the statistics, and consequently, a new method of composite signals was proposed where the outputs of different methods were combined and accounted for the covariance between the statistics. It differed from the other composite statistics by considering the dependencies among the univariate statistics and was termed as De-correlated composite of multiple selection signals (DCMS) (Ma et al., 2015; Lotterhos et al., 2017). This statistic is found to be more effective than earlier methods and helps in prioritizing the candidate genes affecting major economic traits in various species that are helpful to medicine, agriculture, and animal breeding. It is preferred over the univariate statistics as the latter retains high local resolution. Besides these beneficial attributes, DCMS estimation is also possible with better accuracy even with less demography information. Selection signature analysis in Russian and Swedish local cattle using de-correlated

DCMS revealed functional variants under selection related to production, reproduction, and adaptation (Yurchenko et al., 2018; Ghoreishifar et al., 2020).

Under the vast physio-geographical region of the Indian subcontinent, Sahiwal is a valuable *indicine* cattle adapted to tropical environments with desirable milk attributes, and it is apparent that the genomic regions of Sahiwal cattle will be under intense putative selection for centuries. Literature on these aspects is still scanty so far. Therefore, to understand the molecular mechanism affecting the quantitative and other important traits in these cattle, the present investigation was undertaken to explore signatures of selection in the genome of Sahiwal cattle. Simultaneously, annotation of genes and quantitative trait loci was also carried out for prioritizing the candidate genes having major effects on various production and adaptive traits including coat color in Sahiwal cattle using the medium-density genotyping assay with INDUSCHIP2, curated from Illumina BovineSNP50 Bead Chip and developed by the National Dairy Development Board (NDDB) for genotyping indicine cattle (NDDB, 2019).

MATERIALS AND METHODS

Animals and Genotyping

The study was conducted on 193 Sahiwal cattle belonging to the germplasm unit of AICRP maintained at ICAR-National Dairy Research Institute, Karnal, India. The entire population of Sahiwal cattle was further categorized into two subpopulations, viz., founder/unrelated animals ($n = 41$) and those farm-born/related ($n = 152$). These farm-born animals were born in the germplasm unit between 2003 and 2016. All the cattle were genotyped with INDUSCHIP2 consisting of 53,648 SNPs. This genotyping chipset is developed by the NDDB, India, which is customized from commercially available Illumina Bovine SNP chip (BovineSNP50K v3 Bead Chip) to genotype native cattle breeds and their crosses for implementing the genomic selection schemes in small and organized herds in India (NDDB, 2019).

The quality control (QC) of SNP data was implemented in Plink v1.9 program (Purcell et al., 2007). SNPs with a genotype call rate lower than 0.95 and a minor allele frequency less than 0.05 and those SNPs with Hardy–Weinberg equilibrium below 0.001 were removed. In addition, SNPs with duplicated position, located on the sex chromosome and with an unidentified position on UMD3.1 assembly, were excluded using the `–exclude` option. Furthermore, the highly related individual information was obtained from the `–genome` command in the form of PI-HAT indices. Individual pairs with unusually high PI-HAT values were discarded from the analysis to minimize the bias of sample size. Quality control of genotypes was again performed for phasing of haplotypes with the *Shapeit* program (Delaneau et al., 2012) to get high-quality SNPs. In total, 37,594 SNPs were considered after QC for final analysis.

Principal Component Analysis

Principal component analysis (PCA) was carried out in R environment with the *snprlate* package (Zheng et al., 2012) to explore the structure and clustering of the samples with the help

of plotting the genotypes of all the individuals, belonging to two subpopulations of Sahiwal cattle.

De-Correlated Composite of Multiple Selection Signals

As outlined by Yurchenko et al. (2018), the current study used the DCMS method to combine all the five statistics, viz., F_{ST} , Haplotype Homozygosity (H1), Modified Haplotype Homozygosity (H12), Tajima's D index, and Nucleotide diversity (π) (Nei and Li, 1979; Weir and Cockerham, 1984; Tajima, 1989; Garud et al., 2015). The DCMS statistic is computed at any given loci l as follows:

$$DCMS_l = \sum_{t=1}^n \frac{\log \left[\frac{1-p_{lt}}{p_{lt}} \right]}{\sum_{i=1}^n |r_{it}|}$$

Here, p_{lt} refers to the p -value at the l position for each statistic t . The denominator consists of a correlation component (r_{it}), which is the weighing factor at each locus. The weighing factors were genome-wide correlations between all the univariate statistic pairs (Ma et al., 2015); however, the statistic with greater correlation adds less to the calculation. To obtain the DCMS, all the statistics were converted into p -values using one-tailed and two-tailed ranks, where these fractional ranks lie between $1/(n+1)$ and $n/(n+1)$, respectively.

Fixation Index

Fixation index is known as a measure of population differentiation, and it was calculated for each SNP.

For the purpose of estimation of F_{ST} in our study, we have divided the entire Sahiwal population ($n = 193$) into two subpopulations, viz., founder/unrelated animals ($n = 41$) and farm-born/related animals ($n = 152$) based on the estimated average amount of IBD values sharing across all loci, i.e., pairwise relatedness. This is accomplished because F_{ST} is a parameter that measures genetic structure in a subdivided population. As per Wright (1951), F_{ST} is also the probability that alleles drawn randomly from a subpopulation are “identical by descent” (IBD), relative to an ancestral population.

The F_{ST} analysis was carried out between these two subpopulations of Sahiwal cattle using the `–fst` and `–within` functions of Plink1.9. Zero F_{ST} values were converted to zeros and the F_{ST} values of each SNP were smoothed with the *runmed* function in the R program.

After estimation of F_{ST} values, the other univariate statistics such as haplotype homozygosity (H1), modified haplotype homozygosity (H12), Tajima's D index, and Nucleotide Diversity (π) were estimated. Subsequently, all the five statistics were combined into one single framework of DCMS.

Haplotype Homozygosity Statistics (H1 and H12)

Phasing of each chromosome was carried out separately with the *SHAPEIT2 version* program (Delaneau et al., 2012) with default parameters like conditional states (`–states 400`) and the effective population (`–effective-size 108`), which was calculated using SNePV1.1 program with the default parameters (Barbato et al., 2015). A bovine recombination map was used to rectify the variation due to the recombination rate along the autosomal

genome (Ma et al., 2015). An R script was used to convert the phased haplotypes into the format as required by the H12_H1H2.py program¹. H1 and H12 statistics were obtained for all the SNPs by using the parameters of window size of 14 SNPs and step size of 1 (-window 14 -jump 1) (Garud et al., 2015).

Tajima's D and Nucleotide Diversity (π)

The vcftools program was used to compute Tajima's D and π (π) statistics (Danecek et al., 2011); both statistics were calculated for each chromosome separately (Danecek et al., 2011). Tajima's D index was obtained with the parameter of non-overlapping sliding windows of 300 MB (-Tajima D 300). The p -values were assigned to each SNP within the bin. All the missing values were changed to zero. The π values for all SNPs were computed with the function -site-pi function. The raw p -values were smoothed with the *runmed* function implemented in the R program with a window of 31 SNPs and constant end rule ($k = 31$; *endrule* = "constant") as described by Yurchenko et al. (2018).

DCMS Estimation

All five statistic parameters (H1, H12, Tajima's D index, π , and F_{ST}) for each SNP were combined to a new composite signal as DCMS. Based on the functional ranks, the left-tailed test was applied to Tajima's D values and π , and the right-tailed test was applied to H1 and H12 and F_{ST} statistic, respectively, using *stat_to_p-value* function in the MINOTAUR package in R environment (Verity et al., 2017). Later, the *covNAMcd* function ($\alpha = 0.75$, *nsamp* = 50,000) was applied from the *rrcovNA* R package (Todorov et al., 2011) and a correlation matrix of $n \times n$ order was calculated, which will be input to DCMS function of MINOTAUR R package (Verity et al., 2017) and the genome-wide DCMS values were computed. These DCMS values were transformed to a normal distribution with the robust linear model (*rlm*) using the *MASS* R package (Venables and Ripley, 2002) as outlined in Yurchenko et al. (2018). Then, DCMS statistics fitted to normal distribution were transformed to p -values by the *pnorm* function (*lower.tail* = FALSE, *log.p* = FALSE). Finally, thus obtained p -values were converted into the respective q -values after the Benjamini and Hochberg (1995) correction using the *q-value* R function (Storey and Tibshirani, 2003). The false discovery rate is calculated, which minimizes the error rate from multiple tests.

Identification of Functional Genes and QTL

The genomic regions were considered as significant if q -value is lower than 0.05 for adjacent SNPs. The boundaries of the genomic regions were determined from the SNP with a q value greater than 0.1. The gene and QTL annotations were performed using R package GALLO (Genomic Annotation in Livestock for positional candidate Loci) (Fonseca et al., 2020). The gene and QTL annotation files (.gtf and .gff files) derived from the ARS-UCD1.2 assembly (Rosen et al., 2018) and Animal QTL Database (Hu et al., 2013) were

used for the gene and QTL identification, respectively. The QTL enrichment analysis was also performed for all the QTLs annotated by the chromosome-based method using the same GALLO package. A bootstrap method was implemented to correlate the observed and expected number of QTLs per trait from the cattle QTL database with 1,000 iterations of random sampling. The calculated p -values in the enrichment analysis were also adjusted using FDR (<5%) for multiple testing.

Prioritization of Candidate Genes and Gene Enrichment Analysis

The candidate gene prioritization analysis was implemented in Topp Gene Suite (Chen et al., 2009), and the program requires a training set and a test set of genes. The training set of genes were obtained from GUILDify software by providing the keywords "sperm plasma membrane integrity," "milk yield," "milk fat percent," "lactation persistency," "facial pigmentation," "eye area pigmentation," "body weight," "maternal behavior," and "Bovine Tuberculosis Susceptibility." The gene list obtained from this analysis was used as an input in Topp Gene Suite and the identified genes were considered as a test gene set. The prioritization analysis is a multivariate method that utilizes the functional information from the Gene Ontology terms, human and mouse phenotypes, PubMed publications, and diseases. The significant p -values were obtained by linking all the p -values of a random sample of 5,000 genes. The candidate genes were prioritized after adjustment of p -values with $FDR \leq 5\%$. The prioritized genes were considered as an input in *NetworkAnalystv.3.0* (Zhou et al., 2019). A protein-protein interaction network was also generated, which is based on the string protein-protein interaction database with a confidence score cutoff of 900. Networks with nodes and edges were generated, and the networks with gene ontology terms such as Molecular function, Biological process, and Cellular components were also produced.

RESULTS

Quality Control and PCA

Quality control of genotypes for minor allele frequency, genotype call rate, Hardy-Weinberg equilibrium, and duplicated genotype parameters had excluded 12,842 SNPs and left the final dataset with 37,594 genotypes. The effective population size was calculated as 52 in Sahiwal population based on genotype data using SNeP V1.1 software.

The PCA was performed in the final dataset and showed that all the individuals in the dataset were homogeneous. Hence, the selection signature analysis was carried out by considering all the individuals as one group after exclusion of related individuals from the analysis with the help of the -genome command in Plink. This function had excluded 41 related individuals from the original dataset, having pairwise PI_HAT values above 0.1. Final analysis was performed on 152 individuals from the final dataset with 37,594 SNPs.

¹<https://github.com/ngarud/SelectionHapStats>

De-Correlated Composite of Multiple Selection Signals

The DCMS values were calculated for all 37,594 SNPs; the *p*-values were corrected ($FDR < 0.05$) and fitted to normal distribution. The selection analysis revealed 14 significant genomic regions with their average length observed as 652.06 ± 830.18 KB, ranging from 60.05 to 3,444.44 KB length. The total size of the genomic region is 9.78 Mb, with 117 total number of protein-coding regions found in the study (Table 1). Our study identified the four most significant genomic regions on BTA 6 (17567590:18290048, *q*-value = $1.76E-08$), BTA 20 (30375415:30375415, *q*-value = $8.73E-07$; 22144338:22425353, *q*-value = $2.68E-05$), and BTA 23 (39175206:39671814, *q*-value = $1.79E-08$), respectively (Table 1). The findings of gene annotation (Table 2) revealed a wide range of genes related to different types of traits, viz., growth (*CPNE4* and *RALGAP1*), survival (*MIER3*), adaptation (*GIPC2*), heat tolerance (*DNAJB4*), milk protein content (*MRPL3*, *NUDT16*, and *NEK11*), milk yield (*ADGRL4* and *PTGFR*), milk production (*RPS6KA2*, *ZNF469*, and *GHR*), milk fat secretion (*HADH*, *CYP2U1*, *SGMS2*, *SLC25A21*, and *MAP3K1*), carcass traits (*COL25A1*), sperm membrane integrity (*OSTC*), resistance to bovine tuberculosis and John's disease (*RNF144B* and *PAPSS1*), coat color (*PDGFRA*, *KIT*, and *KDR*), eye area pigmentation (*LEF1*), reproduction (*DENND1A* and *NKX2-1*), lactation persistency (*MRPS30*, *NNT*, *CCL28*, *HMGCS1*, *NIM1K*, *ZNF131*, and *CCDC152*), body height (*NHLRC1*), and mineral concentration (*FAM8A1*).

Putative Signatures of Selection on Other Chromosomes

Heat tolerance

The indigenous cattle (*B. indicus*) are best known for their heat tolerance among the tropically adapted species. The putative signals in our study were also identified on BTA 1, 3, 9, and 18. Among these *DNAJB4* is one of the functional candidate genes related with heat tolerance located on BTA 3, and this gene has a straightforward role in the cellular response during heat shock. This gene is associated with conserving the integrity of cytoskeleton, controls the protein folding, and removes the altered or misfolded proteins (Collier et al., 2008).

Coat color

Coat color in cattle is driven by complex molecular mechanisms and rendered it as a breed-specific characteristic in nature. In our study, we could identify three key genes *PDGFRA*, *KIT*, and *KDR* (Platelet-derived growth factor receptor alpha, *KIT* proto-oncogene, and Receptor tyrosine kinase insert domain receptor) located at 17-Mb regions on BTA 6. These three genes represent the cluster of tyrosine kinase receptor genes. In a study on the coat color patterns in the Nellore–Angus crossbred population, the red coat color is attributed to the three genes located in the region and is mostly associated with the *KIT* gene (Hulsman Hanna et al., 2014).

Milk and related traits

Sahiwal cattle along with other indigenous cattle are more revered in India due to their better milk attributes in terms of health perspective. Although studies on comparative milk profiles of Sahiwal and Holstein Friesian cattle suggested that milk composition of Sahiwal is slightly better than the Holstein–Friesian cattle in terms of fatty acid compositions (high unsaturated fatty acid 38.6%, low saturated fatty acid 61.4%, higher percent of monounsaturated and polyunsaturated fatty acids) (Sharma et al., 2018). The DCMS analyses in our study identified three genes, viz., Mitochondrial ribosomal protein L3 (*MRPL3*), Nudixhydrolase (*NUDT16*), and NIMA related kinase 11 (*NEK11*) located on BTA 1 at 139 Mb position. Interestingly, results of another study on weighted single-step genome-wide association analyses in Holstein and Holstein \times Jersey crossbred dairy cattle coincided with this region (Raschia et al., 2020). This study associated *MRPL3*, *NUDT16*, and *NEK11* genes with milk composition traits. *ADGRL4* is an important functional candidate gene identified on BTA 3 related to the milk yield. Other studies on signatures of selection in the crossbred cattle reported a region that coincided with our study and is associated with milk production (Li et al., 2010; Singh et al., 2020). In our study, additionally one more candidate gene located on BTA 9 and related to milk production was observed (Mustafa et al., 2018). Multi-trait meta-analysis in Nordic cattle detected several loci with pleiotropic effects on milk production and mastitis resistance (Cai et al., 2020). We identified another candidate gene Zinc finger protein (*ZNF469*), associated with milk production and mastitis resistance in cattle in the intergenic region of BTA 18. *MAP3K1* (mitogen activated protein kinase kinase1) is also known as *MEKK1* with single intronic indel codes Serine/Threonine kinase and was related to the *MAP3K* signaling pathway. It was also involved in breast cancer susceptibility in humans. Hence, it is assumed that this gene might play a role in the function of bovine mammary gland (Cuevas et al., 2006). Our analysis of selection signature could find out a candidate gene *MAP3K1* at 20 Mb on BTA 20, which is significantly associated with milk production traits in cattle. Two important candidate genes, Solute carrier family 25 member 1 (*SLC25A1*) and forkhead box A2 (*FOXA2*), are involved in the milk fat synthesis. *SLC25A1* is known to be associated with the oleic and total monounsaturated fatty acid synthesis, and this gene is actively involved in two KEGG pathways such as metabolic and n-glycan biosynthesis pathways (Ibeagha-Awemu et al., 2016). Likewise, *FOXA1* plays a role in the synthesis of cholesterol fat in the milk of cattle (Do et al., 2017). Copine7 (*CPNE7*), SPG7 matrix AAA peptidase subunit (*SPG7*), and FA complementation group A (*FANCA*) are a group of genes identified on BTA 18 and are associated with lipid metabolic, process cardiac system, and nervous system development. This region has a major pleiotropic effect in Chinese local cattle, which are diversified under adaptive selection (Xu et al., 2019).

Reproduction

Two candidate genes identified in this study (*DENND1A* and *NKX2-1*) were related to reproduction traits, while *DENND1A*

TABLE 1 | Summary of the genomic regions determined by the de-correlated composite of multiple selection signals (DCMS) in Sahiwal cattle.

Breed	N regions	Average \pm SD (KB)	Min (KB)	Max (MB)	N SNP	Total size (Mb)	N genes
Sahiwal	14	652.06 \pm 830.18	60.05	3.44	221	9.78	117

N Regions, number of segments identified using the DCMS method as significant genomic regions harboring signatures of selection; *Min*, minimum size of significant genomic regions; *Max*, maximum size of significant genomic regions; *N* SNPs, number of SNPs identified within the regions detected as signatures of selection using the DCMS method (*q*-value < 0.05); *N* genes, number of protein coding genes identified within the significant regions detected by the DCMS method.

TABLE 2 | Gene annotation for the significant genomic regions (autosomes) under putative selection identified by DCMS analyses in Sahiwal cattle.

Region (Mb)	<i>q</i> -value	Candidate gene	Trait
1:139.09–139.23	0.0058	KCNH, CPNE4, MRPL3, NUDT16, NEK11, and RF00026	Growth
1:139.49–139.81			Milk protein content
3:65.95–66.60	0.018	ADGRL4, IFI44, IFI44L, RF00568, PTGFR, GIPC2, RF00026, DNAJB4, FUBP1, and NEXN	Milk yield
			Endothelial metabolism
			Antiviral property
			Heat stress response
			Adaptation
			Reproduction
6:17.56–18.29	1.76E-08	COL25A1, RF00156, ETNPPL, OST RPL34, LEF1, HADH, CYP2U1, SGMS2, and PAPSS1	Carcass traits
			Fatty acid, lipid metabolism
			Milk fat secretion
			Eye area pigmentation
			Immune regulation
			Mycobacterium resistance paratuberculosis
6:71.23–71.84	0.001	LNK1, CHIC2, GSX2, PDGFRA, KIT, and KDR	Meat quality (meat tenderness, protein ubiquitination)
		LNK1, CHIC2, GSX2, PDGFRA, KIT, and KDR	Coat color
9:64.47–65.16	0.02	RF00099, RF000278, SYNCRIO, SNX14, and TBX18	Adaptation
9:10.21–10.22	0.01	RF00026, C9H6, PDE10A, TBXT, SFT2D1, MPC1, RPS6KA2, RNASET2, and FGFR1OP	Milk production
			Residual feed intake
11:94.29–94.36	0.03	OR5C1, OR1K1, PDCL, RC3H2, RF00579, ZBTB6, ZBTB26, RABGAP1, GPR21, STRBP, RF00026, CRB2, DENND1A, and RF00402	Reproduction
18:14.00–14.54	0.03	ZNF469, ZFPM1, ZC3H18, IL17C, CYBA, MVD, SNAI3, RNF166, CTU2, PIEZO1, CDT1, APRT, CALNS, TRAPPC2L, CBFA2T3, ACSF3, CDH15, SLC22A31, ANKRD11, SPG7, RPL13, RF00324, CPNE7, DPEP1, CHMP1A, CDK10, SPATA2L, VPS9D1, ZNF276, and FANCA	Milk yield and Mastitis resistance
			Reproduction
			Heat tolerance and Immunity
20: 21.99–2205	0.0004	RF00406, GPBP1, MIER3, SETD9, RF00003, MAP3K1, and RF00026	Milk protein and fat traits
20: 22.14–22.42	2.68E-05		Survival
20:30.37–31.55	8.73E-07	MRPS30, RF00026, NNT, PAIP1, C20H5, TMEM267, CCL28, HMGCS1, NIM1K, ZNF131, RF00302, CCDC152, and GHR	Lactation persistency, teat and udder structure,
			Mammary function and mastitis
			Milk production
21:46.46–49.91	0.04	INSM2, RALGAP1, RF00026, BRMS1L, MBIP, NKX2-1, NKX2-8, PAX9, SLC25A21, FOXA1, TTC6, SEC23A, GEMIN2, TRAPPC6B, PNN, and FBXO33	Reproduction
			Milk protein composition
			Lethal embryonic phenotype
23:39.17–39.67	1.79E-08	RF00001, RF00012, RNF144B, DEK, KDM1B, TPMT, NHLRC1, RF00026, KIF13A, NUP153, CAP2, RBM24, and STMND1	Bovine tuberculosis susceptibility
			Spermatogenesis
			Growth
			Body height
			Inflammation and immunity
			Stature

was found to be associated with a number of embryos produced by the Holstein donor cows (Jaton et al., 2018). *NKX2-1* gene produces a transcription factor that controls the activity of thyroid-specific genes and is involved in morphogenesis. In a genome-wide association study, the *NKX2-1* gene is found to be associated with calving to first service and days open traits in Canadian Dairy Holstein cattle (Nayeri et al., 2016). These

findings revealed the genetic control of the reproductive traits in the studied Sahiwal population.

Growth, survival, and adaptation

Copine4 (*CPNE4*), GIPC PDZ domain containing family member 2 (*GIPC2*), and Ral GTPase activating protein catalytic subunit alpha 1 (*RALGAP1*) are the top candidate genes located

in the peak value of the regions on BTA 1, 3, and 21, respectively, and are related to growth, survival, and adaptation, respectively. *CPNE4* is identified on BTA 1, and this region is associated with body size, muscle, and bone development in cattle (de Simoni Gouveia et al., 2014; Barbato et al., 2020). In a large multibreed genome-wide association study on milk production in dairy cattle, *MIER3* gene was located close to the most significant SNP related to survival (Raven et al., 2014). Interestingly, this gene was also identified in our study. Sahiwal is a well-known indicine breed adapted to tropical environment and involved in the development of various synthetic breeds across the world (Australian Milking Zebu, Jamaica Hope, etc.). These synthetic cattle are also suitable for adaptation and resistance to tropical diseases (Ilatsia et al., 2012; Rehman et al., 2014). Temperature–humidity index (THI), a parameter combining temperature and humidity, is generally used along with other physiological variables, viz., respiration rate and rectal temperature, for quantification of heat stress in ruminants, including cattle.

In a genome-wide association study, *GIPC2* is located close to the most significant SNP, which is found to be associated with the adaptation in Columbian cattle (De León et al., 2019). In a similar manner, this gene was identified in our study, signifying its role in the adaptation of Sahiwal cattle for which they are better known.

QTL Identification and Enrichment Analysis

The QTL identification revealed that significant genomic regions consist 54.6% of milk-type QTLs in Sahiwal cattle and other QTL types such as production, exterior, reproduction, health, and meat and carcass, which were annotated and accounted for 15.79, 14.25, 8.71, 3.1, and 3.55%, respectively (Figure 1). These QTLs were mapped to BTA 1, 3, 6, 9, 11, 18, 20, 21, and 23. The QTL enrichment analysis has shown 17 significant (FDR-corrected p -value ≤ 0.05) QTLs on chromosomes BTA 1, 3, 6, 11, 20, 21, and 23, which are associated with exterior, health, milk, production, and reproduction traits (Table 3). The top most significant QTLs were mapped on BTA 6, 20, and 23, which are associated with milk fat percentage, eye area pigmentation, lactation persistency, facial pigmentation, and bovine tuberculosis susceptibility (Figure 2).

Gene Enrichment Analysis and Prioritization of Candidate Genes

The network constructed from the Network Analyst software consisted of 131 nodes and four candidate genes, KIT, KDR, LEF1 and MAP3K1, that are associated with skin pigmentation and milk fat percent. The gene ontology terms enriched in the analysis were GO: BP (124), GO: MF (46), and GO: CC (14), which are associated with lipid synthesis and skin pigmentation.

DISCUSSION

The present study is the first comprehensive report on genomic selection signatures wherein putative genomic intervals were explored using the DCMS method and major candidate genes for different traits of interest were identified in Sahiwal cattle.

Selection signatures in Sahiwal were identified by the DCMS method, which combined the p -values from five different statistics into a new statistic, in contrast to the studies that considered the overlap in the genomic regions among the different methods. One earlier study utilized two complementary tests such as iHS and F_{ST} (Mustafa et al., 2018) to detect the loci under selection in Sahiwal cattle. However, the compound measure of DCMS gives precise and unbiased information about the genomic regions under selection by integrating the univariate statistic p -values (Grossman et al., 2010; Ma et al., 2015; Lotterhos et al., 2017). This method is helpful in accurate prioritization of candidate variants that will be useful to understand the mechanism of selection signatures that determine phenotypes in Sahiwal cattle. As a first step, we conducted IBD and PCA to check the presence of genetic clusters in the Sahiwal cattle. Our findings revealed that even if all the samples were homogeneous, they belonged to two subpopulations within the herd. Hence, we carried out the within group selection signature analysis.

Sahiwal cattle, which are distributed in the plains of Northern region of India, were under intense artificial selection for many generations. This breed is known for its potential for milk production and survives better even in harsh tropical climate. This breed could combat infectious diseases due to better immunity levels (Ilatsia et al., 2012). However, the information on the putative loci in the genome that controls these traits is not deciphered so far, which hindered our understanding regarding the mechanism of selection in these cattle. Genomic scans of Sahiwal cattle using the DCMS method captured a number of putative regions of selection associated with economic traits like growth, facial pigmentation, eye area pigmentation, milk production and composition, reproduction, body height, adaptation and survival. Our results suggested that the genome of Sahiwal cattle was under the pressure of recent ongoing selection.

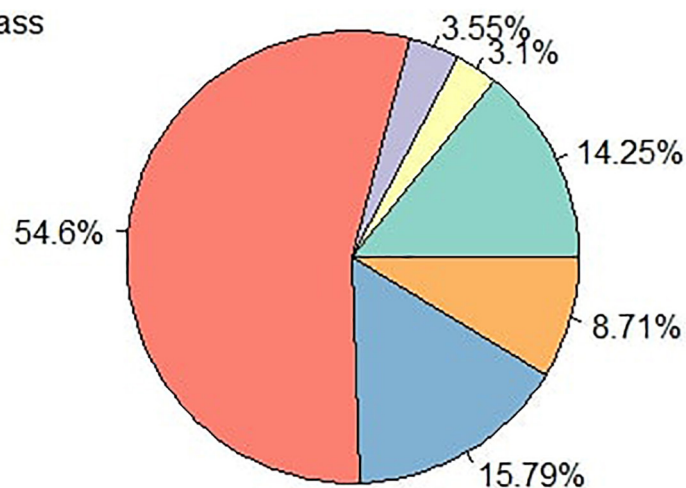
Indian cattle breeds are categorically classified into three major groups based on their utilities, such as milch, dual purpose, and draught. Sahiwal is a well-known milch breed of India. Sahiwal has a lean conformation with brown coat color and better milk production attributes. In the present study, significant genomic regions and genes related to major economic traits under selection in Sahiwal were identified.

Herein, the four most significant genomic regions were identified in Sahiwal cattle: one each on BTA 6 (17567590:18290048, q -value = $1.76E-08$) and BTA 23 (39175206:39671814, q -value = $1.79E-08$), while two were on BTA 20 (30375415:30375415, q -value = $8.73E-07$; 22144338:22425353, q -value = $2.68E-05$). The most significant candidate genes identified in our study in the region BTA 6 (17567590:18290048, q -value = $1.76E-08$) were collagen type XXV alpha 1 chain (*COL25A1*), Ethanalamine-phosphate phospho-lyase (*ETNPPL*), Oligo-saccharyltransferase complex non-catalytic subunit (*OSTC*), Ribosomal protein L34 (*RPL34*), Lymphoid enhancer binding factor 1 (*LEF1*), Hydroxy acyl-CoA dehydrogenase (*HADH*), Cytochrome P450, family2, subfamily U, polypeptide 1 (*CYP2U1*), Sphingomyelin synthase2 (*SGMS2*), and 3'-phosphoadenosine 5'-phosphosulfate synthase 1 (*PAPSS1*).

COL25A1 gene is known as collagen gene located at 17 and 18 Mb regions. This gene is associated with carcass

A

- Exterior
- Health
- Meat and Carcass
- Milk
- Production
- Reproduction



B

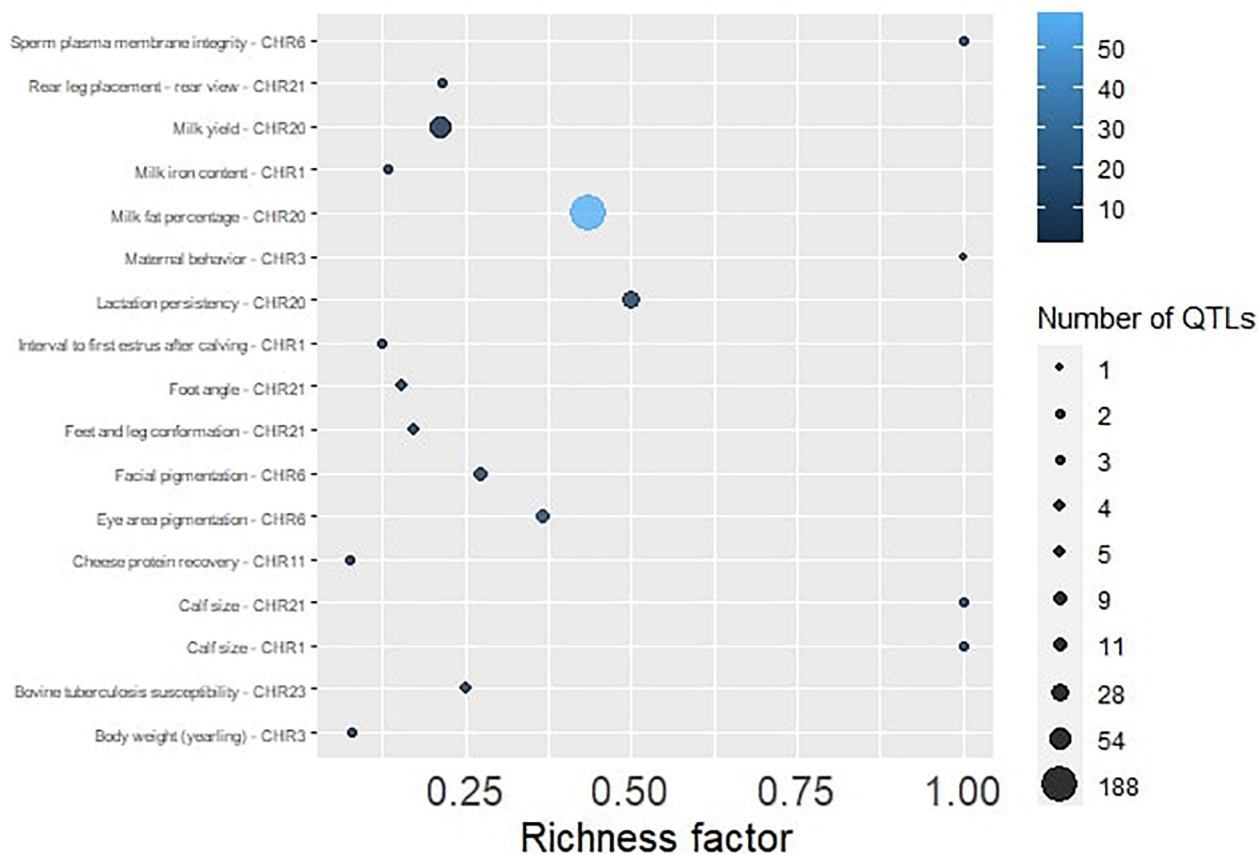


FIGURE 1 | (A) The Pie plot showing the proportion of six QTL classes annotated in the significant genomic regions in Sahiwal cattle. **(B)** The QTL enrichment analysis determined the important traits enriched in the significant genomic regions. The number of observed QTLs for a class is based on the area of the circle. The color gradient denotes the p -value scale.

TABLE 3 | The enriched QTLs annotated in the putative genomic regions of the Sahiwal cattle.

Trait	Chromosome	QTLs (number)	Annotated QTLs (number)	p-value	FDR-corrected p-value
Exterior	6	11	175	1.73E-13	1.76E-11
	6	9	175	6.19E-10	3.15E-08
	21	4	70	0.000873	0.016187
	21	4	70	0.00141	0.023975
	21	3	70	0.002188	0.034331
	3	1	11	0.003696	0.044355
Health	23	5	11	3.88E-07	1.58E-05
Milk	20	188	467	7.06E-62	1.44E-59
	11	3	8	8.21E-05	0.002095
	20	54	467	0.000466	0.009506
	1	2	14	0.002738	0.03989
Production	20	28	467	7.86E-11	5.35E-09
	3	2	11	0.003558	0.044355
Reproduction	6	3	175	2.93E-06	9.97E-05
	1	2	14	2.71E-05	0.000791
	21	2	70	0.000372	0.008422
	1	2	14	0.003119	0.042418

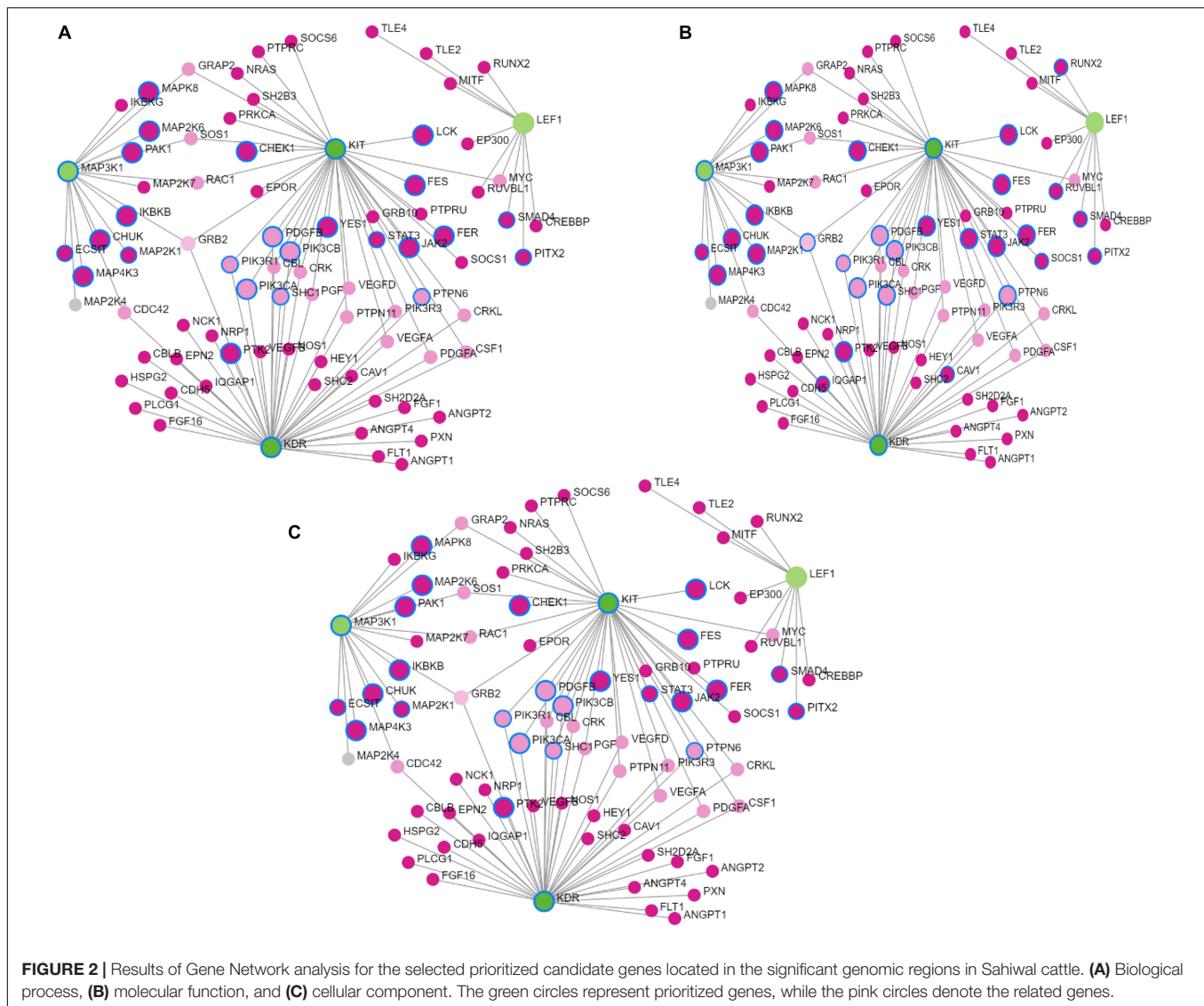
traits, and it is confirmed in a genome-wide study in Italian cattle breeds (Piedmontese, Marchigiana, Italian Holstein, Italian Brown, and Italian Pezzata Rossa breeds), as a similar peak is observed in Piedmontese beef cattle of Italy (Mancini et al., 2014). The *OSTC* gene is responsible for sperm membrane integrity in Holstein-Friesian bulls (Kamiński et al., 2016). Cattle breeds were distinguished from each other due to specific pigmentation markings.

The pigmentation particularly around the eyes in cattle is termed as ambilateral circumocular pigmentation. This pattern makes those animals less susceptible to squamous cell carcinoma of the eye, and the *LEF1* gene is known for this eye pigmentation area in cattle (Pausch et al., 2012). *HADH* (Hydroxy acyl-CoA dehydrogenase), *CYP2U1* (cytochrome P450 family 2 subfamily U member 1), and *SGMS2* (sphingomyelin synthase 2) genes are associated with fatty acid and lipid metabolism, and thus help in fat globule secretion in milk (Li et al., 2010; Grilz-seger et al., 2019). The *PAPSS1* gene, which is associated with resistance to Johne's disease, is also located in this region (Mallikarjunappa et al., 2018).

Our study could also locate another putative selection signature in Sahiwal, located on BTA 23 (39175206: 39671814, q -value = 1.79E-08) containing few important candidate genes related to bovine tuberculosis susceptibility, spermatogenesis, body height, and mineral content. The strongest candidates included Ring finger protein 144B (*RNF144B*), Thiopurine S-methyltransferase (*TPMT*), and NHL repeat containing E3 ubiquitin protein ligase 1 (*NHLRC1*). The role of *RNF144B* toward bovine tuberculosis susceptibility was also reported in Holstein-Friesian cattle (Raphaka et al., 2017). This gene is associated with the expression of Nuclear-factor-kappa-B-inhibitor alpha (NF- κ B) in human macrophages and also controls the function of various genes involved in a wide range of cellular processes like inflammation and immunity. In addition, it controls the apoptosis and cell proliferation. This protein

coding gene is also conserved in other species, especially in humans (Zhou et al., 2016). It is important to note that the gene *TPMT* located on BTA 23 at the 39-Mb region is associated with porcine infections and affects the susceptibility to parasitic burden in pigs (Gaur et al., 2014). The tropical climate favors the prevalence of parasitic infections. These gene polymorphisms could affect the growth and production performance. Thus, this candidate gene could play a role in cattle production systems. Other important genes located on BTA 23 in Sahiwal cattle were NHL repeat containing E3 ubiquitin protein ligase 1 (*NHLRC1*) and family with sequence similarity 8 member A1 (*FAM8A1*). These functional candidate genes identified in this location were known to be associated with body conformation traits like body stature and mineral content, especially magnesium (Tizioto et al., 2015; Yan et al., 2020).

Another genomic region predicted to be under putative selection was located on BTA 20 at the 30-Mb interval, and a panel of candidate genes identified in this region were Mitochondrial ribosomal protein S30 (*MRPS30*), C-C motif chemokine ligand 28 (*CCL28*), 3-hydroxy-3-methylglutaryl-CoA synthase 1 (*HMGCS1*), and Growth hormone receptor (*GHR*). *MRPS30* is also known as programmed cell death 9 (*PDCD9*) with a role in breast cancer susceptibility (Fletcher et al., 2011). This gene has a role in cell apoptosis and found to be associated with lactation persistency in cattle (Appuhamy et al., 2009), whereas the *CCL28* gene is associated with the health of mammary gland by homing and inundation of IgA antibodies during the early lactation period (Hieshima et al., 2003). The *GHR* gene affects the milk yield and the lactation process by controlling the activity of growth hormone (Moisio et al., 1998; Rahmatalla et al., 2011). Similarly, the *HMGCS1* gene is associated with lactation persistency in cattle as this gene affects the synthesis of milk cholesterol and lipid (Rikitake et al., 2001; Murray et al., 2003).



QTL Identification and Enrichment

The major fraction of QTL annotation in our study belonged to “Milk” type, which accounts for 54.6% of the total QTLs. The milk-type QTLs comprise milk protein percentage, milk yield, milk glycosylated kappa-casein percentage, milk protein yield, milk stearic acid content, milk protein content, milk margaric acid content, milk arachidic acid content, milk caprylic acid content, milk mid-infrared spectra, milk oleic acid content, milk lactose yield, milk phosphorylated alpha-S2-casein percentage, milk pentadecylic acid content, cheese protein recovery, milk *cis*-9-Eicosenoic acid, milk *trans*-10-Octadecenoic acid, and 305-days milk yield. The QTL enrichment analysis was performed to obtain the unbiased information about the significant QTLs present in the population rather than simply performing QTL annotation. The top five significant QTLs enriched are located on BTA 6, 20, and 23. The QTLs enriched include milk fat percentage, eye area pigmentation, lactation persistency, facial pigmentation, and bovine tuberculosis susceptibility. These

results are consistent with our gene annotation findings, where the genes related to these characteristics were observed in the significant genomic regions. Sahiwal cattle are managed under selective breeding in the germplasm unit with an objective to improve the milk production and composition traits. The average milk yield in Sahiwal cattle is in the range of 1,500 and 3,000 kg. However, the production performance of a few top yielders is up to 4000 kg in a single lactation under organized herds. The milk fat in Sahiwal cattle ranges from 4.6 to 5.2% and SNF ranges from 8.9 to 9.3% (Joshi et al., 2001). The present QTL enrichment analysis also revealed that milk fat QTLs were annotated ($1.44E-59$) in the population. Interestingly, the top significant QTL identified on BTA 20 in our study is related to the milk fat percentage. The coat color and pigmentation patterns were breed specific in *B. indicus* cattle. The coat color in Sahiwal is one of the desirable characteristic features. Coat color might be under selection by breeders for a longer period. The findings

of QTL enrichment are in the same direction. According to the literature, cattle breed with white color coat at the facial region are more susceptible to the squamous cell carcinoma (Guilbert and Wahid, 1948) compared to others, as they are more exposed to UV radiation.

Bovine tuberculosis is an important zoonotic disease in cattle, which causes huge economic losses globally. Srinivasan et al. (2018) examined the incidence of bovine tuberculosis in India through a meta-analysis study, which revealed that approximately 20% of cattle population were affected by this disease. Our study could find out major genes and QTLs associated with bovine tuberculosis susceptibility in individuals.

Prioritized Functional Candidate Genes

The functional enrichment analysis of the top 20 prioritized genes revealed significant gene ontology terms that include biological processes, molecular function, and cellular component related to coat color, eye area pigmentation, and fatty acid and lipid metabolism. The *KIT*, *KDR*, *LEF1*, and *MAP3K1* were the top candidate genes prioritized, which are associated with coat color and milk fat percent in Sahiwal cattle. The findings in our study are helpful in comprehending the genetic control of a number of traits in Sahiwal cattle. This information is also useful in warranting the better genomic prediction for economic traits in Sahiwal cattle. The findings of our study also pointed out that the genes mapped onto BTA 6 and 20 had pleiotropic effects and had significant associations with other relevant economic traits.

CONCLUSION

In summary, we have identified 14 significant genomic regions that are representative of putative signatures of selection in Sahiwal cattle. The methodology of DCMS was used in the computation of *p*-values from different univariate statistics into a composite signal, and the most significant regions were mapped onto BTA 6, 20, and 23. Our results are consistent with the domestication and selection history of these cattle. BTA 6 harbors the selection signatures, consisting of key candidate genes that are associated with coat color, milk fat percent, sperm membrane integrity, and carcass traits. These findings corroborate with the objectives of the genetic improvement program in the Sahiwal herd, where the purebred Sahiwal cattle with higher breeding values for milk composition and production traits are selected to produce the next-generation offspring. Similarly, two putative genomic regions at the intervals of 22 and 30 Mb on BTA 20 are found to contain genes associated with the variation in milk fat, milk yield, and lactation persistency in these cattle. Another important region on BTA 23 at the 17-Mb region harbors genes related to susceptibility to bovine tuberculosis and tolerance to parasitic infestations. The highly significant genomic regions

helped in shaping the demography and trait architecture in this breed and established it as one of the most suitable milch cattle breeds in the tropics. Prioritization analysis could identify key candidate genes that control pigmentation and milk fat percent. Our study reveals several loci that are associated with quantitative traits, disease resistance, and adaptation to tropical climate in the Sahiwal genome that might have been affected by years of selection. The present investigation highlights the importance of *B. indicus* cattle, which perform better in terms of both production and adaptation to tropical climate due to their unique genetic mechanism that is deciphered. The future looks more promising since the implementation of this genomic information in the ongoing breed improvement program might well enhance the genetic progress in Sahiwal cattle.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The animal study was reviewed and approved by Institute Animal Ethics committee, NDRI (NDRI approval 43-IAEC-18-8).

AUTHOR CONTRIBUTIONS

AM and SM had conceived the idea. SI, SN, and SM performed the entire bioinformatics analysis. SI, AM, and SN wrote the manuscript. SM done editing and proofreading. AM supervised the entire project. All authors read the manuscript and approved the final manuscript.

FUNDING

This study was supported by ICAR-National Dairy Research Institute, Karnal, and the work was carried out as part of the Ph.D. research program of the first author who was funded by the Sri Venkateswara Veterinary University (SVVU), Tirupati.

ACKNOWLEDGMENTS

We are grateful to the Director, ICAR-National Dairy Research Institute, Karnal, and Sri Venkateswara Veterinary University, Tirupati, Andhra Pradesh, for all the necessary help and support for this study.

REFERENCES

- Ajmone-Marsan, P., Garcia, J. F., and Lenstra, J. A. (2010). On the origin of cattle: how aurochs became cattle and colonized the world. *Evol. Anthropol. Issues News Rev.* 19, 148–157. doi: 10.1002/evan.20267
- Andersson, L., and Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.* 5, 202–212. doi: 10.1038/nrg1294
- Appuhamy, J. A. D. R. N., Cassell, B. G., and Cole, J. B. (2009). Phenotypic and genetic relationships of common health disorders with milk and fat yield

- persistencies from producer-recorded health data and test-day yields. *J. Dairy Sci.* 92, 1785–1795. doi: 10.3168/jds.2008-1591
- Barbato, M., Hailer, F., Upadhyay, M., Del Corvo, M., Colli, L., Negrini, R., et al. (2020). Adaptive introgression from indicine cattle into white cattle breeds from Central Italy. *Sci. Rep.* 10:1279. doi: 10.1038/s41598-020-57880-4
- Barbato, M., Orozco-terWengel, P., Tapio, M., and Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* 6:109. doi: 10.3389/fgene.2015.00109
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Cai, Z., Dusza, M., Gulbrandsen, B., Lund, M. S., and Sahana, G. (2020). Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genet. Sel. Evol.* 52:19. doi: 10.1186/s12711-020-00538-6
- Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427
- Collier, R. J., Collier, J. L., Rhoads, R. P., and Baumgard, L. H. (2008). Invited review: genes involved in the bovine heat stress response. *J. Dairy Sci.* 91, 445–454. doi: 10.3168/jds.2007-0540
- Cuevas, B. D., Winter-Vann, A. M., Johnson, N. L., and Johnson, G. L. (2006). MEKK1 controls matrix degradation and tumor cell dissemination during metastasis of polyoma middle-T driven mammary cancer. *Oncogene* 25, 4998–5010. doi: 10.1038/sj.onc.1209507
- DAHDF (2018/2019). *20th Livestock Census. Department of Animal Husbandry Dairying & Fisheries*. New Delhi: Ministry of Fisheries, Animal Husbandry & Dairying, Government of India.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330
- De León, C., Manrique, C., Martínez, R., and Rocha, J. F. (2019). Genomic association study for adaptability traits in four colombian cattle breeds. *Genet. Mol. Res.* 18:gmr18373. doi: 10.4238/gmr18373
- de Simoni Gouveia, J. J., da Silva, M. V. G. B., Paiva, S. R., and de Oliveira, S. M. P. (2014). Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37, 330–342. doi: 10.1590/S1415-47572014000300004
- Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785
- Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature* 418, 700–707. doi: 10.1038/nature01019
- Do, D. N., Bissonnette, N., Lacasse, P., Miglior, F., Sargolzaei, M., Zhao, X., et al. (2017). Genome-wide association analysis and pathways enrichment for lactation persistency in Canadian Holstein cattle. *J. Dairy Sci.* 100, 1955–1970. doi: 10.3168/jds.2016-11910
- Felius, M. (1995). *Cattle Breeds: An Encyclopedia*. Doetinchem: Misset.
- Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., et al. (2011). Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl. Cancer Inst.* 103, 425–435. doi: 10.1093/jnci/djq563
- Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S., et al. (2009). The genome response to artificial selection: a case study in dairy cattle. *PLoS One* 4:e6595. doi: 10.1371/journal.pone.0006595
- Fonseca, P. A. S., Suárez-Vega, A., Marras, G., and Cánovas, Á (2020). GALLO: an R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *Gigascience* 9:gial149. doi: 10.1093/gigascience/gial149
- Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet* 11:e1005004. doi: 10.1371/journal.pgen.1005004
- Gaur, U., Xiong, Y. Y., Luo, Q. P., Yuan, F. Y., Wu, H. Y., Qiao, M., et al. (2014). Breed-specific transcriptome response of spleen from six to eight week old piglet after infection with *Streptococcus suis* type 2. *Mol. Biol. Rep.* 41, 7865–7873. doi: 10.1007/s11033-014-3680-x
- Ghoreishifar, S. M., Eriksson, S., Johansson, A. M., Khansefid, M., Moghaddaszadeh-Ahrabi, S., Parna, N., et al. (2020). Signatures of selection reveal candidate genes involved in economic traits and cold acclimation in five Swedish cattle breeds. *Genet. Sel. Evol.* 52:52. doi: 10.1186/s12711-020-00571-5
- Grilz-seger, G., Druml, T., Neuditschko, M., Dobretsberger, M., Horna, M., and Brem, G. (2019). High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. *BMC Genomics* 20:174.
- Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327, 883–886. doi: 10.1126/science.1183863
- Guilbert, H. R., and Wahid, A. (1948). Observations on pigmentation of eyelids of Hereford cattle in relation to occurrence of ocular epitheliomas. *J. Anim. Sci.* 7, 426–429. doi: 10.2527/1948.74426x
- Hieshima, K., Ohtani, H., Shibano, M., Izawa, D., Nakayama, T., Kawasaki, Y., et al. (2003). CCL28 has dual roles in mucosal immunity as a chemokine with broad-spectrum antimicrobial activity. *J. Immunol.* 170, 1452–1461. doi: 10.4049/jimmunol.170.3.1452
- Hu, Z. L., Park, C. A., Wu, X. L., and Reecy, J. M. (2013). Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res.* 41, D871–D879. doi: 10.1093/nar/gks1150
- Hulsman Hanna, L. L., Sanders, J. O., Riley, D. G., Abbey, C. A., and Gill, C. A. (2014). Identification of a major locus interacting with MC1R and modifying black coat color in an F2 Nellore-Angus population. *Genet. Sel. Evol.* 46:4. doi: 10.1186/1297-9686-46-4
- Ibeagha-Awemu, E. M., Peters, S. O., Akwanji, K. A., Imumorin, I. G., and Zhao, X. (2016). High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Sci. Rep.* 6:31109. doi: 10.1038/srep31109
- Ilatsia, E. D., Roessler, R., Kahi, A. K., Piepho, H. P., and Zárate, V. (2012). Production objectives and breeding goals of Sahiwal cattle keepers in Kenya and implications for a breeding programme. *Trop. Anim. Health Prod.* 44, 519–530. doi: 10.1007/s11250-011-9928-8
- Jaton, C., Schenkel, F. S., Sargolzaei, M., Cánova, A., Malchiodi, F., Price, C. A., et al. (2018). Genome-wide association study and in silico functional analysis of the number of embryos produced by Holstein donors. *J. Dairy Sci.* 101, 7248–7257. doi: 10.3168/jds.2017-13848
- Jensen, J. D., Foll, M., and Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Mol. Ecol.* 25, 1–4. doi: 10.1111/mec.13493
- Joshi, B. K., Singh, A., and Gandhi, R. S. (2001). Performance evaluation, conservation and improvement of Sahiwal cattle in India. *Anim. Genet. Resour. Inf.* 31, 43–54. doi: 10.1017/s1014233900001474
- Kamiński, S., Hering, D. M., Oleński, K., Lecewicz, M., and Kordan, W. (2016). Genome-wide association study for sperm membrane integrity in frozen-thawed semen of Holstein-Friesian bulls. *Anim. Reprod. Sci.* 170, 135–140. doi: 10.1016/j.anireprosci.2016.05.002
- Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* 01, 539–559. doi: 10.1146/annurev.genom.1.1.539
- Li, H., Wang, Z., Moore, S. S., Schenkel, F. S., and Stothard, P. (2010). “Genome-wide scan for positional and functional candidate genes affecting milk production traits in Canadian Holstein Cattle”, in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany.
- Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., and Verity, B. (2017). Composite measures of selection can improve the signal-to-noise ratio in genome scans. *Methods Ecol. Evol.* 8, 717–727. doi: 10.1111/2041-210X.12774
- Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity (Edinb.)* 115, 426–436. doi: 10.1038/hdy.2015.42
- Mallikarjunappa, S., Sargolzaei, M., Brito, L. F., Meade, K. G., Karrow, N. A., and Pant, S. D. (2018). Short communication: uncovering quantitative trait loci associated with resistance to *Mycobacterium avium* ssp. paratuberculosis

- infection in Holstein cattle using a high-density single nucleotide polymorphism panel. *J. Dairy Sci.* 101, 7280–7286. doi: 10.3168/jds.2018-14388
- Mancini, G., Gargani, M., Chillemi, G., Nicolazzi, E. L., Marsan, P. A., Valentini, A., et al. (2014). Signatures of selection in five Italian cattle breeds detected by a 54K SNP panel. *Mol. Biol. Rep.* 41, 957–965. doi: 10.1007/s11033-013-2940-5
- Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. doi: 10.1038/nature16152
- Maule, J. (1990). *The Cattle of Tropics*. First. Edinburgh: University of Edinburgh Centre for Tropical Veterinary Medicine.
- Moisio, S., Elo, K., Kantanen, J., and Vilkki, J. (1998). Polymorphism within the 3' flanking region of the bovine growth hormone receptor gene. *Anim. Genet.* 29, 55–57. doi: 10.1046/j.1365-2052.1998.00254.x
- Moon, S., Kim, T. H., Lee, K. T., Kwak, W., Lee, T., Lee, S. W., et al. (2015). A genome-wide scan for signatures of directional selection in domesticated pigs. *BMC Genomics* 16:130. doi: 10.1186/s12864-015-1330-x
- Muhuyi, W. B., Lokwaleput, I., and Ole Sinkeet, S. N. (1999). Conservation and utilisation of the Sahiwal cattle in Kenya. *Anim. Genet. Resour. Inf.* 26, 35–44. doi: 10.1017/s1014233900001176
- Murray, R. K., Granner, D. K., Mayes, P. A., and Rodwell, V. W. (2003). *Harper's Illustrated Biochemistry*, 26th Edn. New York, NY: Lange Medical Books/McGraw Hill.
- Mustafa, H., Khan, W. A., Kuthu, Z. H., Eui-Soo, K., Ajmal, A., Javed, K., et al. (2018). Genome-wide survey of selection signatures in Pakistani cattle breeds. *Pak. Vet. J.* 38, 214–218. doi: 10.29261/pakvetj/2018.051
- Nayeri, S., Sargolzaei, M., Abo-Ismael, M. K., May, N., Miller, S. P., Schenkel, F., et al. (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet.* 17:75. doi: 10.1186/s12863-016-0386-1
- NDDB (2019). *Annual Report 2018-19. National Dairy Development Board, Anand-388001*. Available online at: <https://www.nddb.coop/sites/default/files/NDDB-AR-2019-ENGLISH-24022020.pdf> (accessed March 22, 2021).
- Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5269–5273. doi: 10.1073/pnas.76.10.5269
- Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. B. Biol. Sci.* 365, 185–205. doi: 10.1098/rstb.2009.0219
- Pausch, H., Wang, X., Jung, S., Krogmeier, D., Edel, C., Emmerling, R., et al. (2012). Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. *PLoS One* 7:e36346. doi: 10.1371/journal.pone.0036346
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Rahmatalla, S. A., Müller, U., Strucken, E. M., Reissmann, M., and Brockmann, G. A. (2011). The F279Y polymorphism of the GHR gene and its relation to milk production and somatic cell score in German Holstein dairy cattle. *J. Appl. Genet.* 52, 459–465. doi: 10.1007/s13353-011-0051-3
- Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2016). A meta-assembly of selection signatures in cattle. *PLoS One* 11:e0153013. doi: 10.1371/journal.pone.0153013
- Raphaka, K., Matika, O., Sánchez-Molano, E., Mrode, R., Coffey, M. P., Riggio, V., et al. (2017). Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet.* 18:27. doi: 10.1186/s12863-017-0493-7
- Raschia, M. A., Nani, J. P., Carignano, H. A., Amadio, A. F., Maizon, D. O., and Poli, M. A. (2020). Weighted single-step genome-wide association analyses for milk traits in Holstein and Holstein x Jersey crossbred dairy cattle. *Livest. Sci.* 242:104294. doi: 10.1016/j.livsci.2020.104294
- Raven, L. A., Cocks, B. G., and Hayes, B. J. (2014). Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics* 15:62. doi: 10.1186/1471-2164-15-62
- Rehman, Z., Khan, M. S., and Aslam Mirza, M. (2014). Factors affecting performance of Sahiwal cattle—a review. *J. Anim. Plant Sci.* 24, 1–12.
- Rikitake, Y., Kawashima, S., Takeshita, S., Yamashita, T., Azumi, H., Yasuhara, M., et al. (2001). Anti-oxidative properties of fluvastatin, an HMG-CoA reductase inhibitor, contribute to prevention of atherosclerosis in cholesterol-fed rabbits. *Atherosclerosis* 154, 87–96. doi: 10.1016/S0021-9150(00)00468-8
- Rosen, B., Bickhart, D., Schnabel, R., Koren, S., Elsik, C., Zimin, A., et al. (2018). Modernizing the bovine reference genome assembly. *Proc. World Congr. Genet. Appl. Livest. Prod.* 3:802.
- Sharma, R., Ahlawat, S., Aggarwal, R. A. K., Dua, A., Sharma, V., and Tantia, M. S. (2018). Comparative milk metabolite profiling for exploring superiority of indigenous Indian cow milk over exotic and crossbred counterparts. *J. Food Sci. Technol.* 55, 4232–4243. doi: 10.1007/s13197-018-3360-2
- Singh, A., Mehrotra, A., Gondro, C., Romero, A. R., da, S., Pandey, A. K., et al. (2020). Signatures of selection in composite Vrindavani Cattle of India. *Front. Genet.* 11:589496. doi: 10.3389/fgene.2020.589496
- Singh, P. K., Kumar, D., and Varma, S. K. (2005). Genetic studies and development of prediction equations in Jersey x Sahiwal and Holstein-Friesian x Sahiwal Half Breds. *Asian-Australas. J. Anim. Sci.* 18, 179–184. doi: 10.5713/ajas.2005.179
- Srinivasan, S., Easterling, L., Rimal, B., Niu, X. M., Conlan, A. J. K., Dudas, P., et al. (2018). Prevalence of Bovine tuberculosis in India: a systematic review and meta-analysis. *Transbound. Emerg. Dis.* 65, 1627–1640. doi: 10.1111/tbed.12915
- Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- Tizioto, P. C., Taylor, J. F., Decker, J. E., Gromboni, C. F., Mudadu, M. A., Schnabel, R. D., et al. (2015). Detection of quantitative trait loci for mineral content of Nelore longissimus dorsi muscle. *Genet. Sel. Evol.* 47:15. doi: 10.1186/s12711-014-0083-3
- Todorov, V., Templ, M., and Filzmoser, P. (2011). Detection of multivariate outliers in business survey data with incomplete information. *Adv. Data Anal. Classif.* 5, 37–56. doi: 10.1007/s11634-010-0075-2
- Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., et al. (2013). Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One* 8:e64280. doi: 10.1371/journal.pone.0064280
- Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th Edn. New York, NY: Springer.
- Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., and Lotterhos, K. E. (2017). minotaur: a platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Mol. Ecol. Resour.* 17, 33–43. doi: 10.1111/1755-0998.12579
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072
- Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N. Y.)* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x
- Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15, 323–354.
- Xu, L., Zhao, G., Yang, L., Zhu, B., Chen, Y., Zhang, L., et al. (2019). Genomic patterns of homozygosity in Chinese local cattle. *Sci. Rep.* 9:16977. doi: 10.1038/s41598-019-53274-3
- Yan, Z., Wang, Z., Zhang, Q., Yue, S., Yin, B., Jiang, Y., et al. (2020). Identification of whole-genome significant single nucleotide polymorphisms in candidate genes associated with body conformation traits in Chinese Holstein cattle. *Anim. Genet.* 51, 141–146. doi: 10.1111/age.12865
- Yurchenko, A. A., Daetwyler, H. D., Yudin, N., Schnabel, R. D., Vander Jagt, C. J., Soloshenko, V., et al. (2018). Scans for signatures of selection in Russian cattle breed genomes reveal new candidate genes for environmental adaptation and acclimation. *Sci. Rep.* 8:12984. doi: 10.1038/s41598-018-31304-w
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal

- component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606
- Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* 47, W234–W241. doi: 10.1093/nar/gkz240
- Zhou, Q., Jackson, T., Elhagdakhny, S., Townsend, P., Crosbie, E., and Sayan, B. (2016). Targeting the E3-ubiquitin ligase RNF144B to inhibit proliferation in oestrogen receptor negative endometrial cancer cells. *Eur. J. Cancer* 61, S168–S169. doi: 10.1016/s0959-8049(16)61596-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Illa, Mukherjee, Nath and Mukherjee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Changthangi Pashmina Goat Genome: Sequencing, Assembly, and Annotation

Basharat Bhat^{1*}, Nazir A. Ganai¹, Ashutosh Singh², Rakeeb Mir³, Syed Mudasir Ahmad¹, Sajad Majeed Zargar⁴ and Firdose Malik⁵

¹ Division of Animal Biotechnology, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar, India, ² Department of Life Science, Shiv Nadar University, Greater Noida, India, ³ Department of Biotechnology, Baba Ghulam Shah Badshah University, Rajouri, India, ⁴ Division of Plant Biotechnology, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar, India, ⁵ Division of Temperate Sericulture, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Srinagar, India

OPEN ACCESS

Edited by:

Zhe Zhang,
South China Agricultural University,
China

Reviewed by:

Lei Zhou,
China Agricultural University, China
Priscila S. N. de Oliveira,
Federal University of São Carlos,
Brazil

*Correspondence:

Basharat Bhat
bb284@snu.edu.in

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 14 April 2021

Accepted: 22 June 2021

Published: 20 July 2021

Citation:

Bhat B, Ganai NA, Singh A, Mir R,
Ahmad SM, Majeed Zargar S and
Malik F (2021) Changthangi Pashmina
Goat Genome: Sequencing,
Assembly, and Annotation.
Front. Genet. 12:695178.
doi: 10.3389/fgene.2021.695178

Pashmina goats produce the world's finest and the most costly animal fiber (Pashmina) with an average fineness of 11–13 microns and have more evolved mechanisms than any known goat breed around the globe. Despite the repute of Pashmina goat for producing the finest and most sought-after animal fiber, meager information is available in the public domain about Pashmina genomics and transcriptomics. Here we present a 2.94 GB genome sequence from a male Changthangi white Pashmina goat. We generated 294.8 GB (>100X coverage) of the whole-genome sequence using the Illumina HiSeq 2500 sequencer. All cleaned reads were mapped to the goat reference genome (2,922,813,246 bp) which covers 97.84% of the genome. The Unaligned reads were used for *de novo* assembly resulting in a total of 882 MB non-reference contigs. *De novo* assembly analysis presented in this study provides important insight into the adaptation of Pashmina goats to cold stress and helps enhance our understanding of this complex phenomenon. A comparison of the Pashmina goat genome with a wild goat genome revealed a total of 2,823 high impact single nucleotide variations and small insertions and deletions, which may be associated with the evolution of Pashmina goats. The Pashmina goat genome sequence provided in this study may improve our understanding of complex traits found in Pashmina goats, such as annual fiber cycling, defense mechanism against hypoxic, survival secret in extremely cold conditions, and adaptation to a sparse diet. In addition, the genes identified from *de novo* assembly could be utilized in differentiating Pashmina fiber from other fibers to avoid falsification at marketing practices.

Keywords: Pashmina goat, whole genome sequence, goat SNP, Pashmina fiber, cold stress

BACKGROUND

The domestic goat (*Capra hircus*) is an Asian animal distributed across all ecologies ranging from cold arid to hot humid. It serves as an important source of meat, milk, skin, fiber, and manure. Goats exhibit several traits and diseases that are similar to those of humans, which is the reason why goats are being extensively used as an animal model for biomedical research (Fulton et al., 1994). Modern domestic goats have been domesticated from *Capra aegagrus*

(Dong et al., 2015; Sheikh et al., 2016). Over centuries, different goat breeds have been established through evolution and genetic selection, exhibiting traits such as (1) coat color variations, (2) adaptability to different climatic conditions, (3) change in size, (4) development of fine fiber (Dong et al., 2015).

An important breed of extremely cold Temperate Himalayan region of India is Changthangi Pashmina goat which produces world's most sought-after natural animal fiber, **Pashmina** (Bhat B. et al., 2019). These goats are also referred as Pashmina goats or Cashmere goats. Pashmina goats are double hair coated with an outer coat of long coarse guard hair and an inner coat of shorter fine Pashmina fiber. The guard hair develops from primary hair follicles (PHFs) and Pashmina from secondary hair follicles (SHF) (Ansari-Renani et al., 2011). These goats survive in extreme climatic conditions [$+35^{\circ}\text{C}$ (short summer) and -40°C (long winter)] at an altitude of 4,000–5,500 m above mean sea level (AMSL) under cold and hypoxic conditions. To adapt to the harsh climatic conditions like cold, hot, arid, dry, and poor grazing conditions, these goats have acquired many special abilities and attributes. The growth of a double hair coat in Pashmina goats may be one of the mechanisms of protection from cold, and the result of the response triggered by the thermoregulatory center of the brain.

It is believed that the Pashmina goats were originated in the Himalayan ranges and migrated to different regions of central Asia ranging from China, Mongolia, Iran, Russia, Afghanistan, and India (Ryder, 1966). China is the leading Pashmina producers followed by Mongolia, Iran, New Zealand, and Britain. Even though India—Kashmir contributes $<1\%$ of the world's Pashmina production, because of its fineness and quality it holds a unique position in the world's Pashmina trade (Shakyawar et al., 2013). Globally Pashmina is sold under the geographical indication (GI) tract of *Cashmere*, which belongs to India—**Kashmir**. Changthangi Pashmina goats produce the world's finest and the most costly Pashmina fiber with an average fineness of 11–13 microns (Bumla et al., 2012).

Pashmina fiber is the primary source of income for the nomadic population of Ladakh (Sheikh et al., 2016). Apart from the economic value of the Pashmina goat, this goat serves as an important source of food (milk and meat), skin, and manure in the region. This goat is a safe and secure form of investment and stable means of income in the cold and arid deserts of Ladakh. These animals often provide the only practical means of utilizing vast areas of natural grasslands in the areas where crop production is uneconomical. Adapted to the harsh environmental conditions, the Changthangi goat is a unique genetic resource of the country. However, due to lack of breeding policy and population structure over a long period of time, it is suspected that inbreeding in these indigenous breeds may pose serious threat (Ganai et al., 2011).

In this study, we report the first Pashmina goat genome sequence. We generated 294.8 GB ($>100\times$ coverage) raw reads from a Changthangi Pashmina goat using Illumina HiSeq 2500 sequencer. We deciphered 2.94 GB of the Pashmina goat genome, revealing 26,687 protein-coding genes, 842 miRNAs, 188 lncRNAs, and 1879 snRNAs. We compared the Pashmina

goat genome with the wild goat genome revealed important genetic variants related to the evolution of Pashmina goats.

METHODS

Ethics Statement

This study was approved by the Institutional Animal Welfare and Ethics Committee of the Sher-e-Kashmir University of Agricultural Science and Technology of Kashmir (SKUAST-K). All experiments and methods were performed in accordance with relevant guidelines and regulations. Experimental goats were housed in SKUAST-Kashmir goat farm located in the northern Himalayas (Satakna, Ladakh) at an altitude of 5,000 m AMSL.

Sampling, Genome Sequencing, and Assembly

Genomic DNA was extracted from the blood of a 26 months old male Changthangi Pashmina goat. A whole-genome sequencing (WGS) library was prepared with the Illumina-compatible NEXTflex Rapid DNA sequencing kit (BIOO Scientific, Austin, Texas, U.S.A.) as per the manufacturer's guidelines. Genomic DNA was sheared using Covaris S2 sonicator (Covaris, Woburn, Massachusetts, USA) to generate approximate fragment size distribution from 200 to 400 bp. Here, the fragment size distribution was checked on Agilent Bioanalyzer and subsequently purified using Hiprep magnetic beads (Magbio). Purified fragments were end-repaired, adenylated, and ligated to Illumina multiplex barcode adaptors as per NEXTflex Rapid DNA sequencing kit protocol. Adapter-ligated DNA was purified and size selected using Hiprep beads. Resultant fragments were amplified for four cycles of PCR using Illumina-compatible primers provided in the NEXTflex Rapid DNA sequencing kit. The final PCR product (i.e., sequencing library) was purified with Hiprep beads, followed by a library-quality control check. Illumina compatible sequencing library was initially quantified by Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and its fragment size distribution was analyzed on Agilent TapeStation. Sequencing and base calling were performed according to the Illumina recommendations.

Using the Illumina HiSeq 2500 platform, a total of 294.8 GB (150 bp reads) high-quality data (100X coverage of the estimated genome size) were generated. The raw reads were pre-processed to remove the adapter sequences, low-quality reads, and low-quality bases filtration toward 3'-end using cutadapt program v3.1 (Martin, 2011). Filtered reads were mapped to *C. hircus* reference genome assembly ARS1 downloaded from National Center for Biotechnology Information (NCBI) using bowtie2 v2.4.2 (Langmead and Salzberg, 2012). The Unmapped paired-end reads to reference were used for contig assembly using Abyss *de novo* assembler v2.0 (Jackman et al., 2017). The complete assembly was obtained by merging the reference-assisted and *de novo* assembled consensus sequences. The completeness and correctness of genome-assembly were evaluated with the Benchmarking Universal Single-Copy Orthologs (BUSCO) program v5.1.2 (Simão et al., 2015).

Genome Annotation

The repetitive elements were identified in the final assembled draft genome using RepeatMasker v4.0.9 (Tarailo-Graovac and Chen, 2009). The transfer RNAs (tRNAs) were predicted using the tRNAscan-SE program v2 (Lowe and Chan, 2016), with default parameters for eukaryotic genomes. Ribosomal RNAs (rRNAs) were collected based on homology information from *Homo sapiens*, *Bos taurus*, *Bos mutus* and *Ovis aris* rRNAs using the BlastN program v2.11.0+ (Ye et al., 2006). Other non-coding RNA were predicted from the assembled genome with Infernal v1.1.2 (Nawrocki and Eddy, 2013) using the Rfam database v13 (Griffiths-Jones et al., 2003), with an *E*-value cutoff of 0.001. Only non-truncated CM hits detected in the first pass of the pipeline with an inclusion threshold of ≤ 0.001 and score ≥ 45 were reported. Overlapping hits with lower score hits were filtered out.

Protein-coding gene identification was performed using Augustus (Stanke and Waack, 2003) and Maker v2.31.10 (Cantarel et al., 2008) programs. A total of 26,687, protein-coding genes were identified using a 2-pass schema. The first round of gene prediction was carried out using AUGUSTUS with *H. sapiens*, *B. taurus*, *B. mutus*, and *O. aris* as reference models for the hard-masked assembled genome. The predicted gene model from AUGUSTUS was taken as input along with transcripts and the available gene model of *C. hircus* for the second round of gene prediction in the MAKER tool. The MAKER tool provides evidence-based gene modeling. After performing the gene prediction through MAKER, we selected the following filtering criteria for the selection of the final gene model.

- Predicted gene should have a start and stop codon.
- Length of the gene should be > 300 bps.
- Gene does not contain more than 1% of N's.

Predicted proteins were annotated using homology-based prediction by searching against mammalian gene sequences utilizing the BLAST program.

- At first the protein sequences were similarity searched against the UniProt Bovidae (91,402) and Caprine family (91,402) protein database using BLASTP program with an *e*-value of 0.00001 for gene ontology (GO) and annotation (Consortium, 2014; Shaik et al., 2019).
- The unmapped genes were homology searched against NCBI *C. hircus* (GCF-001704415.1) proteins (42,687) using the BLASTP program with an *E*-value cut-off of 0.00001.
- The unannotated sequences were further annotated against NCBI non-redundant (NR) database and Pfam database with default parameters. A total of 90% of predicted genes were annotated.

The predicted proteins were uploaded to the KEGG (Kyoto Encyclopedia of Genes and Genomes)—KAAS (KEGG Automatic Annotation Server) (Moriya et al., 2007) server for pathway identification using *B. taurus*, *B. mutus*, *C. hircus*, *Bos indicus* and *Ovis aries* as reference organisms.

Variant Detection and Annotation

Variant identification was done using the Genome Analysis Toolkit (GATK) v4.0.7.0 (McKenna et al., 2010) applied on

bowtie output. As recommended by the GATK practice, Picard tools v2.25.1 (Pic, 2019) were used to add read group information, mark duplicates, and index a sorted BAM file. As per the GATK pipeline following steps were performed for variant calling from genomic data; split and trim to reassign mapping quality, local realignment, InDel realignment, and BaseQualityScore recalibration. For variant discovery, HaplotypeCaller and GenotypeGVCFs were used followed by filtering variants. Identified variants were divided into different functional classes based on their genomic distribution. *C. hircus* gene annotation file was used to determine if a SNP is located within mRNA start and end positions (genic), CDS, 5'UTR, or 3'UTR. The variants identified were filtered for the coverage of more than 20 (more than 20 reads covering the position) and quality of 30 (Phred scaled quality of more than 30) to include high confidence variants. Variants were annotated using SnpEff program v4.3T (Cingolani et al., 2012). To further evaluate the biological significance of the genes with high impact variations, the pathway analysis were performed using KEGG-KAAS server (Moriya et al., 2007; Shaik et al., 2019).

RESULTS AND DISCUSSION

Genome Sequence and Annotation

We sequenced genome DNA from a 26-month-old male Changthangi Pashmina goat. High-quality DNA extracted from blood was used to construct paired-end sequencing libraries. Using the Illumina HiSeq 2500 platform, a total of 294.8 GB raw data were generated. Out of the total cleaned reads (965,981,627 reads), 80% reads were mapped to the reference genome, covering 97.84% genome. To identify genes specific to Pashmina goats, *de novo* assembly of the unaligned reads were performed. The final genome assembly was obtained by merging the reference-assisted and *de novo* assembled consensus sequences. The final gap closer was executed using GapCloser program (Simpson and Durbin, 2012) with PE-LI libraries which generated a final draft genome of 2.94 GB with an N50 value of 102582650 (**Supplementary Table 1**). Calculation of the completeness of genome assembly using BUSCO program suggested 99.3% of the assembled genome was complete (85.6% complete and single-copy, 13.7% complete and duplicated) remaining 0.7% of genome were fragmented.

A total of 18,656 contigs (**Supplementary Table 1**) were assembled from *de novo* assembly, which may be specific to the Pashmina goat breed. To interpret the biological implication of sequences extracted from *de novo* assembly, all sequences were mapped to the KEGG database using KASS servers (using cut-off FDR corrected *q*-value < 0.01). KEGG pathways analysis deduced that the sequences are predicted to be involved in metabolism pathways like amino-acids (cysteine, methionine, threonine, serine, and glycine), nucleotides metabolism, fatty acid metabolism; and signaling pathways like Calcium signaling, Notch signaling, cAMP signaling, Rap1, and PI3K-Akt signaling. Amino-acids and energy play a vital role in continuous Pashmina fiber growth and follicle initiation. Enrichment analysis suggests the potential role of cAMP, PI3K-Akt, and calcium signaling in regulating cold stress responses in Pashmina goats. Further

TABLE 1 | Repeat element statistics from Pashmina goat genome assembly.

	No. of elements	Length occupied (bp)	%age in genome
SINEs	2,045,717	296,111,061	10.07
MIRs	391,052	55,830,025	1.90
LINEs	1,316,415	746,218,268	25.37
LINE1	567,547	331,707,256	11.28
LINE2	246,431	61,914,088	2.10
L3/CR1	33,442	6,797,367	0.23
RTE	467,971	345,644,042	11.75
LTR elements	349,139	102,247,773	3.48
ERV	72,022	28,212,673	0.96
ERV-L-MaLRs	120,773	39,051,305	1.33
ERV_classI	39,090	13,144,887	0.45
ERV_classII	101,109	18,070,827	0.61
DNA elements:	278,549	55,919,821	1.90
hAT-Charlie	160,945	29,848,276	1.01
ToMar-Tigger	43,361	11,562,004	0.39
Unclassified	4,534	803,290	0.03
Small RNA	250,244	39,202,683	1.33
Satellites	137,840	149,989,331	5.10
Simple repeats	823,810	89,185,768	3.03
Low complexity:	105,341	8,887,201	0.30
Total repeats:		1,450,391,539 bp	49.30

transcriptomic or proteomic studies are required to identify specific genes and pathways mediating cold stress response in Pashmina goats.

A total of 1,450,391,539 bp (49.30% of the genome) of the Pashmina goat genome contain repetitive elements (**Table 1**) which is in accordance with the earlier studies of cattle (*B. taurus*) genome (50%) (Sequencing et al., 2012). In our study, we identified that long interspersed elements cover 25.37% of the whole genome, which is equitable to that of cattle (23%), human (21%), and horse (20%) (Sequencing et al., 2012). The percentage of short interspersed elements is 10.07% which is lower than that of humans (13%) and cattle's (18%) and greater than that of mice (8%) and horses (7%). 40.83% of the total Pashmina goat genome contains different types of interspersed repeats. Also, 8.44% of the total Pashmina genome contains small RNAs, microsatellites, and simple repeats; which should be useful in quantitative trait locus mapping or marker-assisted breeding in modulating economically important traits in specialty animal fiber like Pashmina.

A multi-way approach was carried out for protein-coding gene prediction from the final assembled genome. We have used AUGUSTUS and MAKER tools for gene prediction. The first round of gene prediction was carried out using AUGUSTUS with humans as a reference model for the hard-masked assembled genome. The predicted gene model from AUGUSTUS was taken as input along with transcripts and the available gene model of *C. hircus* for the second round of gene prediction in the MAKER tool. A total of 26,687 protein-coding genes were

TABLE 2 | Pathways affected by high impact SNPs and InDels in Pashmina goat genome.

Pathways	Genes
Cytokine-cytokine receptor interaction	BMP3, FAS, LEPR, CCR3, IL3RA, TNFRSF11A, IFNGR2, IL1A, CXCL17
ECM-receptor interaction	COL6A5, TNN, ITGB5, COL4A1
Notch signaling pathway	PTCRA, DLL3
MAPK signaling pathway	RASGRP2, CACNA1I, FAS, MAPK11, IL1A, CACNA1S
AMPK signaling pathway	ACACB, CPT1B, LEPR
Calcium signaling pathway	CACNA1I, ATP2B3, CACNA1S, MYLK
PPAR signaling pathway	CPT1B, PLIN2
mTOR signaling pathway	GRB10, ATP6V1G1, WDR59
Neuroactive ligand-receptor interaction	GRID1, CHRN3, TAAR8, LEPR, RLN3, C3
Circadian entrainment	PER3, CACNA1I
TNF signaling pathway	FAS, MAPK11
Neurotrophin signaling pathway	IRAK2, MAPK11
Jak-STAT signaling pathway	IL3RA, IFNGR2, LEPR
VEGF signaling pathway	MAPK11
Wnt signaling pathway	APC2, ROR2

identified by combining reference and *de novo* assembled genome (**Supplementary Table 2**).

Variant Detection

GATK workflow identified 14,270,872 SNVs. The transitions to transversions (Ts/Tv) ratio was 2.35 and homozygosity (4,965,247) and heterozygosity (9,306,235) percentage were 34.79 and 65.21%, respectively, both in accordance with the earlier studies of mammalian genome analysis. Of the SNPs, 35% are intronic, 63% intergenic, 0.6% were in 3' and 5' UTR regions. 206,385 SNVs were identified in coding regions with 96,092 as synonymous and 50,355 as non-synonymous variants. Distributions of SNVs and InDels in different chromosomes of the Pashmina goat genome are presented in **Supplementary Figure 1**. A total of 1,423 high-impact SNVs (start lost-103, stop gained-790, and stop lost-530) were also identified which may be associated with the evolution of the Pashmina goat (**Supplementary Table 3**). We identified 1,126,239 InDels, which consisted of 484,468 insertions and 641,871 deletions. The length distribution of identified InDels ranged from -28 to +28 bp and homozygosity (621,032) and heterozygosity (505,207) percentages were 55 and 45%, respectively. Of all InDels, 61.7 and 37% were in intergenic and intronic regions, respectively. 0.01% InDels were divided between UTR regions of genes. A total of 1,400 high-impact InDels were identified which contained 1,176 frameshift, 20 and 5 stop lost, and stop gained InDels, respectively (**Supplementary Table 4**).

Functional analysis suggest genes with high impact SNPs were involved mainly in signaling pathway like Notch, MAPK, AMPK,

Calcium, PPAR, mTOR, TNF, Neurotrophin, Jak-STAT, VEGF, and Wnt (Table 2). These signaling pathways are involved in diverse biological processes and their specific role in Pashmina goats needs to be further elucidated.

Genome-Wide Identification of lncRNAs

Long non-coding RNAs (lncRNAs) are emerging as key regulators for a myriad of biological processes. In the current lncRNA databases, most of the identified lncRNAs are derived from mice and humans (Volders et al., 2014). Several recent studies on *B. taurus* (Huang et al., 2012), *Gallus gallus* (Li et al., 2012), and *Sus scrofa* (Tang et al., 2017) have increased the information pool of lncRNAs. However, meager information is available on lncRNAs for *Caprines* species. To our knowledge, this is the first study to report genome-wide identification of lncRNAs from any goat species. In this study, we systematically identified a comprehensive list of lncRNA (Supplementary Table 5) identified from the Pashmina goat genome and their role in regulating different biological processes.

Recent studies suggest the role of lncRNAs in modulating immunity; in our study, we identified three major classes of immune regulatory lncRNAs. (1) HOX antisense intergenic RNA myeloid 1 (*HOTAIRM1*) is known to be associated with granulocytes, and is a key regulator in myeloid transcriptional regulation, by modulating *HOXA* expression in cis-configuration (Mumtaz et al., 2017). (2) *HOX* transcript antisense RNA (*HOTAIR*) is an oncogenic long non-coding RNA overexpressed in various carcinomas. It recruits various chromatin-modifying enzymes and regulates gene silencing (Bhan and Mandal, 2015). (3) Nuclear Enriched Abundant Transcript 1 (*NEAT1*) known to the innate immune response to viral infection (Mumtaz et al., 2017).

lncRNA also plays a critical role in mediating gene expression during different developmental and differentiation processes. We identified six classes of lncRNA, which are known for mediating developmental traits in different animal models. (1) *KCNQ1* overlapping transcript 1 (*KCNQ1OT1*) plays crucial role in the transcriptional silencing of the *KCNQ1* locus by regulating histone methylation. *KCNQ1OT1* gene inactivation results multiple growth defects in mice (Fatica and Bozzoni, 2014). (2) *HOXA* transcript at the distal tip (*HOTTIP*) knockdown of these genes results in retinal cell development in mice and altered limb morphology in chickens (Fatica and Bozzoni, 2014). (3) *H19* has a role in cell proliferation; *H19* limits body growth by regulating *IGF2* expression. Mice with the loss of *H19* function show an overgrowth phenotype. (4) Metastasis associated lung adenocarcinoma transcript 1 (*MALAT1*) is majorly involved in neural development, in cultured mice hippocampal neurons *MALAT1* knockdown shows decreased dendritic growth and decreased synaptic density. (5) X-inactive specific transcript (*XIST*) acts as a regulator of X-chromosome inactivating in mammals. *XIST* deletion in mice causes a loss of X-chromosome inactivation and female-specific lethality. (6) Myocardial infarction-associated transcript (*MIAT*) is associated with retinal cell fate and myocardial infarction in humans. Other identified ncRNAs (rRNA, snRNA, tRNA,

lncRNAs, and miRNAs) including FASTA sequences are listed in Supplementary Table 5.

Genome-Wide Identification of miRNAs

MicroRNAs (miRNAs) are small (22 nt long), non-coding regulatory RNAs, which can evoke post-translational repression of mRNA levels of target genes. miRNAs are not only involved in transcriptional/post-transcriptional regulation but also regulate response to environmental stresses. In this study, using high-throughput genome sequencing we identified a total of 338 mature miRNAs in the Changthangi Pashmina goat genome. The length of mature miRNAs varies from 17 to 25 nucleotides with an average of 21 nucleotides. The major class of miRNAs 91% falls within the range of 20–23 nucleotides (Supplementary Figure 2). The highest number of miRNAs are observed in the mir-1255 family followed by mir-1302, mir-544, mir-692, mir-154, mir-562, mir-663, mir-684, mir-650, and let-7.

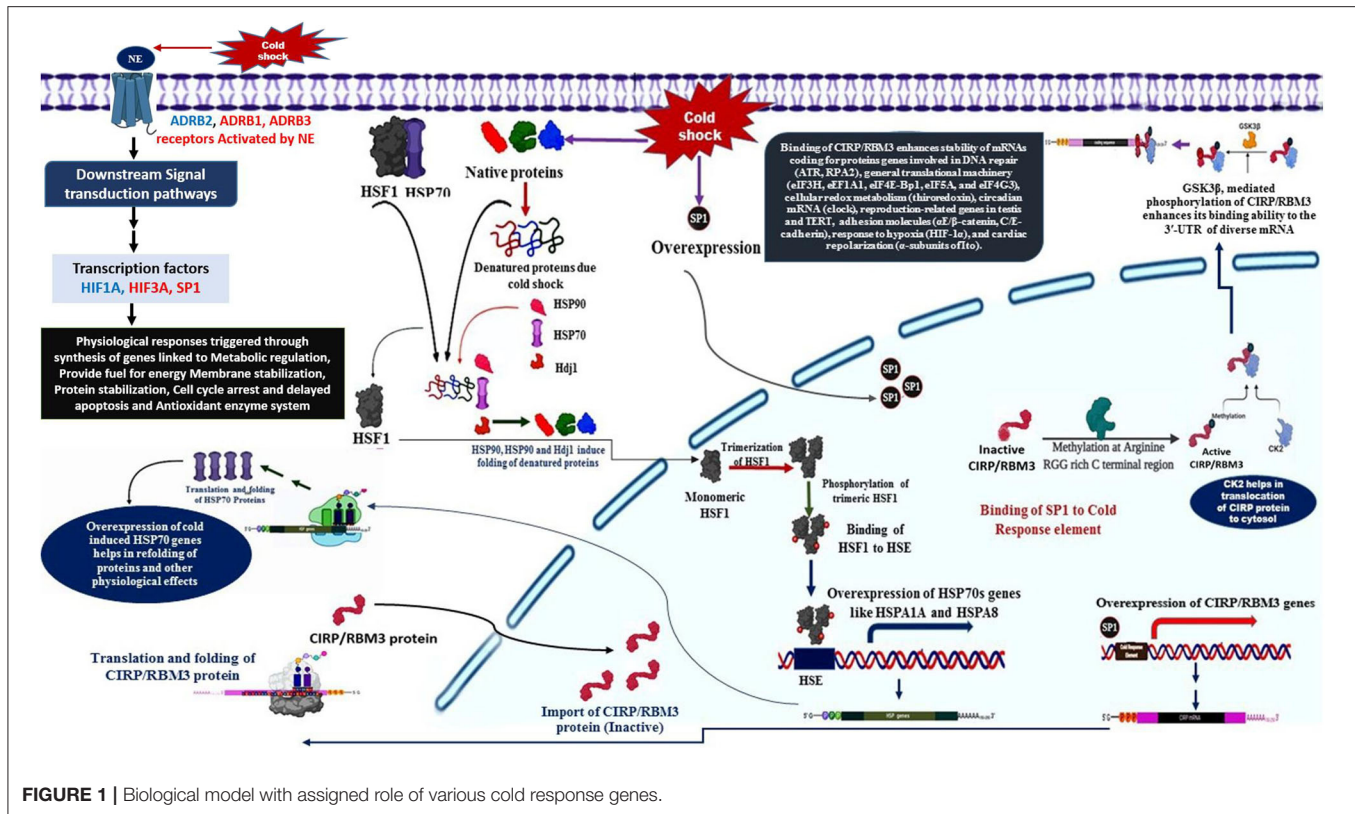
The miR-1255 commonly express in exosome and regulates TGF- β signaling pathway by interacting with SMAD4 gene (Xin et al., 2020). MicroRNA-214, miR-31 and miR-218 controls skin and hair follicle development by modulating Wnt signaling and β -catenin signaling (Mardaryev et al., 2010; Ahmed et al., 2014; Hu et al., 2020; Bhat et al., 2021). miR-128, miR-148, and miR-301 regulate key genes involved in cholesterol-lipoprotein trafficking (Wagschal et al., 2015). miR-187 family regulates key genes (*TGFB1*, *THBS1*, *ACVR18* and *BMP88*) in TGF-beta signaling pathway (Miao et al., 2016; Bhat S. A. et al., 2019). Complete annotation of miRNAs from Pashmina goat genome were listed in Supplementary Table 5.

Cold Stress Response in Pashmina Goat: A Possible Regime

Cold tolerance in Pashmina goats is an extremely complex phenomenon that is influenced by a large number of physiological, biochemical, and endocrine factors. However, the molecular mechanisms underlying the adaptation to cold stress remain largely unknown. The changes that produce a cold-hardy phenotype under cold tolerance situation would involve acclimation and acclimatization in response to low temperatures, rapid cold-hardening, and cold-induced gene expression (Hansen, 2004).

The primary response to cold stress in animals is suggested to be a neuroendocrine response, which triggers the release of catecholamine hormone, usually nor-epinephrine (NE). The cold signal in the form of NE is perceived by beta-2 Adrenergic receptor, a GPCR (Bhat et al., 2017), predominantly localized to the vascular system in comparison to $\beta 1$ and $\beta 3$. This signal perception by $\beta 2AR$ is known to activate multiple intracellular signal transduction pathways that influence various molecular, biochemical, and physiological processes (Figure 1). NE signaling through $\beta 2AR$ regulates blood pressure, heart, and respiratory rate, and body temperature (Chruscinski et al., 2001).

- The cold signal in the form of NE is perceived by β -adrenergic receptors, such as *ADRB1*, *ADRB2*, and *ADRB3*.



- Activation of $\beta 2AR$ triggers several downstream signaling cascades (DSC).
- The DSC lead to the down-regulation of *HIF3A* and also activation of *SP1* and *HIF1A*.
- The activation of *SP1* leads to the over-expression of *CIRP* and *RBM3* genes with the help of *CK2* and *GSK3 β* (Aoki et al., 2002; Yang et al., 2006; De Leeuw et al., 2007; Sumitomo et al., 2012).
- *CIRP* and *RBM3* proteins are predominantly localized in the nucleus, but can migrate to cytoplasm upon stress condition, and acts as an RNA chaperone regulating mRNA stability through its binding signature site in the 3'-UTR of its targets, which includes genes involved in DNA repair (*ATR*, *RPA2*), cellular redox metabolism (thioredoxin), adhesion molecules ($\alpha E/\beta$ -catenin, *C/E-cadherin*), circadian mRNA (clock), reproduction-related genes in testis and *TERT*, response to hypoxia (*HIF-1 α*), general translational machinery (*eIF3H*, *eEF1A1*, *eIF4E-Bp1*, *eIF5A*, and *eIF4G3*), and cardiac repolarization (α -subunits of *Ito*). In addition, *CIRP* can also be secreted into extracellular space through lysosome pathway upon stimulation by *LPS* or hypoxia/reoxygenation (Xia et al., 2012).

Another class of protective agents, the heat shock proteins (HSPs), also contribute significantly to the overwintering cold tolerance (Rinehart et al., 2007). Stress factors, like cold, induces over-expression of heat shock genes responsible for the synthesis of molecular chaperones to refold the misfolded of cold-induced

proteins. Later mechanism is triggered by the stress-induced synthesis of HSFs, which bind to heat shock elements (HSE) consisting of the pentanucleotide motif 5'-nGAAn-3' (Lis and Wu, 1993; Tissieres and Georgopoulos, 1994; Voellmy, 1994).

HSF1 and HSP70 interaction in cold-stress response

- *HSF1* is activated by unfolded proteins resulted due to cold shock, in its inactivated monomeric state *HSF1* is bound to *HSP70* (Santoro, 2000).
- The cold induced unfolded proteins are bound by molecular chaperones, such as *HSP90*, *HSP70*, and *Hsp70* (Santoro, 2000).
- The released *HSF1* is translocated to nucleus, where they undergo trimerization and phosphorylated (Santoro, 2000).
- The phosphorylated trimeric *HSF1* binds to HSE located upstream of HSP genes, resulting in transcriptional activation and synthesis of HSPs, such as *HSP70*, *HSPA1A*, and *HSPA8* genes (Banerjee et al., 2014).
- The transcripts of HSPs are shuttled to the cytosol for translation, the higher expression of HSP, in turn, regulate folding and Activation of a specific class of transcription factors called as heat shock factors (HSFs) (Åkerfelt et al., 2010).

Further, in context to the status of the translation process during cold stress condition, it is suggested that the cAMP/PKA pathway is involved in the dysregulation of translation factors like *EIF6*, *TUFM*, *EIF1AD*, *EIF2B2*, *EIF2B3*, *RPL7*, *EEF2*, *EIF3H*, and *GARS* to modulate the cold stress-responsive protein synthesis. Generally, cold shock results in protein degradation, but an

adaptive response is generated to maintain integrity as well as stability of proteins by selective expression of certain heat shock proteins (e.g., *DNAJA4*, *DNAJB12*, *DNAJC7*, *DNAJC11*, *HSPA8*, and *HSP90B1*).

Cold shock results in changes to the lipid bilayer and composition of the membrane in mammalian cells. The fatty acid and lipid metabolism especially beta-oxidation enzymes (*ACSL1*, *SLC27A1*, *ACADVL*, *HADHA*) are increased to provide fuel in the form *acetyl CoA* for energy generation as well as thermogenesis. Furthermore, HIF1A induced glycolytic enzymes like *G6PD*, *PFKFB3* are up-regulated, resulting in an increase in energy generation during the hypoxic condition. Thus, glucose metabolism and lipid metabolism are regulated in such a way to meet the requirement of energy while maintaining the stability of the lipid membrane. Cold exposure leads to increased production of reactive oxygen species (ROS) which influence HIFs activity and involves in lipid peroxidation (Quirós et al., 2016). Glutathione peroxidase (GPX), catalase (CAT), and peroxiredoxin (PRDX) are activated to neutralize the ROS effect.

Hence, the cross-talk among various signaling pathways mainly hormonal signaling (NE signaling), cAMP/PKA signaling, Src kinase-PI3K/Akt-dependent pathway, Ca²⁺ signaling, ERK1/2 signaling, p38 signaling, and ROS signaling could be responsible for the generation of a stress tolerance response in Pashmina goats by modulating the expression of specific cold stress-responsive genes. **Figure 1** illustrates the role of various cold-responsive genes.

CONCLUSIONS

Here, we report the characterization of the high altitude Pashmina goat genome for the first time. The present study on the genome and annotations may provide Pashmina breeders and other researchers with useful information regarding trait biology and their subsequent improvement. In particular, we highlight pathways that could be involved in cold-stress response and fiber cycling in Pashmina goats. The annotation of coding and non-coding genes provides, for the first time, an understanding of

the gene content in Pashmina goat, which is valuable for future studies on genes, gene structure, and functional genomics.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://www.ncbi.nlm.nih.gov/SRR7739516>.

ETHICS STATEMENT

The animal study was reviewed and approved by Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir (Certificate Number: AU/FVSc/PS-57/2058).

AUTHOR CONTRIBUTIONS

BB and NG designed the study. AS, BB, SA, and NG planned the work-flow. BB performed data analysis. BB, RM, and SM wrote the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by grants from the Indian Council of Agricultural Research, under the National Agricultural Science Fund scheme (Grant ID: NASF/GTR5006/2015-16).

ACKNOWLEDGMENTS

The bioinformatics data analysis was supported by National Agricultural Higher Education Project (NAHEP), which is duly acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.695178/full#supplementary-material>

REFERENCES

- (2019). *Picard Toolkit*. Available online at: <http://broadinstitute.github.io/picard/>
- Åkerfelt, M., Morimoto, R. I., and Sistonen, L. (2010). Heat shock factors: integrators of cell stress, development and lifespan. *Nat. Rev. Mol. Cell Biol.* 11, 545–555. doi: 10.1038/nrm2938
- Ahmed, M. I., Alam, M., Emelianov, V. U., Poterlowicz, K., Patel, A., Sharov, A. A., et al. (2014). MicroRNA-214 controls skin and hair follicle development by modulating the activity of the WNT pathway. *J. Cell Biol.* 207, 549–567. doi: 10.1083/jcb.201404001
- Ansari-Renani, H., Ebadi, Z., Moradi, S., Baghershah, H., Ansari-Renani, M., and Ameli, S. (2011). Determination of hair follicle characteristics, density and activity of Iranian cashmere goat breeds. *Small Rumin. Res.* 95, 128–132. doi: 10.1016/j.smallrumres.2010.09.013
- Aoki, K., Ishii, Y., Matsumoto, K., and Tsujimoto, M. (2002). Methylation of xenopus CIRP2 regulates its arginine-and glycine-rich region-mediated nucleocytoplasmic distribution. *Nucl. Acids Res.* 30, 5182–5192. doi: 10.1093/nar/gkf638
- Banerjee, D., Upadhyay, R. C., Chaudhary, U. B., Kumar, R., Singh, S., Polley, S., et al. (2014). Seasonal variation in expression pattern of genes under hsp70. *Cell Stress Chaperones* 19, 401–408. doi: 10.1007/s12192-013-0469-0
- Bhan, A., and Mandal, S. S. (2015). LncRNA hotair: a master regulator of chromatin dynamics and cancer. *Biochim. Biophys. Acta* 1856, 151–164. doi: 10.1016/j.bbcan.2015.07.001
- Bhat, B., Ganai, N. A., Andrabi, S. M., Shah, R. A., and Singh, A. (2017). TM-aligner: multiple sequence alignment tool for transmembrane proteins with reduced time and improved accuracy. *Sci. Rep.* 7, 1–8. doi: 10.1038/s41598-017-13083-y
- Bhat, B., Singh, A., Iqbal, Z., Kaushik, J. K., Rao, A., Ahmad, S. M., et al. (2019). Comparative transcriptome analysis reveals the genetic basis of coat color variation in pashmina goat. *Sci. Rep.* 9:6361. doi: 10.1038/s41598-019-42676-y
- Bhat, B., Yaseen, M., Singh, A., Ahmad, S. M., and Ganai, N. A. (2021). Identification of potential key genes and pathways associated with the pashmina

- fiber initiation using RNA-seq and integrated bioinformatics analysis. *Sci. Rep.* 11, 1–9. doi: 10.1038/s41598-021-81471-6
- Bhat, S. A., Ahmad, S. M., Ibeagha-Awemu, E. M., Bhat, B. A., Dar, M. A., Mumtaz, P. T., et al. (2019). Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between jersey and kashmiri cattle. *PLoS ONE* 14:e0211773. doi: 10.1371/journal.pone.0211773
- Bumla, N. A., Maria, A., Sasan, J. S., and Khateeb, A. M. (2012). Quality of Indian pashmina fibre in terms of its physico-mechanical properties. *Wayamba J. Anim. Sci.* 459–462. Available online at: <http://www.wayambajournal.com/documents/1348636207.pdf>
- Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., et al. (2008). Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907
- Chruscinski, A., Brede, M. E., Meinel, L., Lohse, M. J., Kobilka, B. K., and Hein, L. (2001). Differential distribution of β -adrenergic receptor subtypes in blood vessels of knockout mice lacking β 1- or β 2-adrenergic receptors. *Mol. Pharmacol.* 60, 955–962. doi: 10.1124/mol.60.5.955
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, Snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695
- Consortium, U. (2014). Uniprot: a hub for protein information. *Nucl. Acids Res.* 43, D204–D212. doi: 10.1093/nar/gku989
- De Leeuw, F., Zhang, T., Wauquier, C., Huez, G., Kruys, V., and Gueydan, C. (2007). The cold-inducible RNA-binding protein migrates from the nucleus to cytoplasmic stress granules by a methylation-dependent mechanism and acts as a translational repressor. *Exp. Cell Res.* 313, 4130–4144. doi: 10.1016/j.yexcr.2007.09.017
- Dong, Y., Zhang, X., Xie, M., Arefnezhad, B., Wang, Z., Wang, W., et al. (2015). Reference genome of wild goat (*Capra aegagrus*) and sequencing of goat breeds provide insight into genetic basis of goat domestication. *BMC Genomics* 16:1. doi: 10.1186/s12864-015-1606-1
- Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606
- Fulton, L. K., Clarke, M. S., and Farris Jr, H. E. (1994). The goat as a model for biomedical research and teaching. *Ilar J.* 36, 21–29. doi: 10.1093/ilar.36.2.21
- Ganai, T., Misra, S., and Sheikh, F. D., (2011). Characterization and evaluation of pashmina producing changthangi goat of Ladakh. *Indian J. Anim. Sci.* 81, 592–599. Available online at: <https://tinyurl.com/rynfknh6>
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. *Nucl. Acids Res.* 31, 439–441. doi: 10.1093/nar/gkg006
- Hansen, P. (2004). Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim. Reprod. Sci.* 82, 349–360. doi: 10.1016/j.anireprosci.2004.04.011
- Hu, S., Li, Z., Lutz, H., Huang, K., Su, T., Cores, J., et al. (2020). Dermal exosomes containing miR-218-5p promote hair regeneration by regulating I-catenin signaling. *Sci. Adv.* 6:eaba1685. doi: 10.1126/sciadv.aba1685
- Huang, W., Long, N., and Khatib, H. (2012). Genome-wide identification and initial characterization of bovine non-coding RNA s from EST data. *Anim. Genet.* 43, 674–682. doi: 10.1111/j.1365-2052.2012.02325.x
- Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., et al. (2017). Abyss 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Res.* 27, 768–777. doi: 10.1101/gr.214346.116
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9:357. doi: 10.1038/nmeth.1923
- Li, T., Wang, S., Wu, R., Zhou, X., Zhu, D., and Zhang, Y. (2012). Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. *Genomics* 99, 292–298. doi: 10.1016/j.ygeno.2012.02.003
- Lis, J., and Wu, C. (1993). Protein traffic on the heat shock promoter: parking, stalling, and trucking along. *Cell* 74, 1–4. doi: 10.1016/0092-8674(93)90286-Y
- Lowe, T. M., and Chan, P. P. (2016). tRNAscan-se on-line: integrating search and context for analysis of transfer RNA genes. *Nucl. Acids Res.* 44, W54–W57. doi: 10.1093/nar/gkw413
- Mardaryev, A. N., Ahmed, M. I., Vlahov, N. V., Fessing, M. Y., Gill, J. H., Sharov, A. A., et al. (2010). Micro-RNA-31 controls hair cycle-associated changes in gene expression programs of the skin and hair follicle. *FASEB J.* 24, 3869–3881. doi: 10.1096/fj.10-160663
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Miao, X., Luo, Q., Zhao, H., and Qin, X. (2016). Genome-wide analysis of miRNAs in the ovaries of Jining grey and Laiwu black goats to explore the regulation of fecundity. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep37983
- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucl. Acids Res.* 35(Suppl. 2), W182–W185. doi: 10.1093/nar/gkm321
- Mumtaz, P. T., Bhat, S. A., Ahmad, S. M., Dar, M. A., Ahmed, R., Urwat, U., et al. (2017). LncRNAs and immunity: watchdogs for host pathogen interactions. *Biol. Proced. Online* 19:3. doi: 10.1186/s12575-017-0052-7
- Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509
- Quirós, P. M., Mottis, A., and Auwerx, J. (2016). Mitonuclear communication in homeostasis and stress. *Nat. Rev. Mol. Cell Biol.* 17:213. doi: 10.1038/nrm.2016.23
- Rinehart, J. P., Li, A., Yocum, G. D., Robich, R. M., Hayward, S. A., and Denlinger, D. L. (2007). Up-regulation of heat shock proteins is essential for cold survival during insect diapause. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11130–11137. doi: 10.1073/pnas.0703538104
- Ryder, M. (1966). Coat structure and seasonal shedding in goats. *Anim. Sci.* 8, 289–302. doi: 10.1017/S000335610003467X
- Santoro, M. G. (2000). Heat shock factors and the control of the stress response. *Biochem. Pharmacol.* 59, 55–63. doi: 10.1016/S0006-2952(99)00299-3
- Sequencing, T. B. C. G., Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., et al. (2012). Genome sequences of wild and domestic Bactrian camels. *Nat. Commun.* 3:1202. doi: 10.1038/ncomms2192
- Shaikh, N. A., Banaganapalli, B., Elango, R., and Hakkeem, K. R. (Eds.). (2019). *Essentials of Bioinformatics Volume 1: Understanding Bioinformatics: Genes to Proteins*. Cham: Springer Publishers.
- Shakyawar, D., Raja, A., Kumar, A., Pareek, P., and Wani, S. (2013). Pashmina fibre-production, characteristics and utilization. *Indian J. Fibre Text. Res.* 38, 207–214. Available online at: <http://nopr.niscar.res.in/bitstream/123456789/19255/1/IJFTR%2038%282%29%20207-214.pdf>
- Sheikh, F., Malik, T. H., Sofi, A., Wani, S., and Ganai, T. (2016). Introduction and performance study of pashmina goats in Kargil district (non traditional area) of Jammu & Kashmir, India. *Indian J. Anim. Res.* 50, 129–132. doi: 10.18805/ijar.7487
- Sim ao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351
- Simpson, J. T., and Durbin, R. (2012). Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* 22, 549–556. doi: 10.1101/gr.126953.111
- Stanke, M., and Waack, S. (2003). Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2), ii215–ii225. doi: 10.1093/bioinformatics/btg1080
- Sumitomo, Y., Higashitsuji, H., Higashitsuji, H., Liu, Y., Fujita, T., Sakurai, T., et al. (2012). Identification of a novel enhancer that binds sp1 and contributes to induction of cold-inducible RNA-binding protein (CIRP) expression in mammalian cells. *BMC Biotechnol.* 12:72. doi: 10.1186/1472-6750-12-72
- Tang, Z., Wu, Y., Yang, Y., Yang, Y.-C. T., Wang, Z., Yuan, J., et al. (2017). Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutionary conservation in *Sus scrofa*. *Sci. Rep.* 7:43166. doi: 10.1038/srep43166
- Tarailo-Graovac, M., and Chen, N. (2009). Using repeat masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 25, 4–10. doi: 10.1002/0471250953.bi0410s25
- Tissieres, A., and Georgopoulos, C. (Eds.). (1994). “Structure and regulation of heat shock gene promoters,” in *The Biology of Heat Shock Proteins and Molecular*

- Chaperones* (Morimoto: Cold Spring Harbor Laboratory Press, Plainview), 375–393.
- Voellmy, R. (1994). Transduction of the stress signal and mechanisms of transcriptional regulation of heat shock/stress protein gene expression in higher eukaryotes. *Crit. Rev. Eukar. Gene Express.* 4, 357–401.
- Volders, P.-J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., et al. (2014). An update on lncipedia: a database for annotated human lncRNA sequences. *Nucl. Acids Res.* 43, D174–D180. doi: 10.1093/nar/gku1060
- Wagschal, A., Najafi-Shoushtari, S. H., Wang, L., Goedeke, L., Sinha, S., Andrew, S. D., et al. (2015). Genome-wide identification of micrornas regulating cholesterol and triglyceride homeostasis. *Nat. Med.* 21, 1290–1297. doi: 10.1038/nm.3980
- Xia, Z., Zheng, X., Zheng, H., Liu, X., Yang, Z., and Wang, X. (2012). Cold-inducible RNA-binding protein (CIRP) regulates target mRNA stabilization in the mouse testis. *FEBS Lett.* 586, 3299–3308. doi: 10.1016/j.febslet.2012.07.004
- Xin, Y., Wang, X., Meng, K., Ni, C., Lv, Z., and Guan, D. (2020). Identification of exosomal miR-455-5p and miR-1255a as therapeutic targets for breast cancer. *Biosci. Rep.* 40:BSR20190303. doi: 10.1042/BSR20190303
- Yang, R., Weber, D. J., and Carrier, F. (2006). Post-transcriptional regulation of thioredoxin by the stress inducible heterogenous ribonucleoprotein A18. *Nucl. Acids Res.* 34, 1224–1236. doi: 10.1093/nar/gkj519
- Ye, J., McGinnis, S., and Madden, T. L. (2006). Blast: improvements for better sequence analysis. *Nucl. Acids Res.* 34(Suppl. 2), W6–W9. doi: 10.1093/nar/gkl164

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Bhat, Ganai, Singh, Mir, Ahmad, Majeed Zargar and Malik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genes and Pathways Affecting Sheep Productivity Traits: Genetic Parameters, Genome-Wide Association Mapping, and Pathway Enrichment Analysis

Seyed Mehdi Esmaili-Fard^{1*}, Mohsen Gholizadeh¹, Seyed Hasan Hafezian¹ and Rostam Abdollahi-Arpanahi²

¹ Department of Animal Science and Fisheries, Sari Agricultural Sciences and Natural Resources University (SANRU), Sari, Iran, ² Department of Animal and Dairy Science, University of Georgia, Athens, GA, United States

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University
of Agricultural Sciences
and Technology, India

Reviewed by:

Majid Khansefid,
AgriBio, La Trobe University, Australia
Wilber Montiel,
Universidad Madero Papaloapan,
Mexico

*Correspondence:

Seyed Mehdi Esmaili-Fard
mehdi.esmailifard@gmail.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 16 May 2021

Accepted: 02 July 2021

Published: 28 July 2021

Citation:

Esmaili-Fard SM, Gholizadeh M,
Hafezian SH and
Abdollahi-Arpanahi R (2021) Genes
and Pathways Affecting Sheep
Productivity Traits: Genetic
Parameters, Genome-Wide
Association Mapping, and Pathway
Enrichment Analysis.
Front. Genet. 12:710613.
doi: 10.3389/fgene.2021.710613

Ewe productivity is a composite and maternal trait that is considered the most important economic trait in sheep meat production. The objective of this study was the application of alternative genome-wide association study (GWAS) approaches followed by gene set enrichment analysis (GSEA) on the ewes' genome to identify genes affecting pregnancy outcomes and lamb growth after parturition in Iranian Baluchi sheep. Three maternal composite traits at birth and weaning were considered. The traits were progeny birth weight, litter mean weight at birth, total litter weight at birth, progeny weaning weight, litter mean weight at weaning, and total litter weight at weaning. GWASs were performed on original phenotypes as well as on estimated breeding values. The significant SNPs associated with composite traits at birth were located within or near genes *RDX*, *FDX1*, *ARHGAP20*, *ZC3H12C*, *THBS1*, and *EPG5*. Identified genes and pathways have functions related to pregnancy, such as autophagy in the placenta, progesterone production by the placenta, placental formation, calcium ion transport, and maternal immune response. For composite traits at weaning, genes (*NR2C1*, *VEZT*, *HSD17B4*, *RSU1*, *CUBN*, *VIM*, *PRLR*, and *FTH1*) and pathways affecting feed intake and food conservation, development of mammary glands cytoskeleton structure, and production of milk components like fatty acids, proteins, and vitamin B-12, were identified. The results show that calcium ion transport during pregnancy and feeding lambs by milk after parturition can have the greatest impact on weight gain as compared to other effects of maternal origin.

Keywords: maternal genes, maternal pathways, GWAS, gene-set analysis, ewe productivity

INTRODUCTION

In sheep breeding, ewe productivity is the most important trait affecting profitability, and genetic progress in this complex trait can lead to more efficient lamb production (Hanford et al., 2003). In some countries such as Iran, where meat is the main sheep product, the productivity of a ewe flock usually has the greatest influence on profitability (Wang and Dickerson, 1991). An

increase in sheep meat production could be achieved by increasing the number and weight of lambs weaned per ewe within a specific year (Duguma et al., 2002). Ewe productivity is normally defined as the total litter weight at weaning per ewe. It is one of the most common composite traits affected by many cooperative components linked to reproduction and growth, including age at puberty, ovulation, pregnancy, parturition, lactation, mothering ability, and lamb survival and growth (Snowder and Fogarty, 2009).

Ewe productivity is often regarded as an overall measure of lamb production capacity by ewes (Bromley et al., 2001). Composite traits are a combination of growth and reproductive traits. Therefore, genetic improvement of ewe productivity is a key target in sheep breeding programs (Duguma et al., 2002). Common composite reproductive traits in sheep are total litter weight at birth and total litter weight at weaning. These parameters can be used as good coordinates for the market, where producers are paid per kilogram of sheep and not per head (Abdoli et al., 2019). Although estimates of genetic parameters have already been reported for composite traits of different Iranian sheep breeds (Abbasi et al., 2012; Abdoli et al., 2019), there are limited reports on the genes and pathways that affect these traits. To our knowledge, there is only one published report about genes and genomic regions associated with composite traits in sheep (Abdoli et al., 2019).

Due to a strict threshold used in GWASs to find significant SNPs, several poorly associated SNPs are always ignored. An alternative strategy is to add gene set analysis as a complement approach after GWAS (Dadousis et al., 2017). In this approach, a set of genes with some common functional features (e.g., being a member of a specific pathway) are identified by significant SNPs of GWAS, although with a less stringent threshold. Then, these genes are tested for over-representation in a specific pathway (Wang et al., 2011). Relevant to this context, there has been a growing interest in gene set enrichment analysis (GSEA) in dairy cattle (Han and Peñagaricano, 2016; Dadousis et al., 2017; Neupane et al., 2018).

The Baluchi sheep is the largest sheep breed in number in Iran and is well-adapted to a wide range of arid subtropical environments from the northeast to the southeast of the country (Moradband et al., 2011). Due to very limited reports on genes and pathways affecting composite traits in sheep, the objective of this study was to use GWAS and GSEA to unravel the genomic architecture underlying ewe productivity in Iranian Baluchi sheep.

MATERIALS AND METHODS

Phenotypic and Genotypic Data

The data set consisted of 1,509 ewes with 3,916 and 3,635 records of birth weight and weaning weight (sheep at 90 days of age), respectively. Progeny birth weight (PBW), litter mean weight at birth (LMWB), and total litter weight at birth (TLWB) were used as maternal composite traits at birth. Also, progeny weaning weight (PWW), litter mean weight at weaning (LMWW), and total litter weight at weaning (TLWW) were considered maternal

composite traits at weaning. The PWW trait was adjusted for birth weight according to the formula:

$$\text{PWW} = (\text{unadjusted PWW} - \text{PBW} / \text{lamb age (day) at weaning}) * 90,$$

and then used for TLWW and LMWW calculation. The LMWB and LMWW are arithmetic mean values of TLWB and TLWW traits. They were calculated for each lambing per ewe. Phenotypic correlation between traits ranged from 0.005 to 0.16. Correlation between the EBVs of traits ranged from 0.13 to 0.99. Correlation table of traits provided in **Supplementary Table 7**. The pedigree file included 4,727 animals with 178 sires, 1,509 dams, and 818 founders. Data were collected from 2004 to 2012 (9-year span) at Abbas Abad Baluchi sheep breeding station, located in Sarakhs city, Khorasan Razavi province, Iran. Descriptive statistics of the studied traits are presented in **Table 1**. The average litter size for all ewes and genotyped ewes were 1.45 and 1.56 lamb, respectively.

Genotype data of 54,241 SNP markers were provided for 91 Baluchi ewes by the animal genetics group of Sari Agriculture Science and Natural Resource University (SANRU), Iran (Gholizadeh et al., 2014). Details of the feeding and management of Baluchi sheep were reported by Gholizadeh and Ghafouri-Kesbi (2015). In sampling the animals for genotyping, two criteria were considered: the selection of unrelated animals based on pedigree information and sampling those that represented the diversity of the breed. Missing markers were imputed using Beagle 5.2 (Browning et al., 2018) on a per chromosome basis. An effective population size equal to 134 was selected based on Tahmoorespur and Sheikhlou (2011). Also, a window size of 1 Mb with an overlap of 200 kb were set. The GenABEL package (Aulchenko et al., 2007) was used for quality control in R software (R Core Team, 2021). Genotyping call rate less than 95% was applied to filter out individuals. Furthermore, SNPs with unknown genomic location, those that were monomorphic or had minor allele frequency less than 0.01, those with genotype call rates less than 93%, and SNPs that departed from the Hardy–Weinberg equilibrium (for a P -value cut-off of 1×10^{-6}) removed from the dataset and 45,342 SNPs were kept for the analysis.

TABLE 1 | Descriptive statistics of the studied traits for genotyped ewes.

Trait	N	Avg	SD	Min	Max	CV	Total N
PBW	436	4.26	0.72	2.30	6.80	0.17	3,916
TLWB	317	5.90	1.77	2.80	13.00	0.30	3,063
LMWB	317	4.40	0.75	2.70	6.80	0.17	3,063
PWW	398	20.31	4.06	9.10	34.60	0.20	3,635
TLWW	294	27.52	8.25	9.90	57.80	0.30	2,869
LMWW	294	20.97	3.98	9.90	34.60	0.19	2,869

PBW, Progeny Birth Weight; TLWB, Total Litter Weight at Birth; LMWB, Litter Mean Weight at Birth; PWW, Progeny Weaning Weight; TLWW, Total Litter Weight at Weaning; LMWW, Litter Mean Weight at Weaning; N, Number of records; Avg, Average; SD, Standard deviation; Min, Minimum value; Max, Maximum value; CV, Coefficient of variation; Total N, Total number of observations used for EBVs calculation.

Genome-Wide Association Study

Here, we regressed progenies' weights on mothers' genotypes. As ewes had several lambing records, we used a repeatability model framework for the association analyses and could consider year and ewes age effects. We also could incorporate multiple records for each ewe in the analyses. Due to the small sample size, we used two different GWAS approaches to understand whether they confirm each other or not. In the first approach, phenotypes were used as the response variable, and in the second approach, EBVs were used as the response variable.

Genome-Wide Association Mapping Using Phenotypes (pGWAS)

For the GWAS using phenotypes, the repeatability model was extended as follows:

$$y = Xb + X_{SNP}\beta_{SNP} + Zu + Wpe + e$$

where y is a vector of observations (ewe's progenies weight); b is a vector of fixed effects, including the lamb's sex, birth type, birth year, and the dam's age at lambing; u is the vector of random direct additive genetic effects, pe is the vector of permanent environmental effects, and e is the vector of random residual effects. The lamb's sex fixed effect was classified for all possible combinations and all traits except the PBW and PWW. The X , Z , and W are design matrices that relate individuals' records to their fixed effects (b), additive genetic effects (u), and permanent environmental effects (pe), respectively. X_{SNP} is the incidence matrix for the SNP markers and β_{SNP} is the regression coefficient. In this case, the random effects have multivariate Gaussian (co)variance:

$$\begin{pmatrix} u \\ pe \\ e \end{pmatrix} \sim N \left[0, \begin{pmatrix} G\sigma_u^2 & 0 & 0 \\ 0 & I_n\sigma_{pe}^2 & 0 \\ 0 & 0 & I_N\sigma_e^2 \end{pmatrix} \right]$$

where G is the genomic relationship matrix, I is an identity matrix, n is the number of genotyped individuals with reproductive records ($n = 91$) and N is the total number of observations for the genotyped individuals ($N = 294-436$, depends on the trait). We can write the extended repeatability model as follows:

$$y = Xb + X_{SNP}\beta_{SNP} + e$$

This model is the same as the model above if,

$$e \sim N(0, V) \text{ where } V = ZGZ'\sigma_u^2 + WW'\sigma_{pe}^2 + I_N\sigma_e^2.$$

In this approach, the P -value for SNP effects that occur in the original model can be calculated from the ratio of the β_{SNP} to its standard error (Wald test). An alternative approach is to use the following score test statistics that can be more computationally efficient, be asymptotically a normal standard, and one that approximates the Wald test,

$$Z = \frac{X'_{SNP}V_o^{-1}(y - X\hat{\beta})}{\sqrt{X'_{SNP}V_o^{-1}X_{SNP}}}$$

but here V_o is computed in the same way as V using a model where the SNP effects ($X_{SNP}\beta_{SNP}$) have been excluded and where $\hat{\beta}$ is computed from the model $y = Xb + X_{SNP}\beta_{SNP} + e$, assuming $e \sim N(0, V_o\sigma_e^2)$. The analyses were performed using the R package RepeatABEL (Rönnegård et al., 2016).

Genome-Wide Association Mapping Using Estimated Breeding Values (eGWAS)

The small number of available genotypes in this study can contribute to the low power of the association analysis, but the use of EBVs can increase the power to some extent as we have a better estimate of the actual genetic variance. EBVs can largely compensate for the limited number of genotypes to get reasonable estimates (Abdoli et al., 2018, 2019; Esmaeili-Fard et al., 2021). In this approach, first, we ran a pedigree-BLUP analysis using the classical repeatability animal model in the BLUPF90 software (Masuda, 2019), and breeding values of all animals (genotyped and not genotyped animals) were estimated for all traits. The lamb's sex, birth type, birth year, and dam's age at lambing were included in the model as fixed effects. Animal direct additive genetic and ewe permanent environmental effects were used as random effects. Variance components were estimated using the Restricted Maximum Likelihood (REML) approach, implemented in the AIREMLF90 software (Masuda, 2019). Accuracy (r) of EBVs were estimated as Henderson (1975) and Hayes et al. (2009):

$$r_i = \sqrt{1 - SE_i^2/\sigma_a^2}$$

where SE_i is the standard error of EBV_i , derived from the diagonal element of the inverted left-hand side in the mixed model equations (Henderson, 1975) and σ_a^2 is the additive genetic variance. Then we weight EBVs using the following formula:

$$W.EBV_i = \frac{1}{1 - r_i^2} * EBV_i$$

Finally, weighted EBVs of genotyped animals were used as the response variable and SNP genotypes were fitted as the fixed effects in a GLM model as follows,

$$W.EBVs = X_{SNP}\beta_{SNP} + e$$

where X_{SNP} is the design matrix that relates weighted EBVs to SNP genotypes and β_{SNP} is the regression coefficient. The GenABEL (Aulchenko et al., 2007) package in the R environment was used for this analysis. Due to the use of the genomic and pedigree-based relationship matrix in GWA analysis, p -values were almost non-inflated ($1.01 \leq \lambda \leq 1.07$) for all traits, however, partial inflation was corrected using the genomic control (GC) method, and all p -values were presented without any inflation ($\lambda = 1$). The CMplot¹ R package was used for drawing Manhattan plots.

We used the simpleM method (Gao et al., 2008) for multiple testing corrections. This method works based on the effective

¹<https://github.com/YinLiLin/R-CMplot>

number of independent tests. The SimpleM, first, computes the eigenvalues from the pair-wise SNP correlation matrix created with composite LD from the SNP dataset and then infers the effective number of independent tests (Meff_G) through principal component analysis. Once Meff_G is estimated, a standard Bonferroni correction is applied to control for the multiple testing. The number of independent tests calculated in our study was 8,164. Based on the average number of independent tests and the P -value cutoff 0.05, we determined 6.12×10^{-6} and 1.67×10^{-4} as genome-wide and chromosome-wide (suggestive) thresholds, respectively.

Gene Annotation

Some well-known databases including BioMart-Ensembl,² UCSC Genome Browser³, and National Center for Biotechnology Information⁴ were used along with the Ovis aries reference genome assembly (Oar_v3.1) to identify candidate genes within a window of 300 kb up and downstream of the significant SNP.

Gene-Set Enrichment Analysis

We performed gene-set enrichment analysis in three steps: (i) the assignment of SNP to the known genes, (ii) the assignment of genes to functional categories, (iii) the association analysis between each functional category and the studied traits.

For each trait, an arbitrary threshold of P -value ≤ 0.05 was applied to determine significant SNP (based on the results of the pGWAS) for enrichment analysis. The Bioconductor R package biomaRt (Durinck et al., 2005, 2009) and the Oar_v3.1 ovine reference genome assembly were used for flagging genes by significant SNP. The SNPs were assigned to genes if they were within the genomic region or 15 kb upstream or downstream of an annotated gene. Genes harboring at least one significant SNP were considered as significantly associated genes.

The Gene Ontology (GO) database (Ashburner et al., 2000) was used for defining the functional sets of genes. The GO database classifies genes into three functional categories (biological process, molecular function, and cellular component) based on their common properties. Finally, the significant association of a particular GO term with maternal composite traits was calculated using Fisher's exact test based on the hypergeometric distribution. The P -value of the g significant genes in the term was computed using the following formula,

$$P\text{-value} = 1 - \sum_{i=0}^{g-1} \frac{\binom{s}{i} \binom{m-s}{k-i}}{\binom{M}{k}}$$

where s is the total number of significant genes associated with a given maternal composite trait at birth or weaning, m is the total number of analyzed genes, and k is the total

number of genes in the term under consideration (Han and Peñagaricano, 2016). The GO enrichment analysis was performed using the R package goseq (Young et al., 2010). GO terms with more than 5 and less than 500 genes were tested. Functional categories with a nominal P -value less than or equal to 0.01 ($p \leq 0.01$) were considered as significant categories. The ggplot2 (Wickham, 2016) R package was used to visualize the GO analysis results as dot plots.

RESULTS

Estimates of Genetic Parameters

Estimates of variance components, heritability (h^2), and repeatability (R) for the traits are shown in **Table 2**. Heritability estimates of the traits ranged from 0.08 in TLWW to 0.25 in LMWB. These estimates were in the range reported by previous authors (Rosati et al., 2002; Vatankhah et al., 2008; Mokhtari et al., 2010; Rashidi et al., 2011; Mohammadi et al., 2013; Yavarifard et al., 2015; Abdoli et al., 2019). Clearly, birth weight traits show greater heritability values than weaning weight traits and indicate that maternal genes have bigger effects on the fetus than on the born lamb.

GWAS Analysis of the Composite Traits at Birth

For maternal composite traits at birth, we searched for maternal genes and pathways that influence the progeny's birth weight during pregnancy. The results of GWAS analysis are shown in a Circular Manhattan plot in **Figure 1**. After FDR correction using the simpleM method, one significant and eight suggestive SNPs were identified for ewe's reproductive traits at birth (**Table 3**). These SNPs are located on chromosomes OAR6, OAR7, OAR15, and OAR23.

Three SNPs including rs422482383, rs423274340, and rs428350449 were identified on OAR15 (19.7–20.3 Mb) which harbors four genes, *RDX*, *FDX1*, *ZC3H12C*, and *ARHGAP20*. The SNPs rs422482383 and rs423274340 were significantly associated with all three traits at birth in both GWAS approaches. The rs422482383 is located within the intron 5 of the *ARHGAP20* gene. Another identified SNP (rs427207318) on OAR15 had a suggestive association with the LMWB trait but did not contain any genes in the 300 kb flanking regions. In addition, we found three marginally suggestive SNPs on OAR7 (rs408063438, rs399067974, and rs400684038) at a distance of 30.2–35.1 Mb. Our BLAST search identified 12 genes in this region, while the SNP rs400684038 was located within the intron 8 of the *TTBK2* gene. Moreover, two SNPs (rs430043751 and rs426428997) were located on OAR23 and OAR6 as they had a suggestive association with TLWB trait.

The SNP rs430043751 on OAR23 was identified in both GWAS approaches and was related to nearly six genes in a 300 kb span, including *EPG5*, *PSTPIP2*, *ATP5F1A*, *HAUS1*, *RNF165*, and *LOXHD1*. This SNP was very close to the threshold line for PBW and LMWB traits in both GWAS approaches. The SNP

² www.ensembl.org/biomart

³ http://genome.ucsc.edu

⁴ https://www.ncbi.nlm.nih.gov

TABLE 2 | Variance components and genetic parameter estimates for composite reproductive traits in Baluchi sheep.

Trait	σ_a^2	σ_{pe}^2	σ_e^2	σ_p^2	$h^2 \pm SE$	$R \pm SE$	EBVs range*	Accuracies avg (sd)**
PBW	0.09	0.03	0.28	0.41	0.23 ± 0.04	0.31 ± 0.02	−0.52 ± 0.51	0.73 (0.06)
TLWB	0.14	0.04	0.43	0.61	0.22 ± 0.04	0.29 ± 0.02	−0.70 ± 0.73	0.73 (0.05)
LMWB	0.10	0.02	0.27	0.39	0.25 ± 0.04	0.30 ± 0.02	−0.56 ± 0.51	0.76 (0.05)
PWW	2.38	2.66	14.84	19.88	0.12 ± 0.04	0.25 ± 0.02	−1.73 ± 2.12	0.60 (0.06)
TLWW	1.76	1.86	19.60	23.23	0.08 ± 0.03	0.15 ± 0.02	−1.41 ± 1.65	0.50 (0.05)
LMWW	1.86	1.25	11.60	14.72	0.13 ± 0.04	0.21 ± 0.02	−2.01 ± 1.69	0.60 (0.06)

PBW, Progeny Birth Weight; TLWB, Total Litter Weight at Birth; LMWB, Litter Mean Weight at Birth; PWW, Progeny Weaning Weight; TLWW, Total Litter Weight at Weaning; LMWW, Litter Mean Weight at Weaning; σ_a^2 , additive genetic variance; σ_{pe}^2 , permanent environmental variance; σ_e^2 , residual variance; σ_p^2 , phenotypic variance; h^2 , heritability; R , repeatability; SE, standard error; *Estimated breeding values range in genotyped animals. **Average and standard deviation of accuracies for genotyped animals.

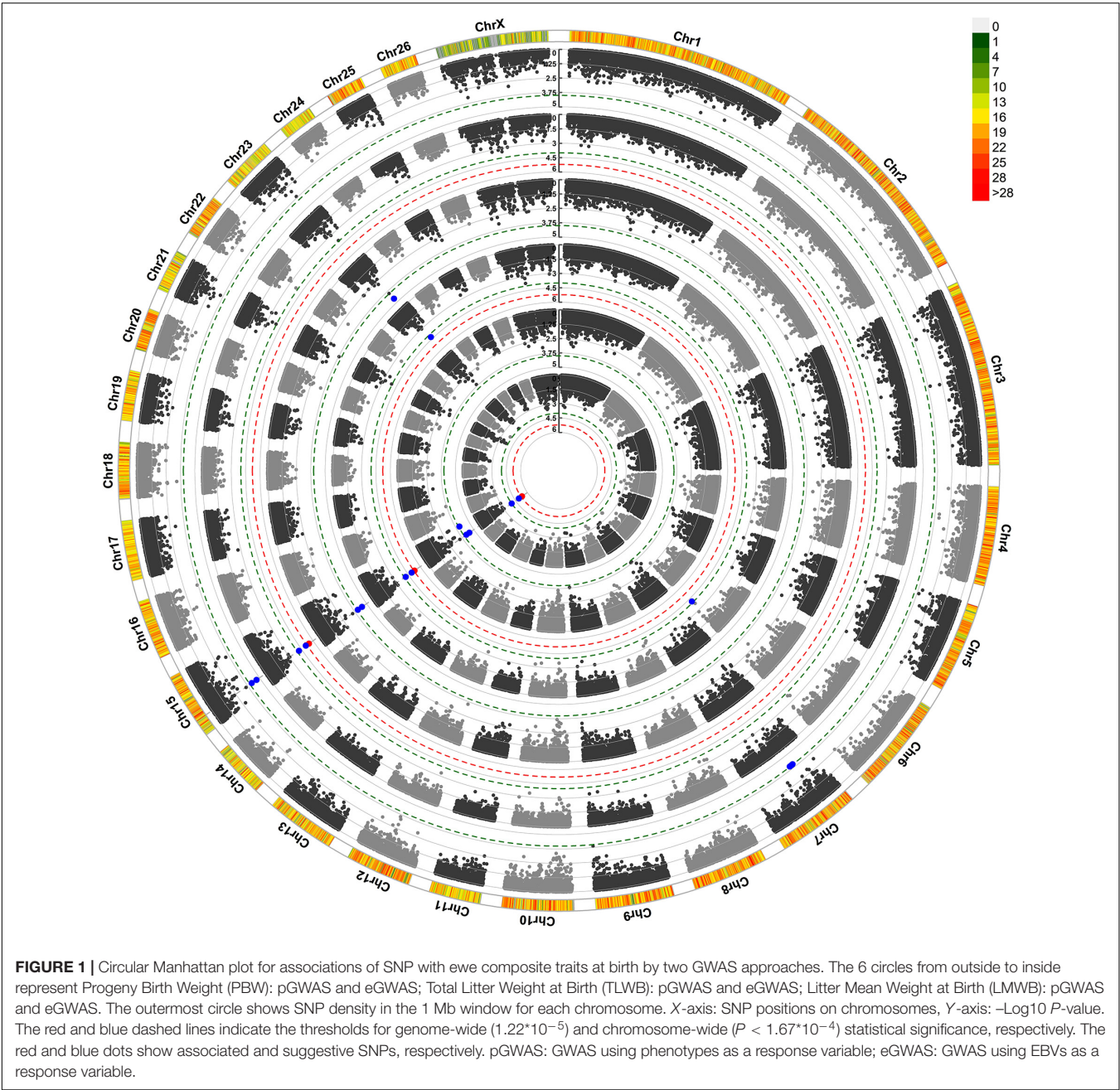


TABLE 3 | Suggested and associated SNPs with ewe composite traits at birth in Baluchi sheep.

Chr	SNP	Position	Genes in 300 kb interval	Analysis method	Adjusted P-value	Trait(s)	MAF
15	rs422482383	20,125,491	RDX, FDX1, ARHGAP20	pGWAS and eGWAS	2.06×10^{-5} 3.34×10^{-6}	PBW, TLWB and LMWB	0.11
15	rs423274340	20,304,472	ARHGAP20	pGWAS and eGWAS	6.43×10^{-5} 8.20×10^{-6}	PBW, TLWB and LMWB	0.10
15	rs428350449	19,740,174	ZC3H12C, RDX, FDX1	eGWAS	6.34×10^{-5}	PBW, TLWB and LMWB	0.12
15	rs427207318	69,550,331	Without gene	pGWAS	8.43×10^{-5}	LMWB	0.08
7	rs408063438	30,207,038	Without gene	pGWAS	9.37×10^{-5}	PBW	0.01
7	rs399067974	32,062,261	ZCRB1, THBS1, FSI1P1	pGWAS	9.37×10^{-5}	PBW	0.01
7	rs400684038	35,113,511	ZNF106, SNAP23, LRRC57, HAUS2, STARD9, CDAN1, TBKB2 , UBR1, TMEM62	pGWAS	9.37×10^{-5}	PBW	0.01
23	rs430043751	46,167,308	EPG5, PSTPIP2, ATP5F1A, HAUS1, RNF165, LOXHD1	pGWAS and eGWAS	2.52×10^{-5} 5.21×10^{-5}	TLWB	0.06
6	rs426428997	109,147,722	Without gene	eGWAS	7.20×10^{-5}	TLWB	0.05

Chr, Chromosome number; pGWAS, GWAS using phenotypes as response variable; eGWAS, GWAS using EBVs as response variable; PBW, Progeny Birth Weight; TLWB, Total Litter Weight at Birth; LMWB, Litter Mean Weight at Birth; MAF, Minor allele frequency. P-values are presented only for the first trait in the Trait(s) column. P-values are adjusted based on the Genomic Control value. Genes with boldface indicate that the significant SNP is located within the gene. SNPs with boldface are significantly associated with the genes but otherwise remain as suggestive SNPs. The regression coefficient of each SNP is provided in **Supplementary Table 14**.

rs426428997 on the OAR6 did not contain any genes in the searched region.

GWAS Analysis of the Composite Traits at Weaning

For maternal composite traits at weaning, we looked for maternal genes and pathways that influence the progeny's weaning weight. The circular Manhattan plot shows associations of SNP markers with traits for both GWAS approaches (**Figure 2**). After FDR correction (0.05) using the simpleM method, a total of 11 SNPs were significantly and suggestively associated with the maternal composite traits at weaning (**Table 4**). These SNPs were located on OAR2, OAR3, OAR5, OAR7, OAR13, OAR16, and OAR25. The results of the two GWAS approaches showed similar profiles with one common significant SNPs on OAR3 (rs428404187).

The most significant SNP (rs428404187, $p = 4.72 \times 10^{-6}$) was located on OAR3 (131.2 Mb) and was significant for the three composite traits at weaning in the pGWAS approach. Besides, this SNP had a suggestive association with LMWW in the eGWAS approach and was found to be located within the intron 1 of the *VEZT* gene. For PWW and TLWW traits, there were no significant or suggestive SNPs using eGWAS. Another common suggestive SNP, rs398620273, and was associated with three composite traits at weaning. This SNP located within the *DTWD2* gene (intron 2) on OAR5. SNP rs412011189 on OAR2 had a suggestive association with PWW and LMWW and was close to the *EIPRI* gene. In addition, we found five suggestively associated SNPs with PWW, including, rs411656768 and rs403459195 on OAR2, rs430218107 and rs419540936 on OAR7, and rs401393221 on OAR13. The SNPs on OAR7 harbor 24 genes in a 300 Kb span, seven of which were RNase genes. SNP on OAR13 was located in the *RSU1* gene (intron 2), while the *VIM* gene is located downstream of this SNP at a

distance of 4.2 kb. Three suggestive SNPs were found to be related to LMWW and were identified on OAR3 (rs404069303), OAR16 (rs409558668), and OAR25 (rs405045517). The SNP on OAR16 was near seven genes in a 300 kb span (*PRLR*, *AGXT2*, *DNAJC21*, *BRIX1*, *RAD1*, *TTC23L*, and *RAI14*). This SNP was located in the ovine gene *TTC23L* (intron 2). Additionally, the *Prolactin receptor* (*PRLR*) gene was located close to this SNP. Another suggestive SNP (rs405045517) was located on OAR25 and harbored three genes (*CDK1*, *FTH1*, and *RHOBTB1*) in the searched region. This SNP was located within the *RHOBTB1* gene (intron 4). Also, the *FTH1* gene was found to be located close to this SNP.

Gene-Set Enrichment Analysis

The results of GWAS were complemented with gene set enrichment analysis using the GO database. In total, 23,462 of the 45,342 SNPs being tested in the GWAS, were located within or 15 kb upstream or downstream of 15,815 annotated genes in the Oar.v3.1 ovine genome assembly. On average, 1,310 out of the 15,815 genes (ranging from 1,291 for LMWW to 1,351 for TLWW) contained at least one significant SNP ($P\text{-value} \leq 0.05$) and were defined as significantly associated with maternal composite traits. GO terms with a nominal $P\text{-value} \leq 0.01$ were reported as significant terms. GWAS results using direct phenotypes (pGWAS) were used for analysis and each trait was analyzed separately.

Gene-Set Enrichment Analysis of the Composite Traits at Birth

Figure 3 shows a set of GO terms that were significantly ($P \leq 0.01$) enriched with genes associated with maternal composite traits at birth. Several GO terms related to the neural system, showed an overrepresentation of significant

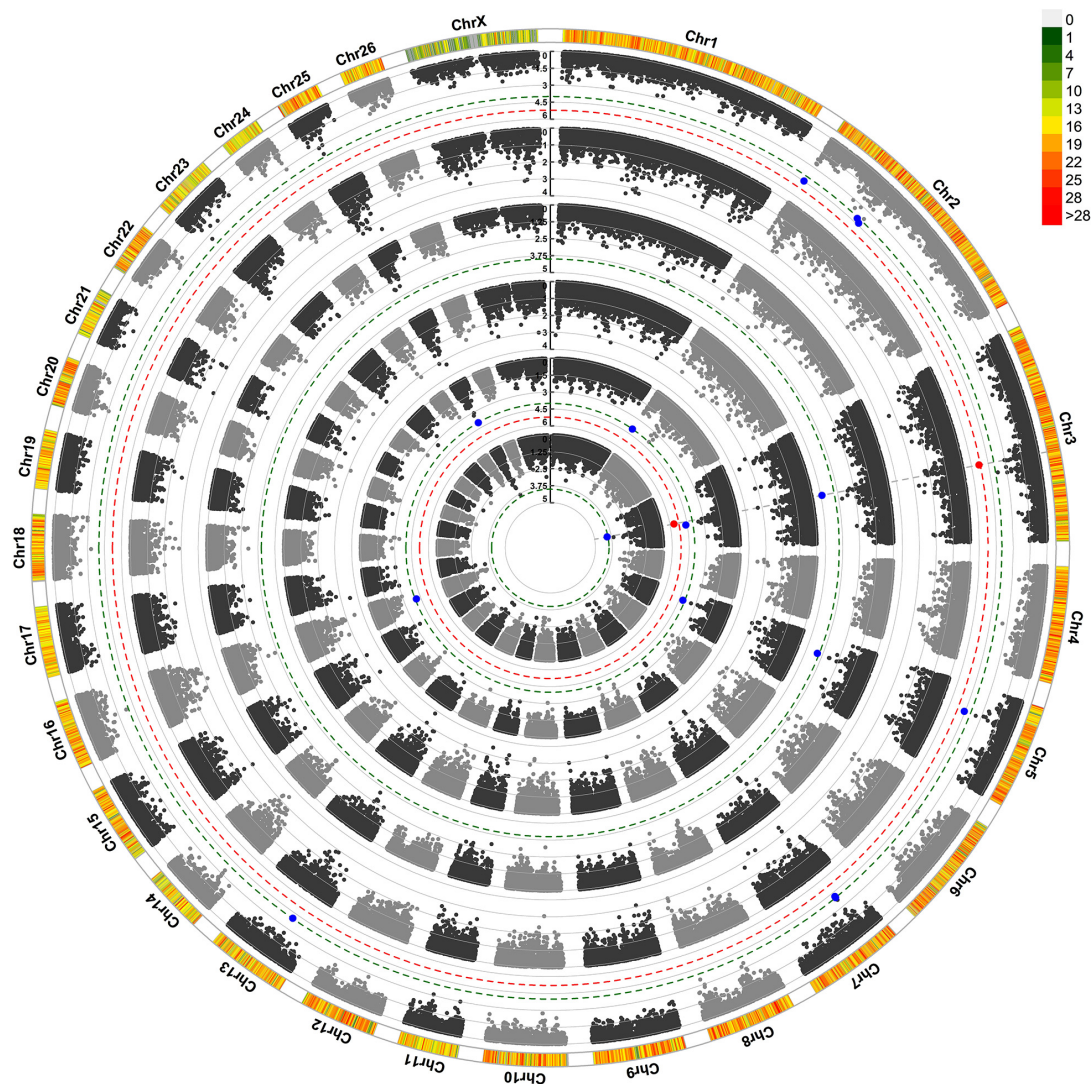


FIGURE 2 | Circular Manhattan plot for associations of SNP with ewe composite traits at weaning by two GWAS approaches. The 4 circles, from outside to inside, represent Progeny Weaning Weight (PWW): pGWAS and eGWAS; Total Litter Weight at Weaning (TLWW): pGWAS and eGWAS; Litter Mean Weight at Weaning (LMWW): pGWAS and eGWAS. The outermost circle shows SNP density in a 1 Mb window for each chromosome. X-axis: SNP positions on chromosomes, Y-axis: $-\log_{10} P$ -value. The red and blue dashed lines indicate the thresholds for genome-wide (1.22×10^{-5}) and chromosome-wide ($P < 1.67 \times 10^{-4}$) statistical significance, respectively. The red and blue dots show associated and suggestive SNPs, respectively. pGWAS: GWAS using phenotypes as response variable; eGWAS: GWAS using EBVs as a response variable.

genes, including *postsynaptic density* (GO:0014069), *Schaffer collateral-CA1 synapse* (GO:0098685), *glutamatergic synapse* (GO:0098978), *synapse* (GO:0045202), *neuron projection development* (GO:0031175), *neurogenesis* (GO:0022008), and many other terms that were not included in **Figure 3** (see **Supplementary Tables 1–3**). The *calcium ion transport* (GO:0000045) was associated with all composite traits at birth. Furthermore, *calcium channel inhibitor activity* (GO:0019855) showed an overrepresentation of significant genes associated with TLWB. Many GO terms related to the immune system also showed significant enrichment of genes associated with composite traits at birth, including *cellular response to chemokine* (GO:1990869), *positive regulation of T-helper 1 type immune*

response (GO:0002827), *positive regulation of interleukin-12 production* (GO:0032735), and *positive regulation of T cell activation* (GO:0050870). Several significant GO terms were related to the signaling process. In particular, *SMAD protein signal transduction* (GO:0060395), *negative regulation of Notch signaling pathway* (GO:0045746), and *regulation of NIK/NF-kappaB signaling* (GO:1901222) showed an overrepresentation of significant genes. In addition, we identified *cell adhesion* (GO:0007155) and *metallopeptidase activity* (GO:0008237) GO terms as significant processes that were associated with the composite traits at birth. Several other GO terms were also significant in terms of composite traits at birth. The full list is provided in **Supplementary Tables 1–3**.

TABLE 4 | Suggested and associated SNPs with ewe composite traits at weaning in Baluchi sheep.

Chr	SNP	Position	Genes in 300 kb interval	Analysis method	Adjusted <i>P</i> -value	Trait(s)	MAF
3	rs428404187	131255497	NDUFA12, NR2C1, FGD6, VEZT , MIR331, METAP2	pGWAS and eGWAS	4.72×10^{-6} 8.60×10^{-5}	PWW, TLWW, and LMWW	0.03
5	rs398620273	32383306	HSD17B4, DMXL1, DTWD2	pGWAS	2.72×10^{-5}	PWW, TLWW, and LMWW	0.28
2	rs412011189	1426911	EIPR1	pGWAS	3.39×10^{-5}	PWW and LMWW	0.2
2	rs411656768	81968762	NFIB	pGWAS	4.74×10^{-5}	PWW	0.11
7	rs430218107	23778602	EDDM3B, ANG1, RNASE1, RNASE6, RNASE4, ANG2, UQCRCF1, RNASE12, RNASE11, RNASE9, RNASE10, PIP4P1, APEX1, OSGEP, KLHL33, TEP1, PARP2, RPPH1, SNORA79B, CCNB1IP1, TTC5, OR11H4, OR11H7, OR11H6	pGWAS	7.42×10^{-5}	PWW	0.42
	rs419540936	23939664		pGWAS	9.79×10^{-5}	PWW	0.29
2	rs403459195	77075145	RPLP0	pGWAS	7.88×10^{-5}	PWW	0.26
13	rs401393221	30320719	PTER, C1ql3, RSU1 , CUBN, VIM	pGWAS	9.50×10^{-5}	PWW	0.29
3	rs404069303	143726957	SNORA62	pGWAS	1.90×10^{-5}	LMWW	0.04
16	rs409558668	39225407	PRLR, AGXT2, DNAJC21, BRX1, RAD1, TTC23L , RAI14	pGWAS	7.95×10^{-5}	LMWW	0.23
25	rs405045517	16553906	CDK1, FTH1, RHOBTB1	pGWAS	9.89×10^{-5}	LMWW	0.09

Chr, Chromosome number; pGWAS, GWAS using phenotype; eGWAS, GWAS using EBVs; PWW, Progeny Weaning Weight; TLWW, Total Litter Weight at Weaning; LMWW, Litter Mean Weight at Weaning; MAF, Minor allele frequency. *P*-values are presented only for the first trait in the Trait(s) column. *P*-values are adjusted based on the Genomic Control value. Genes with boldface indicate that significant SNPs are located within the gene. SNPs with boldface are significantly associated with genes but otherwise are suggestive SNPs. The regression coefficient of each SNP is provided in **Supplementary Table 15**.

Gene-Set Enrichment Analysis of the Composite Traits at Weaning

Figure 4 shows a set of GO terms that were significantly ($p \leq 0.01$) enriched by genes associated with weaning traits. The *Filopodium* (GO:0030175) term was significantly associated with the composite traits at weaning. Moreover, multiple GO terms were linked to protein metabolism, including *protein catabolic process* (GO:0030163), *positive regulation of intracellular protein transport* (GO:0090316), *protein processing* (GO:0016485), and *protein localization to plasma membrane* (GO:0072659). Several GO terms related to GTPase activity were recognized as significant. Among these, *GTPase activator activity* (GO:0005096) showed an overrepresentation of significant genes associated with all composite traits at weaning. Many significant GO terms were related to ion transport and homeostasis and channel activity, including *cellular calcium ion homeostasis* (GO:0006874), *ion channel activity* (GO:0005216), *ion transmembrane transport* (GO:0034220), and *ion transport* (GO:0006811). On the other hand, many GO terms that were related to the metabolism of lipids, cholesterol, and fatty acids showed an overrepresentation of genes associated with the traits

at weaning, including *phospholipid translocation* (GO:0045332), *lipid phosphorylation* (GO:0046834), *cholesterol homeostasis* (GO:0042632), and *fatty acid beta-oxidation* (GO:0006635). In addition, several terms were related to cell proliferation (GO:0008283 and GO:0042127), gene expression (GO:0010628), cell adhesion (GO:0098609), cell junction GO:0005911), Protein kinase activity (GO:0016301), and phosphorylation (GO:0016310). The full list of associated terms with weaning weight traits is provided in **Supplementary Tables 4–6**.

DISCUSSION

GWAS and GSEA of Composite Traits at Birth

Here, we performed a whole-genome scan on Iranian Baluchi sheep for six maternal composite traits. We regressed lambs' weights at birth and weaning on ewes' genotype and tried to identify gene variants (or regions) in the genome of ewes that affect pregnancy outcome (birth weights) and weaning weights of the lambs. To our knowledge, this is the second GWAS on

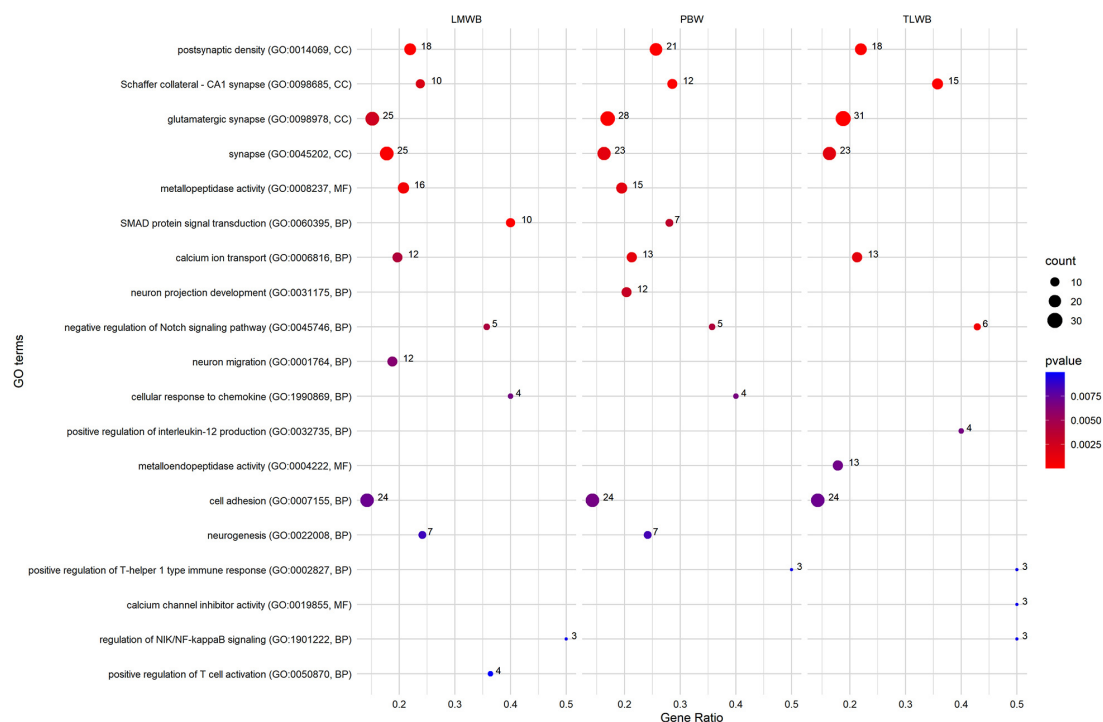


FIGURE 3 | Most related Gene Ontology (GO) terms were significantly ($p \leq 0.01$) enriched using genes associated with maternal composite traits at birth. PBW, Progeny Birth Weight; TLWB, Total Litter Weight at Birth; LMWB, Litter Mean Weight at Birth. Gene Ratio: Number of the significant genes in the category/Number of all genes in the category. Complete associated GO terms with these traits are provided in **Supplementary Tables 1–3**.

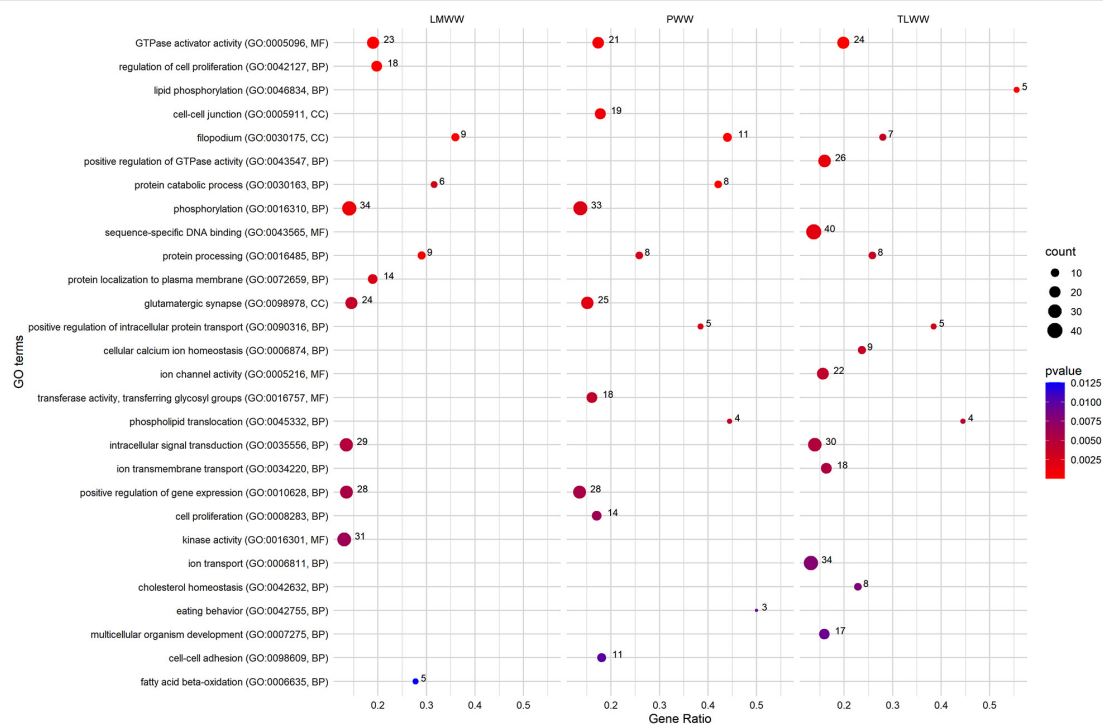
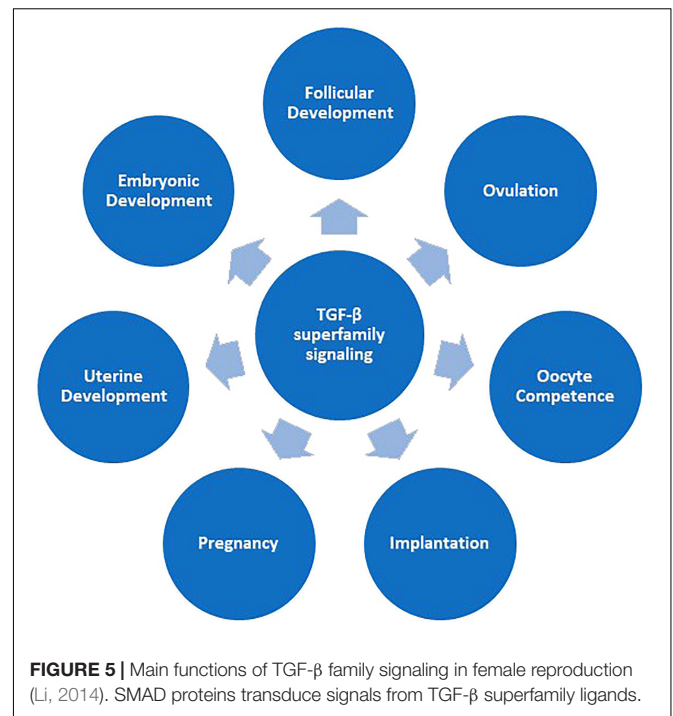


FIGURE 4 | Most related Gene Ontology (GO) terms were significantly enriched ($p \leq 0.01$) using genes associated with ewe composite traits at weaning. PWW: Progeny Weaning Weight; TLWW, Total Litter Weight at Weaning; LMWW, Litter Mean Weight. Gene Ratio: Number of the significant genes in the category/ Number of all genes in the category. Complete associated GO terms with these traits are provided in **Supplementary Tables 4–6**.

these traits in sheep. In the first study, Abdoli et al. (2019) reported five genes neighboring the top SNP (on OAR2, OAR3, OAR15, and OAR16), including *TEX12*, *BCO2*, *WDR70*, *INHBE*, and *INHBC* as possible candidate genes affecting composite traits of the Lori-Bakhtiari sheep. In this study, to attain more consistent findings, two different GWAS approaches were used. Both approaches identified similar regions that may explain some part of the genetic variation in the studied traits. We identified four genes on OAR15 (19.7–20.3 Mb), namely, *RDX*, *FDX1*, *ZC3H12C*, and *ARHGAP20* as maternal genes affecting composite traits at birth. *RDX* (Radixin) is part of the ERM (*EZR-RDX-MSN*) cytoskeleton linker protein family. The expression of ERM proteins in the blastocyst and the uterus has been reported and linked to implantation potential in mice (Matsumoto et al., 2004). *Ferredoxin* (*FDX1*) is an electron transport intermediate that is functional in mitochondrial cytochromes P450 and is found mainly in steroidogenic tissues like testis, adrenal glands, ovaries, and placenta (Sheftel et al., 2010). In this study, the *ARHGAP20* gene was identified as the candidate gene in both GWAS approaches. High expression levels of the *ARHGAP20* gene in the brain have been reported, indicating a role by this gene in neurogenesis (Kalla et al., 2005). The *Zc3h12c* is an endogenous inhibitor of TNF α -induced inflammatory signaling in the human umbilical vein and endothelial cells. It seems that the *Zc3h12c* gene plays a role in immune regulation in pregnancy (Liu et al., 2013). Abdoli et al. (2019) identified an SNP on OAR15 located on 22.02 Mb significantly associated with the TLWB trait, which is close to the region identified in this study and reinforces this possibility that this region on OAR15 is likely to affect fetal development during pregnancy.

Our GWAS analysis identified a region on OAR7 at 30.2–35.1 Mb that contains three suggestively associated SNPs with PBW. This region harbors 12 genes, such as *THBS1* and *TTBK2*. It has been reported that the expression of *THBS1* by placental cells is crucial for the formation of the placental structure (Ostankova et al., 2011). One of these SNPs, rs400684038, is located within the *TTBK2* gene. The *TTBK2* gene encodes a serine-threonine kinase that phosphorylates tau and tubulin proteins and is a critical regulator of processes that initiate the assembly of primary cilia in the embryo (Goetz et al., 2012). Both GWAS approaches identified the SNP rs430043751 on OAR23 associated with TLWW. This SNP harbors six genes, including *EPG5*, *PSTPIP2*, *ATP5F1A*, *HAUS1*, *RNF165a*, and *LOXHD1*. The *EPG5* gene encodes a protein with a crucial role in the autophagy pathway which contributes to early differentiation in human embryonic stem cells (Tra et al., 2011).

Our gene set analysis identified several significantly associated GO terms with maternal composite traits at birth (Figure 3). Interestingly, many GO terms were related to the neural system and showed an overrepresentation of significant genes. Numerous studies have reported that neural alterations occur extensively in pregnant women's brains (Cohen and Mizrahi, 2015; Bridges, 2016). Notably, our identified gene, *ARHGAP20*, has a role in neurogenesis. We identified two pathways (GO:0000045 and GO:0019855) related to calcium ion metabolism. A report suggests that placental calcium transfer increases during pregnancy to match fetal needs and ensure



appropriate fetal skeletal mineralization (Strid and Powell, 2000). However, recent evidence has grown inconsistent about the effects of maternal calcium on birth weight (Thompson et al., 2019). In this regard, Imdad and Bhutta (2012) concluded that calcium supplementation during pregnancy is associated with a reduction in risk of gestational hypertensive disorders and pre-term birth and an increase in birth weight. The *positive regulation of T-helper 1 type immune response* term was significant for all three composite traits at birth. During pregnancy, the fetal expression of paternal major histocompatibility (MHC) antigens renders it foreign, and thus, the maternal immune system must tolerate the semi-allogeneic fetus to support the pregnancy, without causing the mother to become susceptible to infection (Munoz-Suano et al., 2011). A shift in the balance of T_H1/T_H2 cytokine production by maternal peripheral blood leukocytes is regarded as a commonly important feature of successful mammalian pregnancy (Wattegedera et al., 2008). Recently, it has been reported that pregnancy can change the production of Th1 and Th2 cytokines in the maternal thymus in sheep (Zhang et al., 2019). GO terms related to the signaling pathways also showed an overrepresentation of significant genes. *SMAD protein signal transduction* (GO:0060395) was one of these pathways. SMAD proteins transduce signals from the TGF- β superfamily ligands and, as a result, regulate target gene expression. TGF- β superfamily signaling is vital for female reproduction (Figure 5).

It has been reported that SMAD proteins have roles in maintaining the structural and functional integrity of the oviduct and uterus. They are essential for the establishment and maintenance of pregnancy (Rodriguez et al., 2016). Another signaling pathway is Notch signaling which exerts

effects throughout the pregnancy and plays an important role in placental angiogenesis, normal placental function, and trophoblast function (Zhao and Lin, 2012). The *NIK/NF-kappaB signaling* (GO:1901222) works as a transcription factor involved in inflammatory and immune responses (Baldwin, 1996). The effects of NF- κ B and its signaling pathway on the human myometrium during pregnancy and parturition are well-reviewed (Cookson and Chapman, 2010).

GWAS and GSEA of Composite Traits at Weaning

Genes, including *NDUFA12*, *NR2C1*, *FGD6*, *VEZT*, *MIR331*, and *METAP2* were identified on OAR3, specifically around the SNP rs428404187. This SNP is significantly associated with all composite traits at weaning and located within the *VEZT* gene which has a major role in the cellular adhesion process. The cell adhesion process has a widespread effect on the development of mammary glands. It mainly occurs in late pregnancy and partially in the onset of lactation (Shamir and Ewald, 2015). Notably, we identified the *cell-cell adhesion* (GO:0098609) pathway as a significant GO term associated with the PWW trait in our gene set analysis. Another gene in this region, *NR2C1*, is a nuclear steroid hormone receptor. This gene acts as a transcription factor and plays an important role in the differentiation of mammary glands and its development in late pregnancy and during lactation (To and Andrechek, 2018).

The SNP rs398620273 on OAR5 was suggestively associated with all three composite traits at weaning. It is located within the *DTWD2* gene and is likely to be involved in RNA processing (Burroughs and Aravind, 2014). The *HSD17B4* is another gene that was identified around this SNP and plays an important role in feed intake and food conservation (Salleh et al., 2017). In dairy cows, feed intake is a major factor that controls milk production in high-yielding dairy cows in early lactation (Johansen et al., 2018). We identified the *RPLP0* gene on OAR2. Dominant expressions of *RPLP0* occur in mammary vasculature tissues during lactation (Lee et al., 2010). The *SNORA62* gene was identified on OAR3 as a suggestive gene affecting the LMWW trait. Recently, a GWAS of milk fatty acid composition led to the identification of the *SNORA62* gene as a candidate gene affecting fatty acid content in milk (Palombo et al., 2018).

We identified 24 suggestive genes on OAR7 related to PWW and seven of which belong to the pancreatic ribonuclease A family (*RNases*). It seems that this region plays an important role in RNA processing. The *RNASE5* is known as a functional gene in milk production, specifically in milk protein percentage (Raven et al., 2013). We identified a suggestive SNP, rs401393221, within the *RSU1* gene on OAR13. Using meta-analysis and supervised machine learning models on microarray data, the *RSU1* gene has been identified as DEG during the lactation process in both approaches (Farhadian et al., 2018). It is worth noting that the *RSU1* gene is a member of the *milk proteins* KEGG pathway. In addition, we identified *CUBN* and *VIM* genes around this SNP on OAR13. Association between *CUBN* gene and variation in

vitamin B-12 content in bovine milk have been reported (Rutten et al., 2013). *VIM* is a cytoskeletal type III intermediate and has a critical role in the development of mammary glands (Peuhu et al., 2017). A fourfold increase of *VIM* protein in lactating tissues compared to resting tissues reported (Michalczyk et al., 2001).

The SNP rs409558668 was identified on OAR16 with a suggestive association with the LMWW trait. This SNP is located within the *TTC23L* gene and harbors the *PRLR* (Prolactin receptor) gene. The *TTC23L* gene is highly expressed during lactation (Paten, 2014) and is identified as a candidate gene that can affect mastitis in Holstein cows (Tiezzi et al., 2015). The *PRLR* gene is usually expressed in lactating mammary glands and it has been shown that the polymorphism in exon 3 and 7 of the *PRLR* gene is correlated with milk production in Holstein cows (Zhang et al., 2008). Also, we identified the *FTH1* gene on OAR16. The *FTH1* gene encodes the heavy subunit of ferritin. The presence of ferritin in cow's and buffalo's milk has been reported (Farhadian et al., 2018; Arora et al., 2019).

Through gene set analysis for composite traits at weaning, several maternal functional categories were identified. Many GO terms were discovered in association with protein metabolism, protein transport, and fatty acid metabolism. Recently, in a transcriptome analysis study on buffalo milk, the *protein metabolism* (GO:0019538) pathway was identified as a significant term (Arora et al., 2019). Considerable changes occur in the amount of fatty acid synthesis during late pregnancy and lactation. These changes have been reported in a variety of species, like rats, rabbits, pigs, and cows (Larson and Smith, 1974; Abdollahi-Arpanahi et al., 2019). There have been significant observations regarding GO terms linked to calcium and ions metabolism and transport. Many different comparative transcriptome analyses have reported the role of calcium metabolism-related pathways in the lactation process (Arora et al., 2019; Bhat et al., 2019). In the current study, *Phosphorylation* (GO:0016310) was identified as a significant pathway associated with PWW and LMWW traits. On average, caseins comprise 80% of proteins in cow and sheep milk, so, phosphorylation by the Casein Kinase enzyme is a crucial step for milk production in the lactating mammary gland (Bionaz et al., 2012). Notably, the *kinase activity* (GO:0016301) pathway is another significant term for LMWW that catalyzes the transfer of a phosphate group to a substrate molecule. The *cell-cell adhesion* pathway was significantly associated with PWW. The effects of cell adhesion molecules on the lactogenesis process have been reviewed thoroughly in the scientific literature (Morrison and Cutler, 2010). The *VEZT* gene, one of our identified genes in the GWA analysis, is a member of this pathway. The *GTPase activator activity* GO term was identified as a significant pathway related to all three composite traits at weaning. GTPases are known to be involved in numerous secretory processes and play an important role in the translation and translocation of proteins, the secretion of milk fat globules, and, probably, other milk components (Arora et al., 2019). Many GO terms associated with cell proliferation and differentiation were also detected as significant. Most mammary growth takes place through pregnancy. Mammary gland cell proliferation and differentiation have a great impact on milk yield and lactation

persistence (Davis, 2017). Mammary epithelial cells (MECs) are key cells that are present in lactating mammary glands and are responsible for milk production. The number of MECs in the mammary gland and their secretory activity are crucial factors that regulate milk yield (Herve et al., 2016). Milk is essential for lamb survival and growth in the first 3–4 weeks of life and 70% of the difference in weight gain from 3 to 12 weeks is attributed to milk intake (Morgan et al., 2007). In twin lambs, prenatal ewe traits (e.g., ewe weight at mating and set stocking, as well as ewe body condition score at mating and stocking) usually have minimal effects on milk yield and lamb growth until the time of weaning, but milk yield and composition have the greatest proportion of variation in lamb weight gain (Danso et al., 2016) which is consistent with the results of this study.

To date, this is the second GWAS report on composite reproductive traits in sheep. At this point, these genes are merely candidates. The identification of validated causal genetic variants that underlie production traits is one of the main challenges in current livestock genetic research. It is important to point out that the limited number of animals and low to moderate heritability of the traits, actually hinder the detection of strong association signals. Composite traits are complex and try to capture growth, yield, and several different aspects of fertility into a combined selection of traits at the same time. As such, the traits are necessarily polygenic by far in nature and it will require thousands of genotypes to disentangle true signals from background noise. In this study, a good amount of caution was considered for performing analysis and reporting the results. We used two GWAS approaches to confirm the results and reported the *p*-values without any inflation to avoid false positives as much as possible. Of course, we couldn't apply a stringent threshold for all reported SNPs and, hopefully, there are no false positives in the results. However, the Baluchi sheep is a widely used breed in east Iran, but it is not widely distributed in other parts of the world. Finally, while this study does improve our understanding of an interesting but less characterized breed, it will still be useful to see if these results can be broadly applicable to other breeds as well.

CONCLUSION

We used GWAS and GSEA together to find genes and pathways affecting maternal composite traits at birth and weaning in sheep. Several genes including *RDX*, *FDX1*, *ARHGAP20*, *ZC3H12C*, *THBS1*, and *EPG5* were associated with composite traits at birth. These genes play roles in pregnancy, particularly in autophagy, immune response, angiogenesis, and placental formation. Gene set analysis identified *calcium ion transport* as a significant GO term that affecting composite traits at birth. In addition, we identified many genes (e.g., *NR2C1*, *VEZT*, *HSD17B4*, *RSU1*, *CUBN*, *VIM*, *PRLR*, and *FTH1*) as genes affecting composite traits at weaning. Our gene set analysis on these traits identified several significantly related GO terms, e.g., protein processing and transport, phospholipid translocation, ion transport, and cell-cell adhesion. As expected,

most identified genes and GO terms have a role in milk production or in mammary gland development, which means that feeding lambs by milk can have the greatest impact on weight gain as compared to other effects of maternal origin. This suggests that farmers should select ewes with higher milk yields to maximize lamb growth until weaning. Moreover, the results provide a good insight into how maternal genes and pathways influence progeny weight at birth and, subsequently, at weaning.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://figshare.com/> and <https://doi.org/10.6084/m9.figshare.11859996.v1>.

ETHICS STATEMENT

The animal study was reviewed and approved by the Sari Agricultural Sciences and Natural Resources University.

AUTHOR CONTRIBUTIONS

SME-F: conceptualization, data curation, methodology, formal analysis, software, visualization, writing—original draft, and writing—review and editing. MG: methodology, supervision, resources, and writing—review and editing. SH: project administration, supervision, and writing—review and editing. RA-A: methodology, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

SME-F thanks the Abureyhan Campus of the University of Tehran for hosting during the research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.710613/full#supplementary-material>

Supplementary Tables 1–6 | List of significantly ($p \leq 0.01$) associated GO terms with the studied traits.

Supplementary Table 7 | Correlations between the original traits and EBVs of the traits.

Supplementary Tables 8–13 | List of significant genes used for GSEAs.

Supplementary Table 14 | The regression coefficient (SNP effects) of reported SNPs for composite traits at birth.

Supplementary Table 15 | The regression coefficient (SNP effects) of reported SNPs for composite traits at weaning.

REFERENCES

- Abbasi, M. A., Abdollahi-Arpanahi, R., Maghsoudi, A., Torshizi, R. V., and Nejati-Javaremi, A. (2012). Evaluation of models for estimation of genetic parameters and maternal effects for early growth traits of Iranian Baluchi sheep. *Small Rumin. Res.* 104, 62–69. doi: 10.1016/J.SMALLRUMRES.2011.10.003
- Abdoli, R., Mirhoseini, S., Ghavi Hossein-Zadeh, N., Zamani, P., Ferdosi, M., and Gondro, C. (2019). Genome-wide association study of four composite reproductive traits in Iranian fat-tailed sheep. *Reprod. Fertil. Dev.* 31, 1127–1133. doi: 10.1071/RD18282
- Abdoli, R., Mirhoseini, S. Z., Ghavi Hossein-Zadeh, N., Zamani, P., and Gondro, C. (2018). Genome-wide association study to identify genomic regions affecting prolificacy in Lori-Bakhtiari sheep. *Anim. Genet.* 49, 488–491. doi: 10.1111/age.12700
- Abdollahi-Arpanahi, R., Carvalho, M. R., Ribeiro, E. S., and Peñagaricano, F. (2019). Association of lipid-related genes implicated in conceptus elongation with female fertility traits in dairy cattle. *J. Dairy Sci.* 102, 10020–10029. doi: 10.3168/jds.2019-17068
- Arora, R., Sharma, A., Sharma, U., Girdhar, Y., Kaur, M., Kapoor, P., et al. (2019). Buffalo milk transcriptome: a comparative analysis of early, mid and late lactation. *Sci. Rep.* 9:5993.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25.
- Aulchenko, Y. S., Ripke, S., Isaacs, A., and Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–1296. doi: 10.1093/bioinformatics/btm108
- Baldwin, A. S. Jr. (1996). The NF- κ B and I κ B proteins: new discoveries and insights. *Annu. Rev. Immunol.* 14, 649–681. doi: 10.1146/annurev.immunol.14.1.649
- Bhat, S. A., Ahmad, S. M., Ibeagha-Awemu, E. M., Bhat, B. A., Dar, M. A., Mumtaz, P. T., et al. (2019). Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between Jersey and Kashmiri cattle. *PLoS One* 14:e0211773. doi: 10.1371/journal.pone.0211773
- Bionaz, M., Hurley, W., and Loo, J. (2012). “Milk protein synthesis in the lactating mammary gland: insights from transcriptomics analyses,” in *Milk protein*, ed. W. Hurley (Rijeka: InTech), 285–324.
- Bridges, R. S. (2016). Long-term alterations in neural and endocrine processes induced by motherhood in mammals. *Horm. Behav.* 77, 193–203. doi: 10.1016/j.jhbeh.2015.09.001
- Bromley, C. M., Van Vleck, L. D., and Snodder, G. D. (2001). Genetic correlations for litter weight weaned with growth, prolificacy, and wool traits in Columbia, Polypay, Rambouillet, and Targhee sheep. *J. Anim. Sci.* 79, 339–346. doi: 10.2527/2001.792339x
- Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015
- Burroughs, A. M., and Aravind, L. (2014). Analysis of two domains with novel RNA-processing activities throws light on the complex evolution of ribosomal RNA biogenesis. *Front. Genet.* 5:424. doi: 10.3389/fgene.2014.00424
- Cohen, L., and Mizrahi, A. (2015). Plasticity during motherhood: changes in excitatory and inhibitory layer 2/3 neurons in auditory cortex. *J. Neurosci.* 35, 1806–1815. doi: 10.1523/jneurosci.1786-14.2015
- Cookson, V. J., and Chapman, R. (2010). NF- κ B function in the human myometrium during pregnancy and parturition. *Histol. Histopathol.* 25, 945–956.
- Dadousis, C., Pegolo, S., Rosa, G. J. M., Gianola, D., Bittante, G., and Cecchinato, A. (2017). Pathway-based genome-wide association analysis of milk coagulation properties, curd firmness, cheese yield, and curd nutrient recovery in dairy cattle. *J. Dairy Sci.* 100, 1223–1231. doi: 10.3168/jds.2016-11587
- Danso, A. S., Morel, P. C. H., Kenyon, P. R., and Blair, H. T. (2016). Relationships between prenatal ewe traits, milk production, and preweaning performance of twin lambs. *J. Anim. Sci.* 94, 3527–3539. doi: 10.2527/jas.2016-0337
- Davis, S. R. (2017). Triennial lactation symposium/bolfa: mammary growth during pregnancy and lactation and its relationship with milk yield. *J. Anim. Sci.* 95, 5675–5688. doi: 10.2527/jas.2017.1733
- Duguma, G., Schoeman, S. J., Cloete, S. W. P., and Jordaan, G. F. (2002). Genetic and environmental parameters for ewe productivity in Merinos. *S. Afr. J. Anim. Sci.* 32, 154–159.
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525
- Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4:1184. doi: 10.1038/nprot.2009.97
- Esmaeili-Fard, S. M., Gholizadeh, M., Hafezian, S. H., and Abdollahi-Arpanahi, R. (2021). Genome-wide association study and pathway analysis identify NTRK2 as a novel candidate gene for litter size in sheep. *PLoS One* 16:e0244408. doi: 10.1371/journal.pone.0244408
- Farhadian, M., Rafat, S. A., Hasanpur, K., Ebrahimi, M., and Ebrahimie, E. (2018). Cross-species meta-analysis of transcriptomic data in combination with supervised machine learning models identifies the common gene signature of lactation process. *Front. Genet.* 9:235. doi: 10.3389/fgene.2018.00235
- Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* 32, 361–369. doi: 10.1002/gepi.20310
- Gholizadeh, M., and Ghafouri-Kesbi, F. (2015). Estimation of genetic parameters for growth-related traits and evaluating the results of a 27-year selection program in Baluchi sheep. *Small Rumin. Res.* 130, 8–14. doi: 10.1016/j.smallrumres.2015.07.032
- Gholizadeh, M., Rahimi-Mianji, G., Nejati-Javaremi, A., De Koning, D. J., and Jonas, E. (2014). Genomewide association study to detect QTL for twinning rate in Baluchi sheep. *J. Genet.* 93, 489–493. doi: 10.1007/s12041-014-0372-1
- Goetz, S. C., Liem, K. F. Jr., and Anderson, K. V. (2012). The spinocerebellar ataxia-associated gene Tau tubulin kinase 2 controls the initiation of ciliogenesis. *Cell* 151, 847–858. doi: 10.1016/j.cell.2012.10.010
- Han, Y., and Peñagaricano, F. (2016). Unravelling the genomic architecture of bull fertility in Holstein cattle. *BMC Genet.* 17:143. doi: 10.1186/s12863-016-0454-6
- Hanford, K. J., Van Vleck, L. D., and Snodder, G. D. (2003). Estimates of genetic parameters and genetic change for reproduction, weight, and wool characteristics of Targhee sheep. *J. Anim. Sci.* 81, 630–640. doi: 10.2527/2003.813630x
- Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430
- Herve, L., Quesnel, H., Lollivier, V., and Boutinaud, M. (2016). Regulation of cell number in the mammary gland by controlling the exfoliation process in milk in ruminants. *J. Dairy Sci.* 99, 854–863. doi: 10.3168/jds.2015-9964
- Imdad, A., and Bhutta, Z. A. (2012). Effects of calcium supplementation during pregnancy on maternal, fetal and birth outcomes. *Paediatr. Perinat. Epidemiol.* 26, 138–152. doi: 10.1111/j.1365-3016.2012.01274.x
- Johansen, M., Lund, P., and Weisbjerg, M. R. (2018). Feed intake and milk production in dairy cows fed different grass and legume species: a meta-analysis. *Animal* 12, 66–75. doi: 10.1017/s1751731117001215
- Kalla, C., Nentwich, H., Schlotter, M., Mertens, D., Wildenberger, K., Döhner, H., et al. (2005). Translocation t (X; 11)(q13; q23) in B-cell chronic lymphocytic leukemia disrupts two novel genes. *Genes Chromosomes Cancer* 42, 128–143.
- Larson, B. L., and Smith, V. R. (1974). *Lactation. A Comprehensive Treatise. Biosynthesis and Secretion of Milk/Diseases*, Vol. II. Cambridge, MA: Academic Press.
- Lee, N. K., Kim, M. K., Choi, J. H., Kim, E. B., Lee, H. G., Kang, S. K., et al. (2010). Identification of a peptide sequence targeting mammary vasculature via RPLP0 during lactation. *Peptides* 31, 2247–2254. doi: 10.1016/j.peptides.2010.09.008
- Li, Q. (2014). Transforming growth factor β signaling in uterine development and function. *J. Anim. Sci. Biotechnol.* 5:52. doi: 10.1186/2049-1891-5-52
- Liu, L., Zhou, Z., Huang, S., Guo, Y., Fan, Y., Zhang, J., et al. (2013). Zc3h12c inhibits vascular inflammation by repressing NF- κ B activation and pro-inflammatory gene expression in endothelial cells. *Biochem. J.* 451, 55–60. doi: 10.1042/bj20130019

- Masuda, Y. (2019). *Introduction to BLUPF90 suite programs: Standard Edition*. Available online at: <http://nce.ads.uga.edu/wiki/doku.php?id=documentation>
- Matsumoto, H., Daikoku, T., Wang, H., Sato, E., and Dey, S. K. (2004). Differential expression of Ezrin/Radixin/Moesin (ERM) and ERM-associated adhesion molecules in the blastocyst and uterus suggests their functions during Implantation1. *Biol. Reprod.* 70, 729–736. doi: 10.1095/biolreprod.103.022764
- Michalczyk, A., Ackland, M. L., Brown, R. W., and Collins, J. P. (2001). Lactation affects expression of intermediate filaments in human breast epithelium. *Differentiation* 67, 41–49. doi: 10.1046/j.1432-0436.2001.067001041.x
- Mohammadi, K., Nassiri, M. T. B., Rahmatnejad, E., Sheikh, M., Fayazi, J., and Manesh, A. K. (2013). Phenotypic and genetic parameter estimates for reproductive traits in Zandi sheep. *Trop. Anim. Health Prod.* 45, 671–677. doi: 10.1007/s11250-012-0276-0
- Mokhtari, M. S., Rashidi, A., and Esmailzadeh, A. K. (2010). Estimates of phenotypic and genetic parameters for reproductive traits in Kermani sheep. *Small Rumin. Res.* 88, 27–31. doi: 10.1016/j.smallrumres.2009.11.004
- Moradband, F., Rahimi, G., and Gholizadeh, M. (2011). Association of polymorphisms in fecundity genes of GDF9, BMP15 and BMP15-1B with litter size in Iranian Baluchi sheep. *Asian Austr. J. Anim. Sci.* 24, 1179–1183. doi: 10.5713/ajas.2011.10453
- Morgan, J. E., Fogarty, N. M., Nielsen, S., and Gilmour, A. R. (2007). The relationship of lamb growth from birth to weaning and the milk production of their primiparous crossbred dams. *Aust. J. Exp. Agric.* 47, 899–904. doi: 10.1071/ea06290
- Morrison, B., and Cutler, M. L. (2010). The contribution of adhesion signaling to lactogenesis. *J. Cell. Commun. Signal.* 4, 131–139. doi: 10.1007/s12079-010-0099-6
- Munoz-Suano, A., Hamilton, A. B., and Betz, A. G. (2011). Gimme shelter: the immune system during pregnancy. *Immunol. Rev.* 241, 20–38. doi: 10.1111/j.1600-065X.2011.01002.x
- Neupane, M., Kiser, J. N., Bovine Respiratory Disease Complex Coordinated Agricultural Project Research Team, and Neibergs, H. L. (2018). Gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease. *Anim. Genet.* 49, 527–538. doi: 10.1111/age.12718
- Ostankova, Y. V., Klimovskaya, Y. S., Gorskaya, O. A., Kolobov, A. V., Kvetnoi, I. M., Selkov, S. A., et al. (2011). Expression of thrombospondin-1 gene mRNA and protein in the placenta in gestosis. *Bull. Exp. Biol. Med.* 151, 215–218. doi: 10.1007/s10517-011-1292-1
- Palombo, V., Milanese, M., Sgorlon, S., Capomaccio, S., Mele, M., Nicolazzi, E., et al. (2018). Genome-wide association study of milk fatty acid composition in Italian Simmental and Italian Holstein cows using single nucleotide polymorphism arrays. *J. Dairy Sci.* 101, 11004–11019. doi: 10.3168/jds.2018-14413
- Paten, A. M. (2014). *Maternal Nutritional Programming in the Sheep: Effects on Post-Natal Growth, Mammogenesis and Lactation in Adult-Ewe Offspring: A Thesis Submitted for the Degree*. Ph.D. thesis in Animal Science. Palmerston North: Massey University.
- Peuhu, E., Virtakoivu, R., Mai, A., Wärrä, A., and Ivaska, J. (2017). Epithelial vimentin plays a functional role in mammary gland development. *Development* 144, 4103–4113.
- R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: <https://www.R-project.org/>
- Rashidi, A., Mokhtari, M. S., Esmailzadeh, A. K., and Fozi, M. A. (2011). Genetic analysis of ewe productivity traits in Moghani sheep. *Small Rumin. Res.* 96, 11–15. doi: 10.1016/j.smallrumres.2010.11.001
- Raven, L.-A., Cocks, B. G., Pryce, J. E., Cottrell, J. J., and Hayes, B. J. (2013). Genes of the RNASE5 pathway contain SNP associated with milk production traits in dairy cattle. *Genet. Sel. Evol.* 45:25.
- Rodriguez, A., Tripurani, S. K., Burton, J. C., Clementi, C., Larina, I., and Pangas, S. A. (2016). SMAD signaling is required for structural integrity of the female reproductive tract and uterine function during early pregnancy in mice. *Biol. Reprod.* 95, 41–44.
- Rönnegård, L., McFarlane, S. E., Husby, A., Kawakami, T., Ellegren, H., and Qvarnström, A. (2016). Increasing the power of genome wide association studies in natural populations using repeated measures—evaluation and implementation. *Methods Ecol. Evol.* 7, 792–799. doi: 10.1111/2041-210x.12535
- Rosati, A., Mousa, E., Van Vleck, L. D., and Young, L. D. (2002). Genetic parameters of reproductive traits in sheep. *Small Rumin. Res.* 43, 65–74. doi: 10.1016/s0921-4488(01)00256-5
- Rutten, M. J. M., Bouwman, A. C., Sprong, R. C., van Arendonk, J. A. M., and Visker, M. H. P. W. (2013). Genetic variation in vitamin B-12 content of bovine milk and its association with SNP along the bovine genome. *PLoS One* 8:e62382. doi: 10.1371/journal.pone.0062382
- Salleh, M. S., Mazzoni, G., Höglund, J. K., Olijhoek, D. W., Lund, P., Løvendahl, P., et al. (2017). RNA-Seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high- and low-residual feed intake in Nordic dairy cattle. *BMC Genomics* 18:258. doi: 10.1186/s12864-017-3622-9
- Shamir, E. R., and Ewald, A. J. (2015). “Adhesion in mammary development: novel roles for E-cadherin in individual and collective cell migration,” in *Current Topics in Developmental Biology*, ed. A. S. Yap (Amsterdam: Elsevier), 353–382.
- Sheftel, A. D., Stehling, O., Pierik, A. J., Elsässer, H.-P., Mühlenhoff, U., Weber, H., et al. (2010). Humans possess two mitochondrial ferredoxins, Fdx1 and Fdx2, with distinct roles in steroidogenesis, heme, and Fe/S cluster biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11775–11780. doi: 10.1073/pnas.1004250107
- Snowder, G. D., and Fogarty, N. M. (2009). Composite trait selection to improve reproduction and ewe productivity: a review. *Anim. Prod. Sci.* 49, 9–16. doi: 10.1071/ea08184
- Strid, H., and Powell, T. L. (2000). ATP-dependent Ca²⁺ transport is up-regulated during third trimester in human syncytiotrophoblast basal membranes. *Pediatr. Res.* 48:58. doi: 10.1203/00006450-200007000-00012
- Tahmoorepur, M., and Sheikhlou, M. (2011). Pedigree analysis of the closed nucleus of Iranian Baluchi sheep. *Small Rumin. Res.* 99, 1–6. doi: 10.1016/j.smallrumres.2011.01.017
- Thompson, W. D., Tyrrell, J., Borges, M.-C., Beaumont, R. N., Knight, B. A., Wood, A. R., et al. (2019). Association of maternal circulating 25 (OH) D and calcium with birth weight: a mendelian randomisation analysis. *PLoS Med.* 16:e1002828. doi: 10.1371/journal.pmed.1002828
- Tiezzi, F., Parker-Gaddis, K. L., Cole, J. B., Clay, J. S., and Maltecca, C. (2015). A genome-wide association study for clinical mastitis in first parity US Holstein cows using single-step approach and genomic matrix re-weighting procedure. *PLoS One* 10:e0114919. doi: 10.1371/journal.pone.0114919
- To, B., and Andrechek, E. R. (2018). Transcription factor compensation during mammary gland development in E2F knockout mice. *PLoS One* 13:e0194937. doi: 10.1371/journal.pone.0194937
- Tra, T., Gong, L., Kao, L.-P., Li, X.-L., Grandela, C., Devenish, R. J., et al. (2011). Autophagy in human embryonic stem cells. *PLoS One* 6:e27485. doi: 10.1371/journal.pone.0027485
- Vatankhah, M., Talebi, M. A., and Edriss, M. A. (2008). Estimation of genetic parameters for reproductive traits in Lori-Bakhtiari sheep. *Small Rumin. Res.* 74, 216–220. doi: 10.1016/j.smallrumres.2007.02.008
- Wang, C. T., and Dickerson, G. E. (1991). Simulation of life-cycle efficiency of lamb and wool production for genetic levels of component traits and alternative management options. *J. Anim. Sci.* 69, 4324–4337. doi: 10.2527/1991.69114324x
- Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* 98, 1–8. doi: 10.1016/j.ygeno.2011.04.006
- Wattegedera, S., Rocchi, M., Sales, J., Howard, C. J., Hope, J. C., and Entrican, G. (2008). Antigen-specific peripheral immune responses are unaltered during normal pregnancy in sheep. *J. Reprod. Immunol.* 77, 171–178. doi: 10.1016/j.jri.2007.07.003

- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer.
- Yavarifard, R., Ghavi Hossein-Zadeh, N., and Shadparvar, A. A. (2015). Estimation of genetic parameters for reproductive traits in Mehraban sheep. *Czech J. Anim. Sci.* 60, 281–288. doi: 10.17221/8242-CJAS
- Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14.
- Zhang, J., Zan, L., Fang, P., Zhang, F., Shen, G., and Tian, W. (2008). Genetic variation of PRLR gene and association with milk performance traits in dairy cattle. *Can. J. Anim. Sci.* 88, 33–39. doi: 10.4141/cjas07052
- Zhang, L., Zhao, Z., Mi, H., Liu, B., Wang, B., and Yang, L. (2019). Modulation of helper T cytokines in thymus during early pregnancy in ewes. *Animals* 9:245. doi: 10.3390/ani9050245
- Zhao, W.-X., and Lin, J.-H. (2012). Notch signaling pathway and human placenta. *Int. J. Med. Sci.* 9:447. doi: 10.7150/ijms.4593

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Esmaeili-Fard, Gholizadeh, Hafezian and Abdollahi-Arpanahi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



SNPs in Mammary Gland Epithelial Cells Unraveling Potential Difference in Milk Production Between Jersey and Kashmiri Cattle Using RNA Sequencing

Syed Mudasir Ahmad^{1*}, Basharat Bhat^{1*}, Shakil Ahmad Bhat¹, Mifftha Yaseen², Shabir Mir³, Mustafa Raza⁴, Mir Asif Iquebal⁴, Riaz Ahmad Shah¹ and Nazir Ahmad Ganai⁵

OPEN ACCESS

Edited by:

Fabyano Fonseca Silva,
Universidade Federal de Viçosa, Brazil

Reviewed by:

Marcos Inacio Marcondes,
Washington State University,
United States

Ana Paula Sbardella,
Crisis Pecuaría Inteligente, Brazil

*Correspondence:

Syed Mudasir Ahmad
mudasirbio@gmail.com
Basharat Bhat
bb284@snu.edu.in

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 09 February 2021

Accepted: 06 July 2021

Published: 03 August 2021

Citation:

Ahmad SM, Bhat B, Bhat SA,
Yaseen M, Mir S, Raza M,
Iquebal MA, Shah RA and Ganai NA
(2021) SNPs in Mammary Gland
Epithelial Cells Unraveling Potential
Difference in Milk Production Between
Jersey and Kashmiri Cattle Using
RNA Sequencing.
Front. Genet. 12:666015.
doi: 10.3389/fgene.2021.666015

¹ Division of Animal Biotechnology, Faculty of Veterinary Sciences and Animal Husbandry, Sher-e-Kashmir University of Agricultural Sciences and Technology, Srinagar, India, ² Division of Food Science, Faculty of Horticulture, Sher-e-Kashmir University of Agricultural Sciences and Technology, Srinagar, India, ³ Division of Animal Genetics and Breeding, Faculty of Veterinary Sciences and Animal Husbandry, Sher-e-Kashmir University of Agricultural Sciences and Technology, Srinagar, India, ⁴ Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, ⁵ Directorate Planning and Monitoring, Sher-e-Kashmir University of Agricultural Sciences and Technology, Srinagar, India

Deep RNA sequencing experiment was employed to detect putative single nucleotide polymorphisms (SNP) in mammary epithelial cells between two diverse cattle breeds (Jersey and Kashmiri) to understand the variations in the coding regions that reflect differences in milk production traits. The low milk-producing Kashmiri cattle are being replaced by crossbreeding practices with Jersey cattle with the aim of improving milk production. However, crossbred animals are prone to infections and various other diseases resulting in unsustainable milk production. In this study, we tend to identify high-impact SNPs from Jersey and Kashmiri cows (utilizing RNA-Seq data) to delineate key pathways mediating milk production traits in both breeds. A total of 607 (442 SNPs and 169 INDELs) and 684 (464 SNPs and 220 INDELs) high-impact variants were found specific to Jersey and Kashmir cattle, respectively. Based on our results, we conclude that in Jersey cattle, genes with high-impact SNPs were enriched in nucleotide excision repair pathway, ABC transporter, and metabolic pathways like glycerolipid metabolism, pyrimidine metabolism, and amino acid synthesis (glycine, serine, and threonine). Whereas, in Kashmiri cattle, the most enriched pathways include endocytosis pathway, innate immunity pathway, antigen processing pathway, insulin resistance pathway, and signaling pathways like TGF beta and AMPK which could be a possible defense mechanism against mammary gland infections. A varied set of SNPs in both breeds, suggests a clear differentiation at the genomic level; further analysis of high-impact SNPs are required to delineate their effect on these pathways.

Keywords: Jersey and Kashmiri cattle, SNP identification, mammary epithelial cells, RNA Seq data, milk production

INTRODUCTION

Cow milk is an essential natural product which provides a medium for nutrients including growth and immune factors to offspring and a valued raw material for human food (Séverin and Wenshui, 2005; Reinhardt and Lippolis, 2006). It plays an important role in supporting a healthy immune system and provides protection against infections (Goldman, 2000; Séverin and Wenshui, 2005). Milk is produced in the gland by mammary epithelial cells (MEC), which are gradually exfoliated from the epithelium during lactation (Boutinaud et al., 2015). Bovine milk comprises a diverse population of somatic cells including epithelial cells, macrophages, neutrophils, and lymphocytes. The process of lactation in animals consists of several physiological and metabolic changes (Arora et al., 2019). The length of lactation and yield greatly varied among the breeds (Auldist et al., 2007; Mech et al., 2008). Augmenting milk production in cattle is an essential step toward improving the profitability of dairy farms, and the success of dairy forms plays a crucial role in ensuring economic sustainability.

Single nucleotide polymorphisms markers are being rapidly used in selective breeding programmers for improving phenotypic selections of the animals through genomic selection (GS), gene-assisted selection (GAS), and marker-assisted selection (MAS) methods (Georges et al., 1995; Dekkers, 2004; Hayes and Goddard, 2010). It is remarkable that the discovery of the SNP for economic traits has great potential in the genetic improvement of cattle (Pareek et al., 2017). One of the great interests is the ability to improve lactation performance in poor performing breeds. To enhance the lactation performance of dairy animals, the knowledge of gene expression along with SNP profiling and biological pathways and mechanisms that promotes the mammary gland development and lactation is important.

High-throughput RNA sequencing technologies have provided unprecedented prospects in functional and comparative genomic research, including gene expression, genome annotation and pathway analysis, non-coding RNA discovery, and SNP detection and profiling (Bentley, 2006; Morozova and Marra, 2008). It has been widely used to study SNP in cattle breeds like Polish Holstein-Friesian (Pareek et al., 2016), Brangus, Brahman, Nellore, Angus, and Holstein (Dias et al., 2017) and indigenous cattle breeds like Simmental bull, Xuanhan bull, and Shuxuan bull (Wang et al., 2018).

Kashmiri and Jersey cattle are two important milk animals of the Indian northern state, Kashmir which contributes significantly to the total milk production in the state. Kashmiri indigenous cattle are small, hardy, and well adapted to the hilly areas of this region and differ greatly from Jersey cattle in dairy production traits. Whereas Jersey is a well-established exotic dairy breed which has been utilized in crossbreeding programs to enhance the milk production capability of Kashmiri cattle (Bhat et al., 2019).

In the present study, we used the transcriptome data of Jersey and Kashmiri cattle with the objective to identify the single nucleotide polymorphism in coding regions that were associated with milk production traits. Additionally, the study is aimed to benefit the marker-assisted selection programs by cataloging the

differential SNP profile for the improvement of milk production in Kashmiri indigenous cattle.

MATERIALS AND METHODS

Transcriptome data of the Jersey and Kashmiri cattle were downloaded from NCBI SRA database with accession number SRR6324372. The samples were obtained at three different lactation stages from three Jersey and three Kashmiri cattle's breed in similar conditions at Sher-e-Kashmir University of Agricultural Sciences and Technology dairy farm, Mountain Livestock Research Institute (MLRI), Kashmir, India. The sample collection from the healthy animals including sample preparation and RNA sequencing were explained in the study by Bhat et al. (2019). Read quality control was assessed using FASTQC program v0.11.1 (Andrews, 2010). After preprocessing, filtering of low-quality sequences and adaptor trimming were performed using Cutadapt v3.40 (Martin, 2011), high-quality sequencing reads that passed thresholds (PhredScore > 30) were assembled for SNP discovery analysis. A collection of over 40 million high-quality clean reads were obtained for each sample. The cleaned reads were mapped to reference genome assembly ARS-UCD1.2.99 using HISAT2 (Kim et al., 2019). The data preprocessing steps recommended in the GenomeAnalysisToolkit (GATK) best practices workflow was performed before variant identification (Poplin et al., 2018). PCR duplicates were marked with the MarkDuplicates from Picard tools (Picard toolkit, 2019). We also performed local realignment around InDels, checked intron-exon junctions, and recalibrated the base quality scores with GATK. Two different variant callers were used to perform SNP and INDELs discovery across eight Jersey and nine Kashmiri transcriptome samples separately: (i) GATK using the HaplotypeCaller tool in multi-sample calling mode (modality "GATK"); (ii) mpileup from SAMtools v1.4 (Li et al., 2009) in multisample calling mode using default parameters. Final set for analysis contains SNPs and InDelscommon in both datasets.

Bovine genetic variants from dbSNP 2.0 build 153 dated: Aug 8 2019 were incorporated in SNP calling to populate the RS_ID column of the known SNPs. Filtering [base quality score (Q-Score) > 30, mapping quality > 30, and minimum depth > 10] of generated variants and annotation were performed using VCFtools version 0.1.8 and SnpEff program v4.1. To further evaluate the biological significance of the genes with high-impact variations, the KEGG pathway enrichment analysis was performed using nKOBAS server version 3 (Xie et al., 2011; Kanehisa et al., 2021).

RESULTS

In our previous study after analyzing the inter- and intrastage gene expression profiling between Jersey and Kashmiri cattle, we observed a vast diversity in terms of gene expression while huge similarity between the breeds was also witnessed. Differentially expressed genes in both Kashmiri and Jersey were enriched

for multicellular organismal process, receptor activity, catalytic activity, signal transducer activity, macromolecular complex, and developmental processes. Most of the identified pathways responsible for milk production in Jersey include JAK-STAT, p38 MAPK, and PI3 kinase whereas antioxidant genes like RPLPO and RPS28 are highly expressed in Kashmiri cattle (Bhat et al., 2019). The present study based on SNP profiling in coding regions showed interbreed potential difference between Jersey and Kashmiri cattle. The difference in SNPs can help to understand its potential role in controlling the milk production between the two breeds.

Quality Control, Mapping, and Posttreatment

The trimming process removed 0.21% of reads and only 4.2% of duplicated reads were rejected. At the mapping step, around 99.7% of the reads were mapped against the reference genome (*Bos taurus* assembly ARS-UCD1.2.99) (Bhat et al., 2021). A total of 657,299 and 650,556 SNPs and INDELs were identified from Jersey and Kashmir cattle. Chromosomal distribution of SNPs and INDELs were provided in **Supplementary Table 1**, and variant types were provided in **Supplementary Tables 2, 3**. Transitions to transversions ratio (Ts/Tv) in both breeds were around 2.6, with 2,047,112 transitions and 761,864 transversions in Jersey and 1,989,741 transitions and 740,123 transversions in Kashmir cattle. A total of 34.37% missense, 0.296% non-sense, and 65.334% silent mutation were identified in Kashmiri cattle, and 37.701% missense, 0.409% non-sense, and 61.89% silent mutations were identified in Jersey cattle. The common SNPs were filtered out and further analyses were carried out on high-impact SNPs and INDELs specific to Jersey (442 SNPs and 169 INDELs) and Kashmiri (464 SNPs and 220 INDELs) cattle. A total of 351 high-impact variations were identified as frameshift, 400 stop-gained variations, 40 stop lost, 38 start lost, and one SNP in the 5'-UTR region (**Supplementary Tables 4, 5**). SNP distribution on different chromosomes in both Jersey and Kashmiri cattle are shown in **Figure 1**.

Analysis of Genes With SNPs and INDELs

Functional annotation suggests the enrichment (p -value < 0.05) of signaling pathways (like AMP activated protein kinase (AMPK) and tumor growth factor (TGF) beta), insulin resistance, and high antigen processing in Kashmiri cows. Jersey cattle enrichment analysis shows genes with high-impact SNPs were involved mainly in nucleic acid (specifically pyrimidine metabolism) and nucleotide (specifically glycine, serine, and threonine) metabolism pathways. Interestingly, the activity of primary immunodeficiency pathway, nucleotide excision repair pathway, glycerolipid metabolism, and phosphatidylinositol signaling pathway were significantly enriched in Jersey cattle (**Table 1**). Gene ontology (GO) analysis strongly suggests (FDR corrected p -value < 0.05) genes with SNPs in Kashmiri cows were mainly involved in the binding process (enzyme and ribonucleotide binding) and antigen processing (**Figure 2**). In

Jersey cattle, mutated genes were involved mainly in ion binding and ATPase activity (**Figure 2**).

DISCUSSION

Kashmiri cattle are poor performing and differ greatly from Jersey in dairy production characteristics. A comprehensive phenotypic analysis with respect to milk production traits was carried out in our previous study (Bhat et al., 2019) which showed higher milk yield and protein content for Jersey cattle as compared with the Kashmiri indigenous cattle. The average milk production of Jersey and Kashmiri cattle varies between 6 and 10 kg/day and 3 and 5 kg/day, respectively. The protein content in Jersey cattle ranged from 2.91 to 3.36%, and the corresponding values for Kashmiri cattle were 2.81–3.21%. The Jersey is among the top milk producer cattle breeds and is routinely used to upgrade the milk-producing capacity of the Kashmiri cattle by crossbreeding practices. It becomes imperative to understand the difference at multiple levels such as gene expression and SNPs in the coding and regulatory region of these breeds. For this purpose, we have analyzed RNA-seq data and performed the comparative study between both the breeds. A SNP detection analysis was performed using sequencing reads from nine Jersey and eight Kashmiri cows to determine putative polymorphisms in genes involved in the milk production. A total 442 and 464 high-impact SNPs were identified from Jersey and Kashmir cattle, respectively. Out of all the highly confident SNPs identified, a significant portion of them (34.37 and 37.701%) were missense mutations distributed in 29 chromosomes. Furthermore, the functional analysis of the genes showed that they were involved in several crucial biological and physiological processes in lactation. We observed TGF-beta, AMPK, and endocytosis show higher activity in Kashmiri cattle compared with Jersey breed. It is reported that TGF- β 1 expression increases in MECs during of mammary gland involution in mouse (Atwood et al., 1995), goat (Wareski et al., 2001), sow (Motyl et al., 2001), and cow (Zarzyńska et al., 2007). TGF- β 1 plays a key role in the involution of the bovine mammary gland by prompting apoptosis and autophagy in bovine mammary luminal epithelial cells (Kolek et al., 2003; Gajewska et al., 2005). Apoptosis is the foremost kind of cell death responsible for the involution of the bovine mammary gland, but type II programmed cell death (PCD II) is also observed: autophagic cell death. The MEC apoptosis is a marked indication of ending milking at the beginning of the dry period (Zarzyńska and Motyl, 2008). We presume the shorter lactation length in Kashmiri cattle could be one of the reasons for the increased expression of apoptotic signaling pathway toward the beginning of dry period. Moreover, it is known that some milk trait genes (e.g., genes in apoptosis pathway) are not solely expressed in MECs but also by other cell types like leukocytes (Boutinaud et al., 2015).

The levels of TGF- β 1 and TGF- β 2 decrease with advancing chronological age, and colostrum contains the highest levels

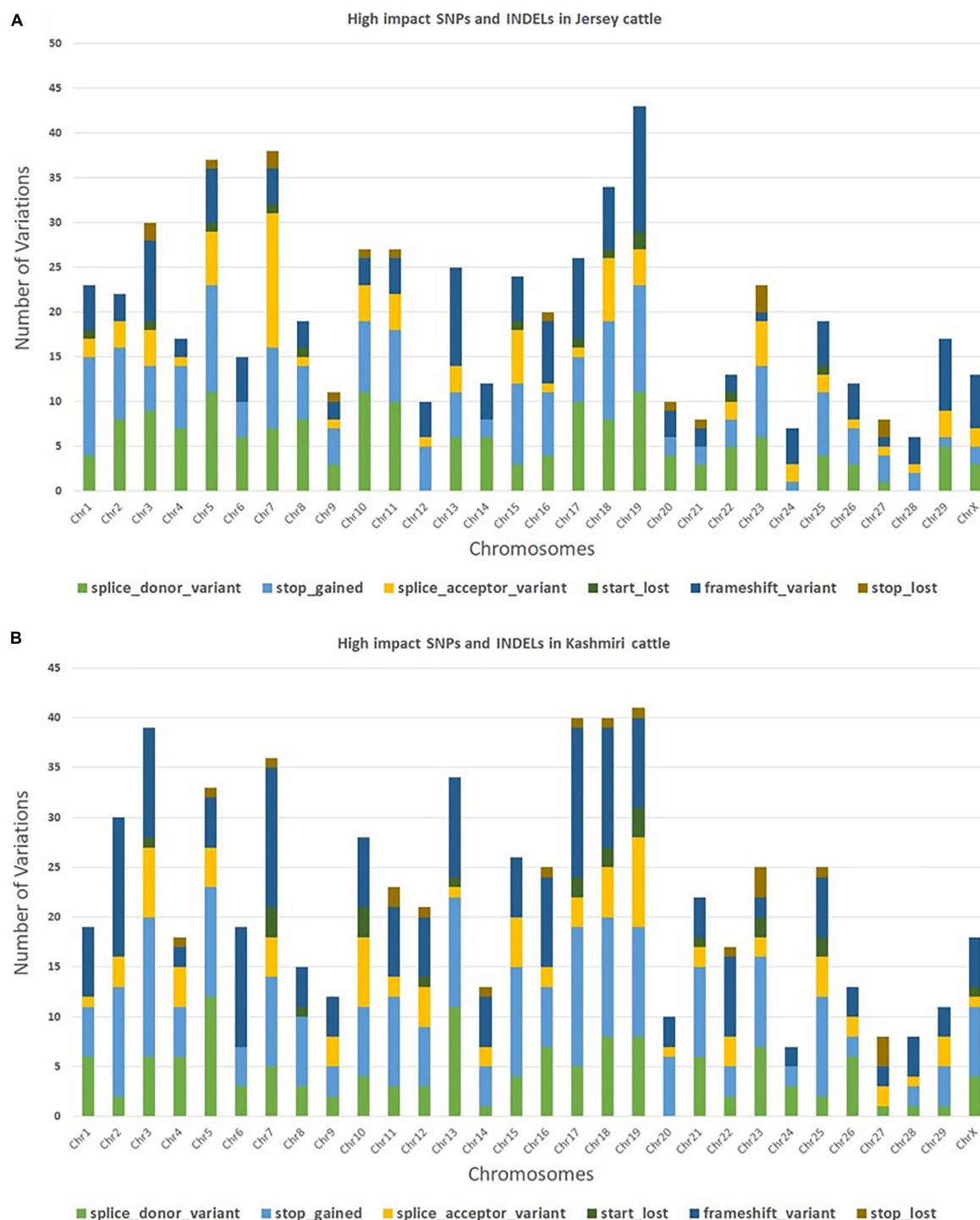


FIGURE 1 | Chromosomal distribution of high impact SNPs and Indels in Jersey (A) and Kashmiri (B) cattle.

for both (Frost et al., 2014). The pronounced role of AMPK is to regulate catabolic and anabolic processes. Negative energy balance (NEB) is generally witnessed in milking cows particularly during lactation. Increase in ADP or AMP is usually related with a NEB. It has been found that during maximum lactation, the

energy intake in dairy cows is not able to fulfill the requirement for milk production. The AMPK signaling has been found greatly active throughout this period, which results in significant reduction in the milk of cows (Eastham et al., 1988). AMPK regulates lipid metabolism by altering the transcription of the

lipogenic gene and the posttranslational modification of the major enzymes involved in lipid synthesis. In addition, AMPK activation was found to repress global protein synthesis *via* inhibition of mTOR signals in bovine mammary epithelial cells

(Appuhamy et al., 2014). These changes along with the effect of lower plane of nutrition during formative phase of the animals and consequent poor body condition may contribute to the lower concentrations of total milk proteins in Kashmiri cattle. This

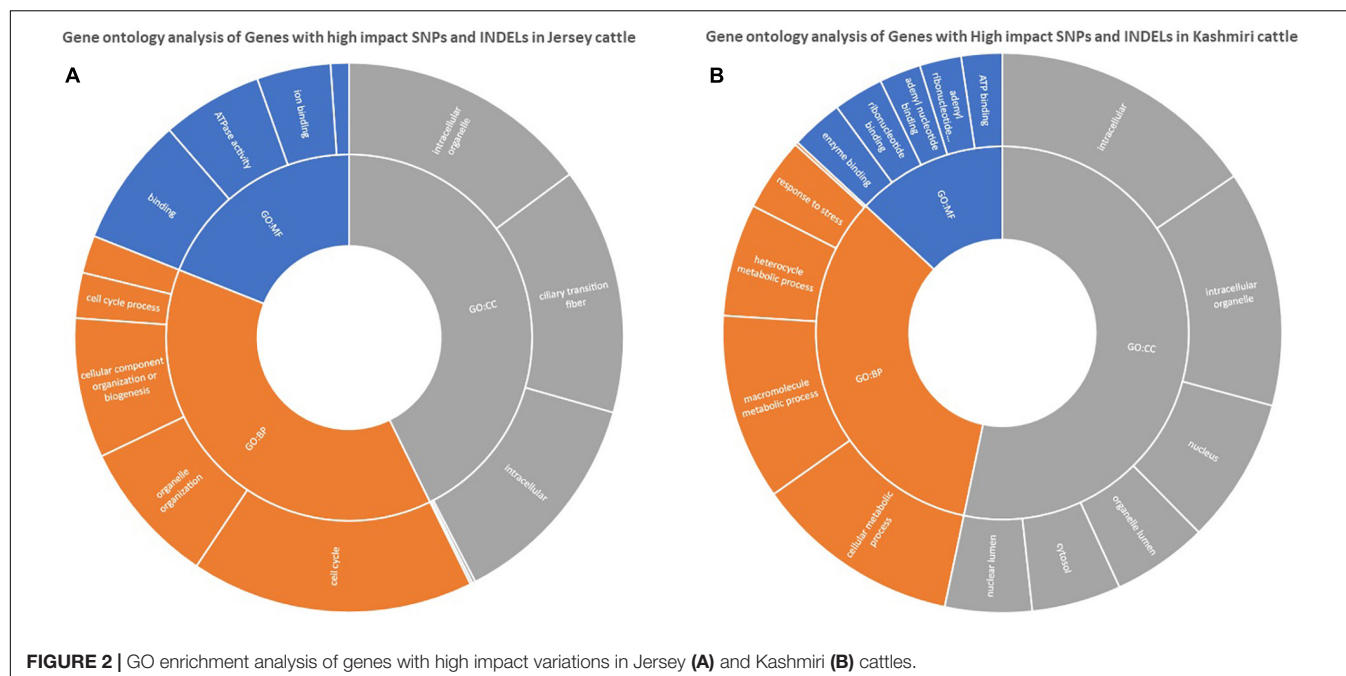
TABLE 1 | Pathways affected by high-impact SNPs in Jersey (A) and Kashmiri (B) cattle.

A. Enriched pathways in Jersey cattle.

Pathway	p-value	Input number	Input
Antifolate resistance	0.002919	5	ENSBTAG00000047764, ENSBTAG00000007599, ENSBTAG00000023309, ENSBTAG00000045751, and ENSBTAG00000031500
Nucleotide excision repair	0.005815	5	ENSBTAG00000032527, ENSBTAG00000021773, ENSBTAG00000014595, ENSBTAG00000007362, and ENSBTAG00000014727
ABC transporters	0.008163	5	ENSBTAG00000047764, ENSBTAG00000045751, ENSBTAG000000006921, ENSBTAG00000023309, and ENSBTAG00000002747
Glycerolipid metabolism	0.013722	5	ENSBTAG00000021695, ENSBTAG00000012060, ENSBTAG00000045746, ENSBTAG00000018201, and ENSBTAG00000011917
Glycine, serine, and threonine metabolism	0.022811	4	ENSBTAG00000021695, ENSBTAG00000031814, ENSBTAG00000031500, and ENSBTAG00000004510
Phosphatidylinositol signaling system	0.026138	6	ENSBTAG00000002010, ENSBTAG00000002350, ENSBTAG00000003809, ENSBTAG00000001030, ENSBTAG00000013116, and ENSBTAG00000020715
Pyrimidine metabolism	0.04586	4	ENSBTAG00000021830, ENSBTAG00000005152, ENSBTAG00000008428, and ENSBTAG00000011527
Primary immunodeficiency	0.048001	3	ENSBTAG00000005280, ENSBTAG00000023144, and ENSBTAG00000006452

B. Enriched pathways in Kashmiri cattle.

Antigen processing and presentation	0.0447	9	ENSBTAG00000045795, ENSBTAG00000006270, ENSBTAG00000002069, ENSBTAG00000038128, ENSBTAG00000019386, ENSBTAG00000012208, ENSBTAG00000020116, ENSBTAG00000005182, and ENSBTAG00000019588
Innate immune system	0.002183	46	ENSBTAG00000047856, ENSBTAG00000033662, ENSBTAG00000015520, ENSBTAG00000002902, ENSBTAG00000010639, ENSBTAG00000046188, ENSBTAG00000020772, ENSBTAG00000021474, ENSBTAG00000038797, ENSBTAG00000031641, ENSBTAG00000045861, ENSBTAG00000012467, ENSBTAG00000005660, ENSBTAG00000003401, ENSBTAG00000003774, ENSBTAG00000019020, ENSBTAG00000017866, ENSBTAG00000017401, ENSBTAG00000018403, ENSBTAG00000016415, ENSBTAG00000015187, ENSBTAG00000005574, ENSBTAG00000010951, ENSBTAG00000027684, ENSBTAG00000002854, ENSBTAG00000023144, ENSBTAG00000049346, ENSBTAG00000001014, ENSBTAG00000016021, ENSBTAG00000001609, ENSBTAG00000020433, ENSBTAG00000012780, ENSBTAG00000018661, ENSBTAG00000006270, ENSBTAG00000021144, ENSBTAG00000007964, ENSBTAG00000002069, ENSBTAG00000019386, ENSBTAG00000012208, ENSBTAG00000020116, ENSBTAG00000005182, ENSBTAG00000044040, ENSBTAG00000053934, ENSBTAG00000047699, ENSBTAG00000019250, and ENSBTAG00000003121
TGF beta signaling pathway	0.008891	12	ENSBTAG00000004154, ENSBTAG00000047856, ENSBTAG00000008300, ENSBTAG00000018127, ENSBTAG00000020053, ENSBTAG00000015720, ENSBTAG00000010662, ENSBTAG00000021145, ENSBTAG00000001609, ENSBTAG00000019289, ENSBTAG00000010649, and ENSBTAG00000006919
Insulin resistance	0.011672	8	ENSBTAG00000009175, ENSBTAG00000016336, ENSBTAG00000017024, ENSBTAG00000001400, ENSBTAG00000017866, ENSBTAG00000017639, ENSBTAG000000039958, and ENSBTAG00000002917
Osteoclast differentiation	0.017121	8	ENSBTAG00000007213, ENSBTAG00000007531, ENSBTAG00000001400, ENSBTAG00000019250, ENSBTAG00000021358, ENSBTAG00000003305, ENSBTAG00000021842, and ENSBTAG00000001609
AMPK signaling pathway	0.018729	8	ENSBTAG00000002883, ENSBTAG00000009175, ENSBTAG00000000754, ENSBTAG00000017024, ENSBTAG00000001400, ENSBTAG000000017866, ENSBTAG000000039958, and ENSBTAG00000002917
Endocytosis	0.046451	11	ENSBTAG00000019386, ENSBTAG00000019072, ENSBTAG00000002069, ENSBTAG00000007964, ENSBTAG00000005182, ENSBTAG00000020116, ENSBTAG00000027684, ENSBTAG00000018959, ENSBTAG00000015720, ENSBTAG00000007120, and ENSBTAG00000010543



statement however, needs to be validated. Similar studies were reported in lactating goats (Cai et al., 2020).

An endocytosis pathway was found to be enriched in the milk of Kashmiri cattle. The mechanism for internalization (endocytosis) and release (exocytosis) are governed by the physiological phase of the mammary gland (McShane and Zerial, 2008). Furthermore, Truchet and Ollivier-Bousquet (2009) reported that the mammary epithelium enrichment organizes the absorption of milk precursors and transport of milk components to produce milk of reasonably constant composition, provided the mammary gland is healthy.

Several important metabolic pathways that were found to be enriched exclusively in Jersey cattle include glycolipid, pyrimidine, and amino acid metabolism (specifically glycine, serine, and threonine) and other important pathways like phosphatidylinositol signaling, Antifolate resistance, ABC transporters pathway, and nucleotide excision repair. Amino acid (AA) metabolism plays a critical role in milk production by regulating maternal endocrine status, blood flow through the lactating mammary gland, and activation of mechanistic (mammalian) target rapamycin (mTOR) signaling (Wu et al., 2010). The higher activity of glycine, serine, and threonine metabolism pathways in Jersey cattle may be particularly important in lactation initiation and higher milk production (Li and Jiang, 2019). TCA cycle and glycine, serine, and threonine metabolism pathways are the most important and work together in the mammary gland for lactation initiation (Sun et al., 2015, 2017). It is a known fact that AAs significantly contribute to liver gluconeogenesis in early lactation. Glycine, serine, and alanine, comprised more than 65% of the liver uptake of glucogenic AA during the periparturient period (Larsen and Kristensen, 2009, 2012). In Jersey cattle, the phosphatidylinositol 3-kinase (PI3K)/protein kinase B (Akt) signaling pathway

emerged as the most enriched pathway. PI3K-Akt pathway mediates glucose absorption into cells and is important for normal insulin-mediated glucose metabolism (Taniguchi et al., 2006). Since Jersey cattle produce milk in higher quantities than Kashmiri cattle, it requires a high-energy source. The enrichment of glycerolipid metabolic pathway in Jersey *via* glycerolipid pathway converts glycerol into glucose thus providing sufficient energy for maintaining high milk production (Sun et al., 2017).

In Kashmiri cattle, we found variations in the major histocompatibility complex (MHC) region. In cattle, polymorphism in the MHC class II genes influences both the magnitude and specificity of antigen-specific T cells to various diseases like mastitis (Hameed et al., 2006). In our previous study, we found that a wide range of proteins like apelin, acid glycoprotein, CD14 antigens, and lactoferrin, involved in immune response and host defense were highly expressed in mammary gland of Kashmiri cattle (Bhat et al., 2020). The role of the MHC in immune response makes a MHC an attractive candidate gene to study associations with disease resistance or susceptibility in cattle. High variability in many MHC genes plays a role in the recognition of bacteria (Apanius et al., 1997). Hedrick and Kim (1999) demonstrate the relation between MHC variation and resistance to bacterial infections. There are several well-documented cases in which specific MHC haplotypes or genotypes provide resistance to bacteria (Hill et al., 1991; Carrington et al., 1999).

We also found variations in genes regulating TGF- β (SMAD4, BMP7) and mTOR signaling in Kashmiri cattle (Supplementary Figures 1, 2) (Bhat et al., 2017). SMAD4 proteins serve as crucial components of TGF- β signaling, which negatively regulates cell growth and promotes apoptosis of epithelial cells. The rate of decline in milk yield with stage of lactation is strongly influenced by the rate of cell death by apoptosis in the lactating gland

(Stefanon et al., 2002). Such changes depict the early backdrop for dryness in Kashmiri cattle. SMAD4 has been found to play important roles in various biological processes. In humans, genetic variants in the SMAD4 have been found to play a protective role in various types of cancers (Wu et al., 2010). A high-impact stop-gained variation on SMAD4 (c.1165C) and frameshift variation on BMP7 (c.693delC) could possibly dysregulate TGF- β pathway in Kashmiri cattle. Further study is required to study the effect of these variations on the TGF- β signaling pathway.

It is widely accepted that mTOR is a key regulator of milk protein synthesis, and most reports were concerned with the role of AAs in the regulation of P-mTOR on Ser2448 in milk protein synthesis (Appuhamy et al., 2012; Apelo et al., 2014). mTORC1 signaling mediates the cell surface level of Wnt receptor frizzled FZD and plays an essential role in embryogenesis and homeostasis Chen et al. (2020). In cattle, the milk protein synthesis was upregulated through activation of the mTOR pathway (Gao et al., 2015). We found double frameshift variations on FZD9 (c.484delG and c.1563delT) in Kashmiri cattle and could be the probable reason for its low milk production.

CONCLUSION

The SNP analysis based on RNA sequencing demonstrated a clear distinction between Jersey and Kashmiri cattle in milk production traits. In Kashmiri cattle, the high-impact SNP variants were involved in adaptive immunity and resistance against mammary gland infectious diseases. Whereas, in Jersey cattle, enriched pathways were mainly involved in maintenance of lactation and production. These findings provide insights in

genetic variation of the breeds that can be used for genomic selection of the animals.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/sra/SRR6324372>.

AUTHOR CONTRIBUTIONS

SA designed the study and wrote the manuscript. BB and MY performed the data analysis. SB and SM helped in the writing of the manuscript. MR and MI helped in data interpretation and analysis. RS and NG did proofreading of the manuscript. All authors revised and approved the final version of the manuscript.

FUNDING

The Department of Biotechnology, Ministry of Science and Technology, Government of India is duly acknowledged for supporting the study under grant BT/PR114 08/AAQ/1/591/2014.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.666015/full#supplementary-material>

REFERENCES

- Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (accessed January 8, 2019).
- Apanius, V., Penn, D., Slev, P. R., Ruff, L. R., and Potts, W. K. (1997). The nature of selection on the major histocompatibility complex. *Crit. Rev. Immunol.* 17, 179–224.
- Apelo, S. A., Knapp, J., and Hanigan, M. (2014). Invited review: current representation and future trends of predicting amino acid utilization in the lactating dairy cow. *J. Dairy Sci.* 97, 4000–4017. doi: 10.3168/jds.2013-7392
- Appuhamy, J. R. N., Knoebel, N. A., Nayananjali, W. D., Escobar, J., and Hanigan, M. D. (2012). Isoleucine and leucine independently regulate mtor signaling and protein synthesis in mac-t cells and bovine mammary tissue slices. *J. Nutr.* 142, 484–491. doi: 10.3945/jn.111.152595
- Appuhamy, J., Nayananjali, W., England, E., Gerrard, D., Akers, R., and Hanigan, M. (2014). Effects of amp-activated protein kinase (ampk) signaling and essential amino acids on mammalian target of rapamycin (mtor) signaling and protein synthesis rates in mammary cells. *J. Dairy Sci.* 97, 419–429. doi: 10.3168/jds.2013-7189
- Arora, R., Sharma, A., Sharma, U., Girdhar, Y., Kaur, M., Kapoor, P., et al. (2019). Buffalo milk 205 transcriptome: a comparative analysis of early, mid and late lactation. *Sci. Rep.* 9:5993.
- Atwood, C. S., Ikeda, M., and Vonderhaar, B. K. (1995). Involution of mouse mammary glands in whole organ culture: a model for studying programmed cell death. *Biochem. Biophys. Res. Commun.* 207, 860–867. doi: 10.1006/bbrc.1995.1265
- Auldust, M. J., O'Brien, G., Cole, D., Macmillan, K. L., and Grainger, C. (2007). Effects of varying lactation length on milk production capacity of cows in pasture-based dairying systems. *J. Dairy Sci.* 90, 3234–3241. doi: 10.3168/jds.2006-683
- Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 16, 545–552.
- Bhat, B., Ganai, N. A., Andrabi, S. M., Shah, R. A., and Singh, A. (2017). Tm-aligner: multiple sequence alignment tool for transmembrane proteins with reduced time and improved accuracy. *Sci. Rep.* 7:12543.
- Bhat, B., Yaseen, M., Singh, A., Ahmad, S. M., and Ganai, N. A. (2021). Identification of potential key genes and pathways associated with the pashmina fiber initiation using rna-seq and integrated bioinformatics analysis. *Sci. Rep.* 11:1766.
- Bhat, S. A., Ahmad, S. M., Ibeagha-Awemu, E. M., Bhat, B. A., Dar, M. A., Mumtaz, P. T., et al. (2019). Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between Jersey and Kashmiri cattle. *PLoS One* 14:e0211773. doi: 10.1371/journal.pone.0211773
- Bhat, S. A., Ahmad, S. M., Ibeagha-Awemu, E. M., Mobashir, M., Dar, M. A., Mumtaz, P. T., et al. (2020). Comparative milk proteome analysis of Kashmiri

- and Jersey cattle identifies differential expression of key proteins involved in immune system regulation and milk quality. *BMC Genomics* 21:161. doi: 10.1186/s12864-020-6574-4
- Boutinaud, M., Herve, L., and Lollivier, V. (2015). Mammary epithelial cells isolated from milk are a valuable, non-invasive source of mammary transcripts. *Front. Genet.* 6:323. doi: 10.3389/fgene.2015.00323
- Cai, J., Wang, D., Zhao, F. Q., Liang, S., and Liu, J. (2020). AMPK-mTOR pathway is involved in glucose-modulated amino acid sensing and utilization in the mammary glands of lactating goats. *J. Anim. Sci. Biotechnol.* 11, 1–12.
- Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., et al. (1999). Hla and hiv-1: heterozygote advantage and b* 35-cw* 04 disadvantage. *Science* 283, 1748–1752. doi: 10.1126/science.283.5408.1748
- Chen, M., Amado, N., Tan, J., Reis, A., Ge, M., Abreu, J. G., et al. (2020). Tmem79/matrin defines a pathway for frizzled regulation and is required for xenopus embryogenesis. *Elife* 9:e56793.
- Dekkers, J. C. (2004). Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.* 82, E313–E328.
- Dias, M. M., Cánovas, A., Mantilla-Rojas, C., Riley, D. G., Luna-Nevarez, P., Coleman, S. J., et al. (2017). SNP detection using RNA-sequences of candidate genes associated with puberty in cattle. *Genet. Mol. Res.* 16:gmr16019522.
- Eastham, P. R., Smith, W. C., Whittemore, C. T., and Phillips, P. (1988). Responses of lactating sows to food level. *Anim. Sci.* 46, 71–77. doi: 10.1017/s0003356100003123
- Frost, B. L., Jilling, T., Lapin, B., Maheshwari, A., and Caplan, M. S. (2014). Maternal breast milk transforming growth factor-beta and feeding intolerance in preterm infants. *Pediatr. Res.* 76, 386–393. doi: 10.1038/pr.2014.96
- Gajewska, M., Gajkowska, B., and Motyl, T. (2005). Apoptosis and autophagy induced by $\text{tgf-}\beta 1$ in bovine 247 mammary epithelial bme-uv1 cells. *J. Physiol. Pharmacol.* 56, 143–157.
- Gao, H.-n., Hu, H., Zheng, N., and Wang, J.-q. (2015). Leucine and histidine independently regulate milk protein synthesis in bovine mammary epithelial cells via mtor signaling pathway. *J. Zhejiang Univ. Sci. B* 16, 560–572. doi: 10.1631/jzus.b1400337
- Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., et al. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139, 907–920. doi: 10.1093/genetics/139.2.907
- Goldman, A. S. (2000). Modulation of the gastrointestinal tract of infants by human milk. Interfaces and interactions. An evolutionary perspective. *J. Nutr.* 130, 426S–431S.
- Hameed, K. G. A., Sender, G., and Mayntz, M. (2006). Major histocompatibility complex polymorphism and mastitis resistance—a review. *Anim. Sci. Pap. Rep.* 24, 11–25.
- Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53, 876–883. doi: 10.1139/g-10-076
- Hedrick, P. W., and Kim, T. J. (1999). “Genetics of complex polymorphisms: parasites and maintenance of mhc variation,” in *Evolutionary Genetics From Molecules to Morphology*, eds R. S. Singh and C. K. Krimbas (Cambridge: Cambridge University Press).
- Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., et al. (1991). Common west african hla antigens are associated with protection from severe malaria. *Nature* 352, 595–600. doi: 10.1038/352595a0
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 49, D545–D551. doi: 10.1093/nar/gkaa970
- Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4
- Kolek, O., Gajkowska, B., Godlewski, M., and Motyl, T. (2003). Antiproliferative and apoptotic effect of $\text{tgf-}\beta 1$ in bovine mammary epithelial bme-uv1 cells. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* 134, 417–430. doi: 10.1016/s1532-0456(02)00249-1
- Larsen, M., and Kristensen, N. B. (2009). Effect of abomasal glucose infusion on splanchnic amino acid metabolism in periparturient dairy cows. *J. Dairy Sci.* 92, 3306–3318. doi: 10.3168/jds.2008-1889
- Larsen, M., and Kristensen, N. B. (2012). Effects of glucogenic and ketogenic feeding strategies on splanchnic glucose and amino acid metabolism in postpartum transition Holstein cows. *J. Dairy Sci.* 95, 5946–5960. doi: 10.3168/jds.2012-5458
- Li, Z., and Jiang, M. (2019). Metabolomic profiles in yak mammary gland tissue during the lactation cycle. *PLoS One* 14:e0219220. doi: 10.1371/journal.pone.0219220
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). 1000 genome project data processing subgroup, the sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Martin, M. (2011). Cutadapt removes adapter sequences from highthroughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.731.200
- McShane, M. P., and Zerial, M. (2008). Survival of the weakest: signaling aided by endosomes. *J. Cell Biol.* 182, 823–825. doi: 10.1083/jcb.2008.07165
- Mech, A., Dhali, A., Prakash, B., and Rajkhowa, C. (2008). Variation in milk yield and milk composition during the entire lactation period in Mithun cows (*Bos frontalis*). *Livest. Res. Rural Dev.* 20, 56–57.
- Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92, 255–264. doi: 10.1016/j.ygeno.2008.07.001
- Motyl, T., Gajkowska, B., Wojewódzka, U., Waręski, P., Rekiel, A., and Płoszaj, T. (2001). Expression of apoptosis-related proteins in involuting mammary gland of sow. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* 128, 635–646. doi: 10.1016/s1096-4959(00)00334-1
- Pareek, C. S., Błaszczyk, P., Dziuba, P., Czarnik, U., Fraser, L., Sobiech, P., et al. (2017). Single nucleotide polymorphism discovery in bovine liver using rna-seq technology. *PLoS One* 12:e0172687. doi: 10.1371/journal.pone.0172687
- Pareek, C. S., Smoczyński, R., Kadarmideen, H. N., Dziuba, P., Błaszczyk, P., Sikora, M., et al. (2016). Single nucleotide polymorphism discovery in bovine pituitary gland using RNA-Seq technology. *PLoS One* 11:e0161370. doi: 10.1371/journal.pone.0161370
- Picard toolkit (2019). *Broad Institute, GitHub Repository*. Broad Institute. Available online at: <http://broadinstitute.github.io/picard/>
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T., Carneiro, M. O., Van der Auwera, G. A., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* [Preprint]. doi: 10.1101/201178 BioRxiv: 201178.
- Reinhardt, T. A., and Lippolis, J. D. (2006). Bovine milk fat globule membrane proteome. *J. Dairy Res.* 73, 406–416. doi: 10.1017/s0022029906001889
- Séverin, S., and Wenshui, X. (2005). Milk biologically active components as nutraceuticals. *Crit. Rev. Food Sci. Nutr.* 45, 645–656. doi: 10.1080/10408690490911756
- Stefanon, B., Colitti, M., Gabai, G., Knight, C. H., and Wilde, C. J. (2002). Mammary apoptosis and lactation persistence in dairy animals. *J. Dairy Res.* 69, 37–52. doi: 10.1017/s0022029901005246
- Sun, H. Z., Shi, K., Wu, X. H., Xue, M. Y., Wei, Z. H., Liu, J. X., et al. (2017). Lactation-related metabolic mechanisms investigated based on mammary gland metabolomics and 4 biofluids’ metabolomics relationships in dairy cows. *BMC Genomics* 18:936. doi: 10.1186/s12864-017-4314-1
- Sun, H. Z., Wang, D. M., Wang, B., Wang, J. K., Liu, H. Y., Guan, L. L., et al. (2015). Metabolomics of four biofluids from dairy cows: potential biomarkers for milk production and quality. *J. Proteome Res.* 14, 1287–1298. doi: 10.1021/pr501305g
- Taniguchi, C. M., Emanuelli, B., and Kahn, C. R. (2006). Critical nodes in signalling pathways: insights into insulin action. *Nat. Rev. Mol. Cell Biol.* 7, 85–96. doi: 10.1038/nrm1837

- Truchet, S., and Ollivier-Bousquet, M. (2009). Mammary gland secretion: hormonal coordination of endocytosis and exocytosis. *Animal* 3, 1733–1742. doi: 10.1017/s1751731109990589
- Wang, W., Wang, H., Tang, H., Gan, J., Shi, C., Lu, Q., et al. (2018). Genetic structure of six cattle populations revealed by transcriptome-wide SNPs and gene expression. *Genes Genom.* 40, 715–724. doi: 10.1007/s13258-018-0677-1
- Wareski, P., Motyl, T., Ryniewicz, Z., Orzechowski, A., Gajkowska, B., Wojewódzka, U., et al. (2001). Expression of apoptosis-related proteins in the mammary gland of goats. *Small Rumin. Res.* 40, 279–289. doi: 10.1016/s0921-4488(01)00178-x
- Wu, D. M., Zhu, H.-X., Zhao, Q.-H., Zhang, Z.-Z., Wang, S.-Z., Wang, M.-L., et al. (2010). Genetic variations in the smad4 gene and gastric cancer susceptibility. *World J. Gastroenterol.* 16:5635. doi: 10.3748/wjg.v16.i44.5635
- Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322.
- Zarzyńska, J., and Motyl, T. (2008). Apoptosis and autophagy in involuting bovine mammary gland. *J. Physiol. Pharmacol.* 59(Suppl. 9), 275–288.
- Zarzyńska, J., Gajkowska, B., Wojewódzka, U., Dymnicki, E., and Motyl, T. (2007). Apoptosis and autophagy in involuting bovine mammary gland is accompanied by up-regulation of TGF-beta1 and suppression of somatotrophic pathway. *Polish J. Vet. Sci.* 10, 1–9.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
- Copyright © 2021 Ahmad, Bhat, Bhat, Yaseen, Mir, Raza, Iquebal, Shah and Ganai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome-Wide Association Studies in Indian Buffalo Revealed Genomic Regions for Lactation and Fertility

Vikas Vohra^{*†}, Supriya Chhotaray[†], Gopal Gowane, Rani Alex, Anupama Mukherjee, Archana Verma and Sitangsu Mohan Deb

Buffalo Breeding Lab, Animal Genetics and Breeding Division, Indian Council of Agricultural Research-National Dairy Research Institute, Karnal, India

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University
of Agricultural Sciences
and Technology, India

Reviewed by:

Gao Huijiang,
Institute of Animal Sciences, Chinese
Academy of Agricultural Sciences,
China

Hadi Atashi,
Shiraz University, Iran
Alessandro Bagnato,
University of Milan, Italy

*Correspondence:

Vikas Vohra
vohravikas@gmail.com

[†] These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 16 April 2021

Accepted: 16 August 2021

Published: 20 September 2021

Citation:

Vohra V, Chhotaray S, Gowane G,
Alex R, Mukherjee A, Verma A and
Deb SM (2021) Genome-Wide
Association Studies in Indian Buffalo
Revealed Genomic Regions
for Lactation and Fertility.
Front. Genet. 12:696109.
doi: 10.3389/fgene.2021.696109

Murrah breed of buffalo is an excellent dairy germplasm known for its superior milk quality in terms of milk fat and solids-not-fat (SNF); however, it is often reported that Indian buffaloes had lower lactation and fertility potential compared to the non-native cattle of the country. Recent techniques, particularly the genome-wide association studies (GWAS), to identify genomic variations associated with lactation and fertility traits offer prospects for systematic improvement of buffalo. DNA samples were sequenced using the double-digestion restriction-associated DNA (RAD) tag genotyping-by-sequencing. The bioinformatics pipeline was standardized to call the variants, and single-nucleotide polymorphisms (SNPs) qualifying the stringent quality check measures were retained for GWAS. Over 38,000 SNPs were used to perform GWAS on the first two principal components of test-day records of milk yields, fat percentages, and SNF percentages, separately. GWAS was also performed on 305 days' milk yield; lactation persistency was estimated through the rate of decline after attaining the peak yield method, along with three other standard methods; and breeding efficiency, post-partum breeding interval, and age at sexual maturity were considered fertility traits. Significant association of SNPs was observed for the first principal component, explaining the maximum proportion of variation in milk yield. Furthermore, some potential genomic regions were identified to have a potential role in regulating milk yield and fertility in Murrah. Identification of such genomic regions shall help in carrying out an early selection of high-yielding persistent Murrah buffaloes and, in the long run, would be helpful in shaping their future genetic improvement programs.

Keywords: Murrah, buffaloes, GWAS, lactation persistency, milk yield, fertility

INTRODUCTION

Buffalo (*Bubalus bubalis*) is an imperative livestock species and act as a key component for improving agricultural economy and supplying milk, meat, and draft power. The buffalo population across the world was recently estimated to be 194 million, 97% of which were present in Asia (FAO, 2014). Buffalo is well known for its high milk quality, with higher fat (6.4–8.0% vs. 4.1–5.0%) and protein (4.0–4.5% vs. 3.4–3.6%) contents than cow milk (Venturini et al., 2014; Khedkar et al., 2016). The concentration of these milk constituents offers a higher economic return of the buffalo milk and increases the demand of value-added products like mozzarella. Buffaloes are an

integral and crucial genetic resource in Indian dairy industry, contributing 49% to the total milk produced according to the basic animal husbandry statistics, 2019¹. About 63% of global buffalo milk production and 95% of Asian buffalo milk production is contributed by Indian buffaloes (FAO, 2014). Buffaloes are more adaptable to harsh environments and often resist various bovine tropical diseases. However, the poor reproductive efficiency of buffaloes limits its potential. Buffaloes exhibit higher age at puberty and maturity, longer postpartum breeding interval, and low conception rates (Gordon, 1996; Warriach et al., 2015; Raina et al., 2016). Buffaloes in the field condition also suffer from short lactation of 252–270 days as compared to the standard 300 days (Ganguli, 1981; Griffiths, 2010). This underlines the scope for improving the milk production and fertility potential through implementation of systematic breeding programs for buffaloes in the country. However, in India, very few works have been done in order to identify the infinitesimally large number of underlying loci regulating the expression of complex traits such as milk production, lactation persistency and fertility.

Genome wide association studies (GWAS) in buffaloes for lactation traits have been mainly limited to the use of bovine single-nucleotide polymorphism (SNP) chip (Wu et al., 2013; Venturini et al., 2014) and Affymetrix's buffalo 90K SNP chip (de Camargo et al., 2015; Iamartino et al., 2017; Liu et al., 2018). There have been very few GWAS conducted in India covering all aspects of production and reproduction performance due to constraints of cost incurred and organized large-scale genotyping programs. High-density SNP panels are a prerequisite for GWAS, which have led to developments of cost-effective and efficient next-generation sequencing (NGS) technologies such as reduced representation of genomic libraries (RRLs) (Van Tassel et al., 2008). The flexibility, robustness, and low cost of double-digestion restriction-associated DNA (RAD) tag genotyping-by-sequencing (ddRAD-GBS) technique renders it suitable for identification of SNPs in any species for GWAS (Davey et al., 2011; Elshire et al., 2011).

Hence, the present study was conducted to identify novel SNPs associated with milk production, composition, lactation persistency, and fertility traits at the genomic level using the genotype-by-sequencing technique in Murrah buffalo, India's major buffalo breed and milch animal of the nation.

MATERIALS AND METHODS

Sampling, Data Recording, and Genotyping

A total of 672 test-day records on each trait, i.e., milk yield (TDMY), fat percentage (TDFP), and solids-not-fat (SNF) percentage (TDSNF) were collected from 96 female Murrah buffaloes reared at LRC, NDRI, Karnal, India (29.68°N and 76.99°E). Records of 96 buffaloes on 305 days' milk yield (305DMY), birth weight (bwt), age at

first calving (AFC), calving interval (CI), and age at sexual maturity (ASM) in months were collected. Other traits such as lactation persistency, postpartum breeding interval in days (PPBI), and breeding efficiency (BE) were derived from the primary phenotype records. Breeding efficiency was calculated for female buffaloes as described by Tomar (1965).

Lactation persistency was estimated by following four different methods:

- (I) Wood (1967) incomplete gamma function, where individuals were classified as persistent and non-persistent as described by Macciotta et al. (2005) considering a positive rate of incline as a favorable condition and a negative rate as an unfavorable condition for persistency
- (II) Johansson and Hansson (1940) method
- (III) Ludwick and Petersen (1943) method
- (IV) Mahadevan (1951) method

DNA was isolated from 96 Murrah buffaloes selected for the study following the standard phenol-chloroform method (Sambrook and Russell, 2006). Samples were further processed using the standard RAD protocol as described by Peterson et al. (2012). DNA double digestion was carried out with *SphI* and *MluCI* restriction enzymes. Adapters (P1 and P2) were prepared as per standard Illumina read multiplexing protocol using an inline barcode along with Illumina index for library preparation. After adapter ligation and size selection, samples were sequenced on a Illumina Hi-Seq 2000 platform.

Variant Calling

The NGS pipeline was standardized after incorporating a few modifications in the standard mpileup variant calling pipeline (Li, 2011), to call variants present in the Murrah population. The Mediterranean buffalo genome, having accession ID GCF_003121395.1, was retrieved from the NCBI dataset and used as a reference genome. Index and sequence dictionary files were created using the Burrows–Wheeler algorithm (BWA) (Li and Durbin, 2010) and PicardTools,² respectively. The quality of paired-end raw FASTQ files generated after sequencing was checked using FastQC (Andrews, 2010), and each report was combined through MultiQC (Ewels et al., 2016). Adapters were marked and trimmed using bbmap (Brian, 2014).³ The BWA-MEM algorithm was used to align the trimmed FASTQ sequences with the reference genome. Aligned files were coordinate-sorted, and duplicate reads were removed. Read group identifiers were updated using PicardTools. The quality of aligned BAM files was checked using qualimap (García-Alcalde et al., 2012). Variants were called using bcftools-mpileup (Li, 2011).

Quality Control of Variants

Only biallelic SNPs having more than 95% genotyping rate were retained for further GWAS. SNPs with a MAF < 0.05 and

¹<https://dahd.nic.in/about-us/divisions/statistics>

²<https://github.com/broadinstitute/picard>

³<https://sourceforge.net/projects/bbmap/>

deviating from the Hardy–Weinberg equilibrium at $p < 0.0001$ were removed from the dataset. SNPs in LD with $r^2 > 0.8$ were also removed. Only autosomal and X chromosome SNPs were retained for the final analysis. All the quality control operations and data preprocessing were performed using PLINK v1.9 (Chang et al., 2015).

Statistical Model

GWAS for milk yield, fat percentage, and SNF percentage were conducted on the principal components (PCs) instead of direct traits. Principal component analysis (PCA) was performed in the R programming environment (v4.0.3) on the 672 test-day records of each trait separately. As the first two PCs cumulatively account for most of the variation in the dataset, they were selected to be GWAS traits. The traits on which GWAS were performed were PCs (PC₁ and PC₂) of TDMYs, TDFPs, TDSNFs, and 305DMY; lactation persistency calculated by four different methods; age at sexual maturity in months; postpartum breeding interval (in days); and breeding efficiency. Genome-wide identity-by-state (IBS) for all pairs of individuals was checked. Multidimensional scaling (MDS) based on SNP information was done to check for the presence of any population stratification (**Supplementary Figure 1**) and was corrected by incorporating the first two MDS components as covariates in the model for GWAS. A genome-wide scan for significant SNPs considering only additive effects was accomplished through a simple regression model using PLINK v1.9 as described by Marees et al. (2018), where residuals were assumed to be normally and independently distributed. A linear regression model was fitted for determining the association between SNPs and continuous traits (Bush and Moore, 2012), while logistic regression was fitted for the binary trait (lactation persistent/non-persistent based on incomplete gamma function). The threshold for genome-wide significance was determined by correcting the p -values of the SNP association test with Benjamini–Hochberg’s false discovery rate (FDR) at 5 and 10% levels (Benjamini and Hochberg, 1995) using the “R” package fuzzySim v3.0 (Barbosa, 2020). A genome-wide significant threshold was set at FDR 5% and a suggestive threshold at 10% by calculating a nominal p -value for the largest index i for which $P_{(i)} \leq (i/m) \times q$, where i = rank of the SNPs, m = no. of individual tests performed, and q = either 0.05 or 0.1 (Glickman et al., 2014). The results were plotted as Manhattan plots and Q-Q plots using the “qqman” package of R.

Linear regression model used for GWAS:

$$y = \beta_0 + x\beta_1 + C_1\beta_2 + C_2\beta_3 + e$$

where y = trait, x = additive effect of SNPs, C_1 = first component of MDS, C_2 = second component of MDS, β_0 = intercept term, β_1 = regression coefficient representing the strength of association between SNP x and trait y , β_2 = regression coefficient of C_1 , β_3 = regression coefficient of C_2 , and e = residuals or noise not explained by SNPs.

The model used for GWAS on 305DMY, however, included four covariates consisting of the first component of MDS, birth weight, age at first calving, and calving interval.

SNP Mapping and Pathway Enrichment

SNPs were identified as genic if present within the genes or intergenic if present within a range of 20 kb from the 5′ and 3′ ends of the gene (da Costa Barros et al., 2018). ARS-UCD1.2/bosTau9 cow assembly was used as the reference genome to identify regions around significant SNPs in the UCSC genome browser (Zimin et al., 2009). Gene ontology (GO) was carried out using gProfiler,⁴ and the GO classifications significant over the Benjamini–Hochberg FDR were selected for pathway enrichment through Cytoscape v3.8.2 (Shannon et al., 2003).

RESULTS

PCA on Test-Day Records

PCA on test-day records proved that the first two PCs cumulatively explain 77.72, 40.77, and 48.06% of the total variation in TDMYs, TDFPs, and TDSNFs, respectively. Eigen values of PC₁ for TDMYs, TDFPs, and TDSNFs were found to be 4.45, 1.67, and 2.30, respectively, while for PC₂ they were 0.98, 1.20, and 1.05, respectively. New synthetic variable PC₁ and PC₂ for the above three traits were constructed for each buffalo utilizing original variables and variable loadings of PCA.

Genotyping

An average of 1.3 million each of forward and reverse reads were obtained per sample. The total number of forward and reverse end reads was 252.56 million with the average read length being 151 base pairs (bp). After quality check of raw data, it was observed that the average GC content was 50.54%. The rate of duplication was quite high, i.e., 80.11%, which was later checked during variant calling. The quality score (Q) throughout the dataset varied from 35 to 40.

Variant Calling and Quality Control

Raw files that were quality checked in FastQC were combined using MultiQC, and the report is provided as **Supplementary Document**. After alignment, BAM files were obtained, and quality was checked. Of the raw reads, 98.11% were mapped accurately to the Mediterranean buffalo reference genome. After removing the duplicate reads, the rate of duplication was reduced to 15.85%. Average mapping quality after alignment was found to be 29.05. A single VCF file was obtained, with 3,854,990 SNPs, out of which 3,792,469 SNPs which were strictly biallelic were retained for further analysis. After applying quality control constraints, 38,560 SNPs present on autosomes and X-chromosomes were retained for further downstream

⁴<https://biit.cs.ut.ee/gprofiler/gost>

TABLE 1 | Details of the top 10 SNPs identified through GWAS on various traits and genes identified through genome scan.

Trait	Chr [†] no.	Position (bp)	p-value	Within	±20 kb
PC ₁ of milk yield	X	7,588,358*	1.91×10^{-07}	<i>GRIA3</i>	
	9	62,699,319*	2.39×10^{-06}	<i>ZNF292</i>	
	16	74,362,861*	3.56×10^{-06}		
	1	182,059,836	1.05×10^{-05}	Uncharacterized LOC112444602	
	6	35,648,660	1.31×10^{-05}		<i>TIGD2</i>
	9	48,559,942	1.38×10^{-05}	<i>GRIK2</i>	
	1	97,587,987	2.68×10^{-05}	<i>LRRC34</i>	<i>LRRC34, ACTRT3</i>
	20	45,808,430	2.69×10^{-05}		
	2	72,321,021	3.65×10^{-05}		
	X	120,205,266	4.13×10^{-05}		
PC ₁ of fat percentage	6	34,234,112	1.96×10^{-05}	<i>CCSER1</i>	
	20	54,742,629	3.10×10^{-05}		
	17	15,895,247	5.70×10^{-05}		
	4	1,579,792	9.68×10^{-05}		
	16	66,495,537	0.0001022		
	14	27,391,536	0.0001391		
	10	45,942,651	0.0001592		<i>DAPK2</i>
	16	46,080,207	0.0001604	<i>CAMTA1</i>	
	11	30,281,532	0.0001669		
	4	39,053,011	0.0001739		
PC ₁ of SNF percentage	14	4,664,0635	6.39×10^{-06}		
	9	104,155,086	6.57×10^{-06}		
	10	4,420,352	1.03×10^{-05}		<i>TICAM2</i>
	11	88,922,443	1.34×10^{-05}		
	12	90,113,650	1.64×10^{-05}		
	3	146,125,164	1.69×10^{-05}		
	8	99,790,321	2.11×10^{-05}		<i>TXN</i>
	14	38,810,289	2.32×10^{-05}	<i>HNF4G</i>	
	14	54,721,238	2.68×10^{-05}	<i>SYBU</i>	
	17	42,173,707	3.79×10^{-05}	Uncharacterized LOC104974614	
PC ₂ of milk yield	1	81,353,445	4.36×10^{-06}		
	7	57,678,016	2.86×10^{-05}	<i>TCERG1</i>	
	4	134,421,018	4.95×10^{-05}		
	16	17,134,516	5.34×10^{-05}		
	1	49,359,549	7.78×10^{-05}		
	4	26,233,428	0.0001182		
	13	60,176,598	0.0001579		<i>ANGPT4</i>
	18	63,008,459	0.0001706		
	2	85,823,516	0.0001802	<i>ANKRD44</i> , uncharacterized LOC112442949	
	21	42,799,270	0.0002008		<i>AKAP6</i>
PC ₂ of fat percentage	3	81,728,404	9.30×10^{-05}	<i>ROR1</i>	
	15	58,026,658	9.67×10^{-05}		<i>CCDC34</i>
	14	5,954,275	0.0001596		
	6	49,443,269	0.0004445		
	23	25,106,184	0.0004533		<i>GSTA1, GSTA2</i>
	15	31,296,423	0.0004619	<i>GRIK4</i>	
	18	61,661,554	0.0004811	<i>CACNG6</i>	
	4	134,458,024	0.0004867		
	7	42,605,598	0.0005285	<i>SH3BP5L, ZNF672</i>	
	16	74,400,826	0.0005499		
PC ₂ of SNF percentage	14	71,295,596	3.83×10^{-05}		<i>TRIQQ</i>

(Continued)

TABLE 1 | (Continued)

Trait	Chr [†] no.	Position (bp)	p-value	Within	±20 kb
305 days' milk yield	12	100,102,349	4.23×10^{-05}		
	1	58,132,034	0.0002283	<i>CFAP44</i>	
	20	53,548,085	0.0003363	<i>CDH18</i>	
	4	15,355,186	0.0003396		
	6	27,219,584	0.0003841		Uncharacterized <i>LOC782977</i>
	22	19,933,711	0.000389		
	5	21,940,540	0.0003954		
	21	56,124,462	0.0004159		
	9	42,153,396	0.0004165	<i>SEC63</i>	
	10	5,804,150	2.93×10^{-05}		
	2	85,890,123	3.90×10^{-05}		<i>ANKRD44</i>
	9	62,699,319	4.52×10^{-05}	<i>ZNF292</i>	
	12	10,144,550	5.91×10^{-05}		
	X	7,588,358	6.08×10^{-05}	<i>GRIA3</i>	
	16	74,362,861	6.59×10^{-05}		
	10	5,804,363	8.51×10^{-05}		
	19	65,955,791	0.0001318		
Persistency based on Wood's function as described by Macciotta	1	97,587,987	0.0001527	<i>MYNN</i>	<i>LRRC34, ACTRT3</i>
	9	48,559,942	0.000153	<i>GRIK2</i>	
	14	2,765,427	0.000693	<i>DENND3</i>	
	9	72,904,775	0.000787		
	3	25,896,446	0.001048		
	10	98,483,390	0.001208		
	21	32,145,325	0.001243	<i>PSTPIP1</i>	
	5	118,217,420	0.001301		
	19	66,731,726	0.001385		
	6	111,661,570	0.001661		
	1	101,471,188	0.00172		
	5	36,948,158	0.001727	<i>ADAMTS20</i>	
	12	61,269,055	0.0003472		
	18	46,542,221	0.0005129	<i>PRODH2, NPHS1, RREL2, 42466</i>	
	10	66,128,637	0.000711		
	5	126,657,828	0.0008095		
	14	46,873,549	0.0008564		
Persistency (Mahadevan, 1951)	22	33,269,149	0.001113	<i>FAM19A1</i>	
	12	72,874,869	0.001206		
	23	19,210,873	0.001381	<i>CLIC5</i>	
	12	32,859,855	0.001488	<i>GPR12</i>	
	13	60,835,727	0.00154	<i>DEFB125</i>	
Persistency (Johansson and Hansson, 1940)	14	45,071,226	1.69×10^{-05}		
	21	37,776,399	2.76×10^{-05}		
	9	62,472,303	3.36×10^{-05}	<i>CFAP206</i>	
	18	46,100,792	0.0001026		
	18	46,100,619	0.0001026		
	18	17,868,492	0.0001162	<i>C18H16orf78</i>	
	2	20,133,507	0.0001273		
	3	55,760,382	0.0001561		
	4	49,950,395	0.0001709		

(Continued)

TABLE 1 | (Continued)

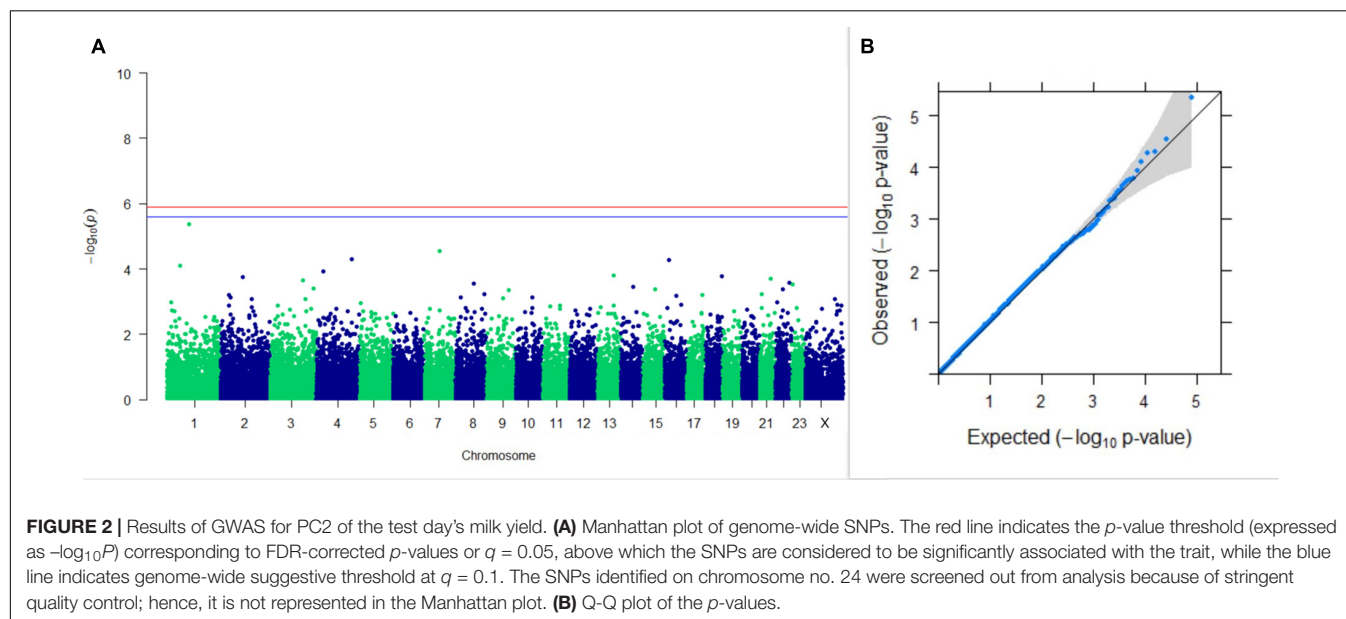
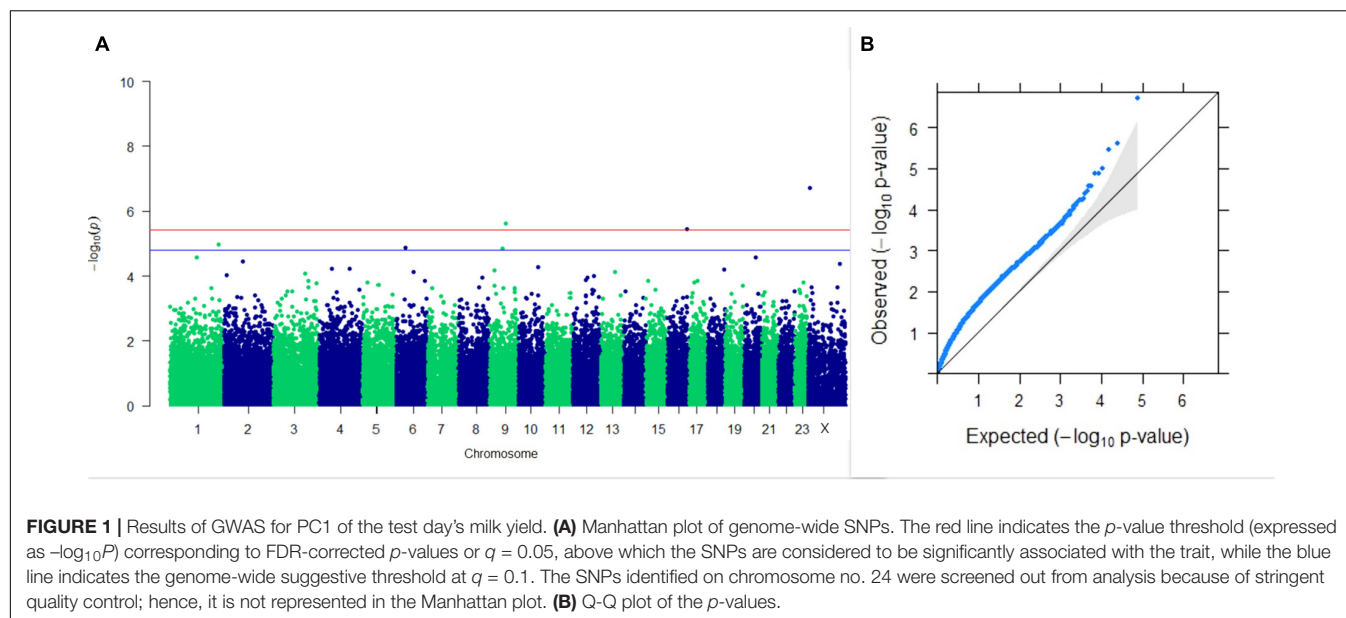
Trait	Chr [†] no.	Position (bp)	p-value	Within	±20 kb
Persistence (Ludwick and Petersen, 1943)	15	11,079,156	0.000183		
	6	2,687,0445	6.07×10^{-05}	STPG2	
	9	62,472,303	8.75×10^{-05}	CFAP206	
	18	46,100,792	0.0001368		
	18	46,100,619	0.0001368		
	4	46,668,425	0.0001908	RINT1, EFCAB10	
	10	65,940,920	0.0002065		
	7	57,678,016	0.0002898	TCERG1	
	25	40,090,282	0.0003082		
	10	35,000,168	0.0003138		
Breeding efficiency (Tomar, 1965)	18	17,868,492	0.0003109	C18H16orf78	
	10	1,310,267	1.70×10^{-05}	APC, uncharacterized LOC112448352	
	11	83,985,064	6.50×10^{-05}		
	12	54,602,811	0.0003442	NDFIP2	
	2	157,056,510	0.0003871		
	X	68,984,061	0.0004459		
	1	40,267,041	0.000448		
	7	117,169,472	0.0005262		
	20	5,188,573	0.0005834	Uncharacterized LOC107131404	
	1	182,894,045	0.0006526		
Age at sexual maturity (in months)	14	70,378,719	0.0009191		PDP1
	2	19,116,988	2.31×10^{-05}	PDE11A	
	7	77,438,396	5.80×10^{-05}		
	18	19,894,177	0.0001711		
	3	9,170,745	0.0001738	SLAMF6	
	17	58,295,542	0.0001793		uncharacterized LOC104974658
	6	32,022,004	0.0001861	GRID2	
	1	2,246,590	0.0003256	uncharacterized LOC104970778	
	15	75,434,743	0.0003302		
	1	96,574,489	0.0003995	EIF5A2, RPL22L1	
Postpartum breeding interval (in days)	10	69,569,504	0.0004181		
	2	25,677,863	4.28×10^{-05}	ERICH2	
	4	31,943,618	5.96×10^{-05}	IGF2BP3	MALSU1
	14	31,151,162	0.0001547	PPP1R42	TCF24
	2	69,670,013	0.0002049		CCDC93
	6	31,285,865	0.0002096	GRID2	
	11	39,293,274	0.0002278		
	7	84,151,709	0.0002888	EDIL3	
	7	84,151,714	0.0002947	EDIL3	
	13	11,218,416	0.0003107	LOC112449367	
	7	17,313,196	0.0003504		Uncharacterized LOC101904981, LOC112447353

[†]Chr = chromosome, *SNPs at the positions were significantly associated with the trait at the 5% level of significance using Benjamini-Hochberg's FDR.

analysis with the final genotyping rate of 98.24%. All the SNPs present on autosome no. 24 failed to pass quality control constraints; hence, in the final analysis, the 24th autosome has been removed.

GWAS Results

GWAS were performed with 38,560 SNPs, and the genome-wide significant threshold was set at the 5% FDR level. GWAS were performed on the two foremost



variation explaining PCs (PC_1 and PC_2) of TDMY, TDFP, and TDSNF.

GWAS on Lactation Traits

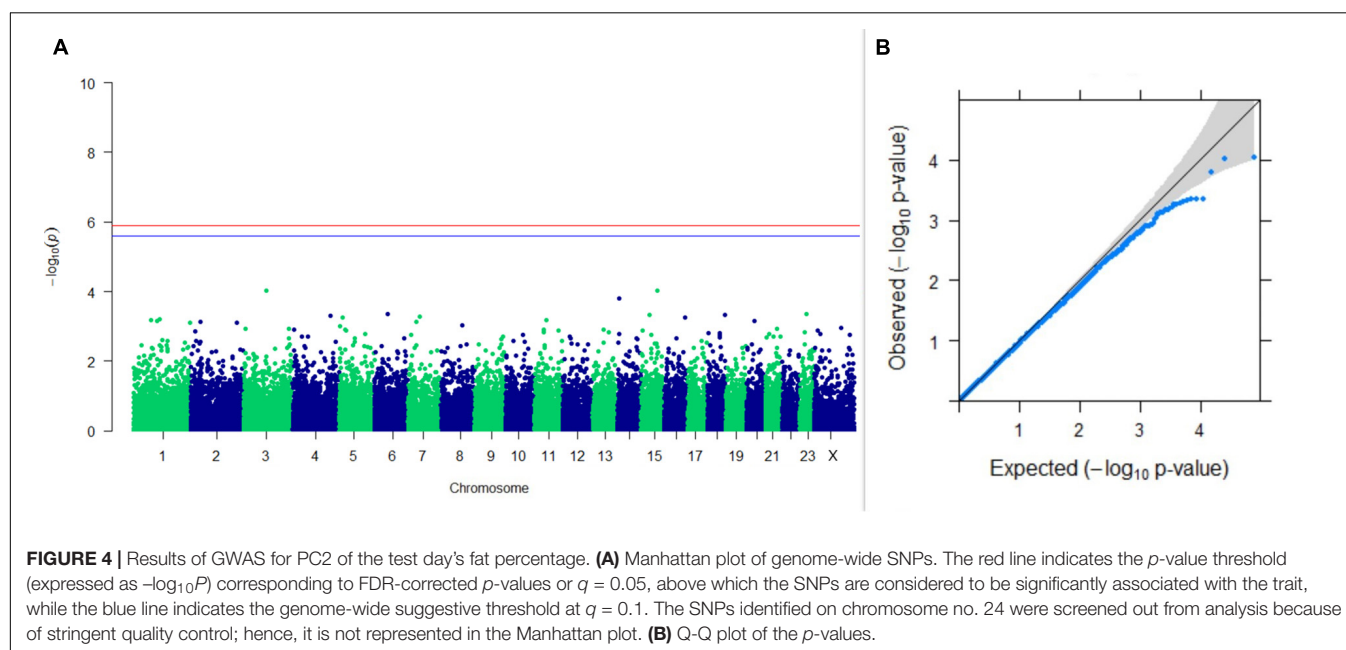
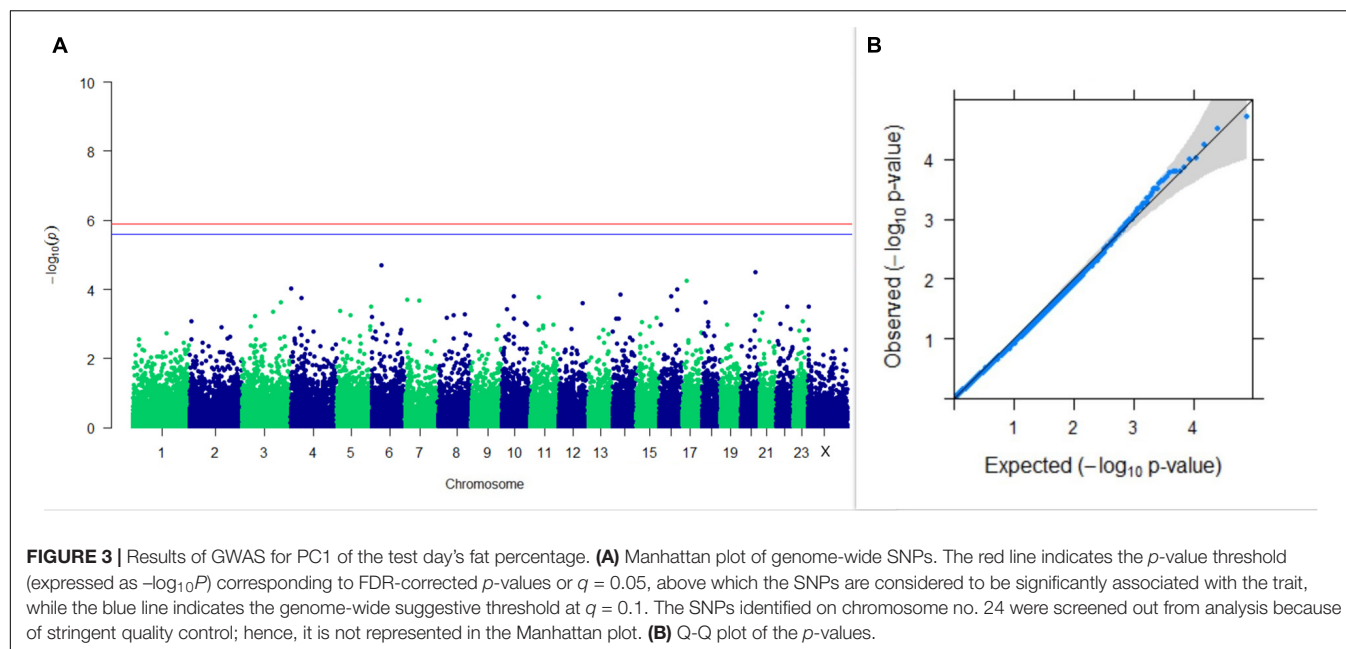
Three SNPs present on chromosomes X (7588358 bp), 9 (62699319 bp), and 16 (74362861 bp) were significantly associated with PC_1 of TDMYs with FDR-corrected p -values of 0.007369, 0.04579, and 0.04579, respectively. No SNPs were found to be significantly associated with PC_2 TDMYs. GWAS on PC_1 and PC_2 of TDFPs and TDSNFs also could not establish any significant association between the traits and SNPs. However, the top 10 SNPs having the lowest p -values in the test of association with different lactation traits, along with their position and genes within a ± 20 -kb region, are listed in **Table 1**. The

Manhattan and Q-Q plots (panels A and B, respectively) for association results of PC_1 -TDMYs, PC_2 -TDMYs, PC_1 -TDFPs, PC_2 -TDFPs, PC_1 -TDSNFs, and PC_2 -TDSNFs are given in **Figures 1–6**, respectively.

GWAS on 305DMY revealed that there was no significant association between SNPs and the trait. It was observed that five SNPs that appeared in the list with GWAS for PC_1 and PC_2 of the test day's milk yield were also in the top SNP list for 305DMY. The Manhattan and Q-Q plots for association results of 305DMY are given in **Figures 7A,B**, respectively.

GWAS on Lactation Persistency

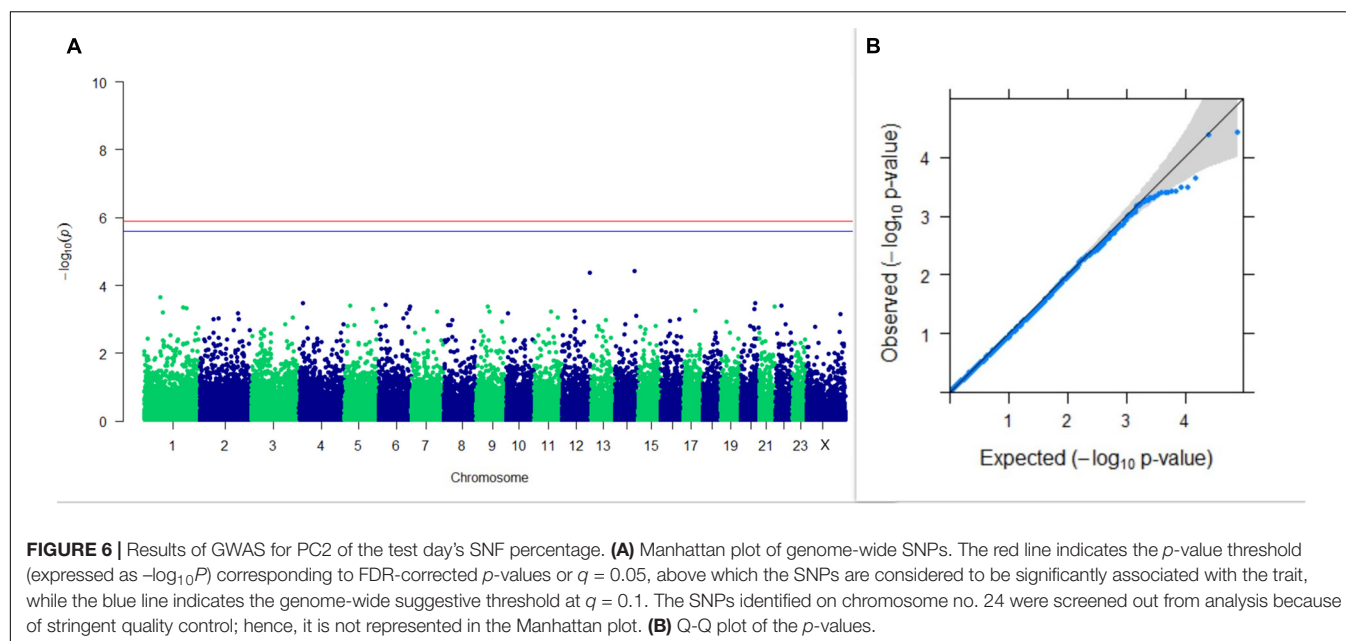
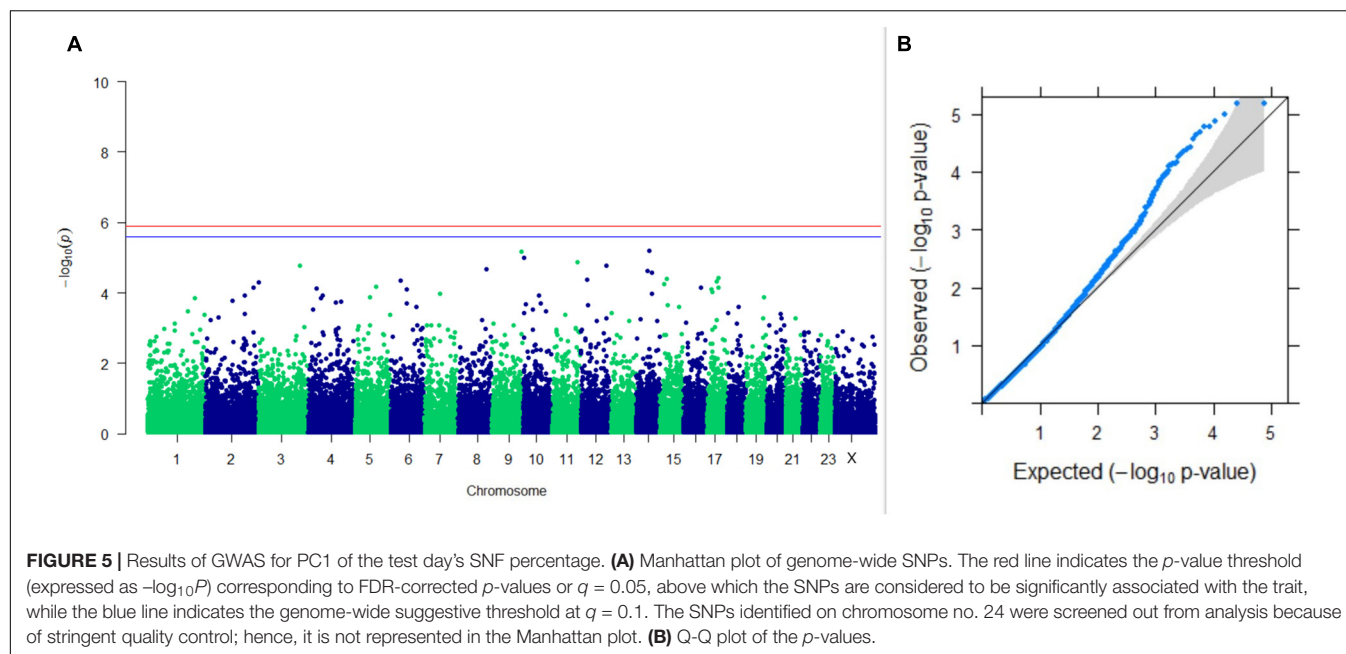
GWAS on lactation persistency estimated by four different methods revealed no significant association between SNPs and



the trait generated. The top 10 SNPs having the lowest p -values along with their genomic position and genes within a ± 20 -kb region are listed in **Table 1** for all four methods. It was observed that four SNPs were found in common between the top SNPs listed for persistency estimated by methods II and III. Among the four SNPs, one was on chromosome 9 at 62472303 bp, and three were present on chromosome 18 at 46100792, 46100619, and 17868492 bp. The Manhattan and Q-Q plots (panels A and B, respectively) for association results of persistency estimated by four different methods (I, II, III, and IV) are given in **Figures 8–11**, respectively.

GWAS on Fertility Traits

Upon performing GWAS on age at sexual maturity, postpartum breeding interval, and breeding efficiency, no significant association of any SNPs were observed for the traits; however, the top 10 SNPs with the lowest p -values and genes within a ± 20 -kb region are listed in **Table 1**. The Manhattan and Q-Q plots (panels A and B, respectively) of GWAS on age at sexual maturity, postpartum breeding interval, and breeding efficiency are given in **Figures 12–14**, respectively.

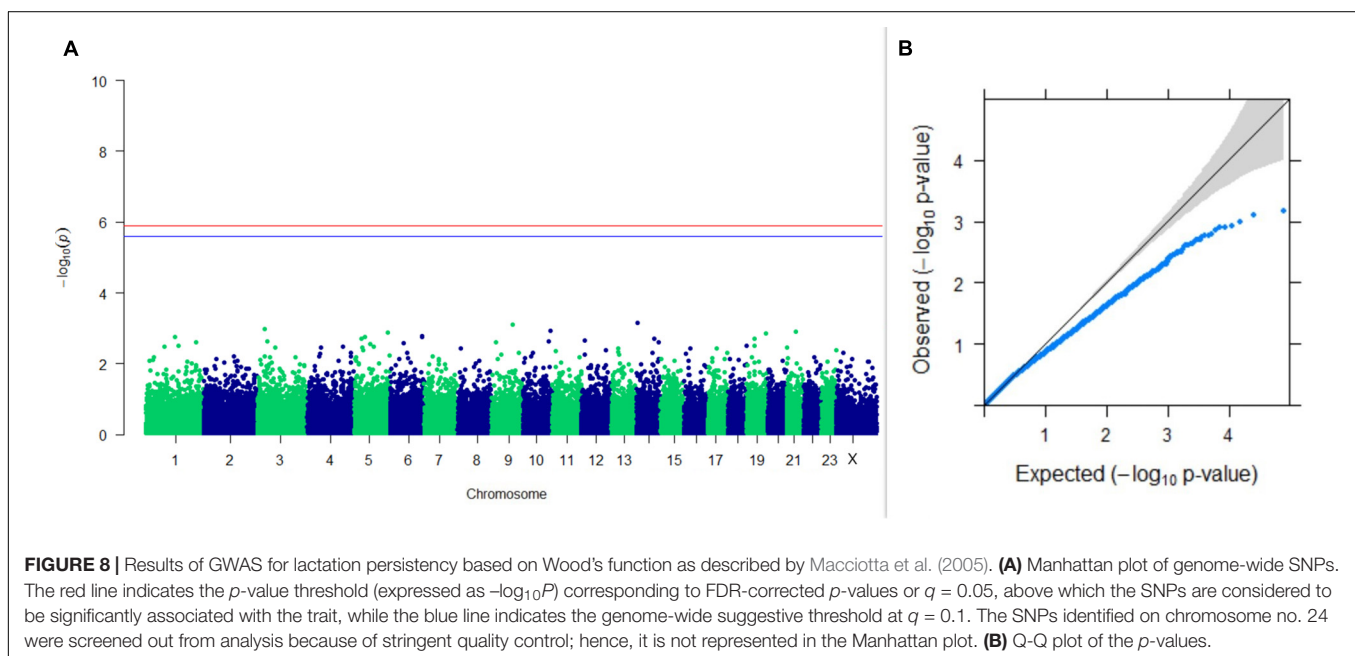
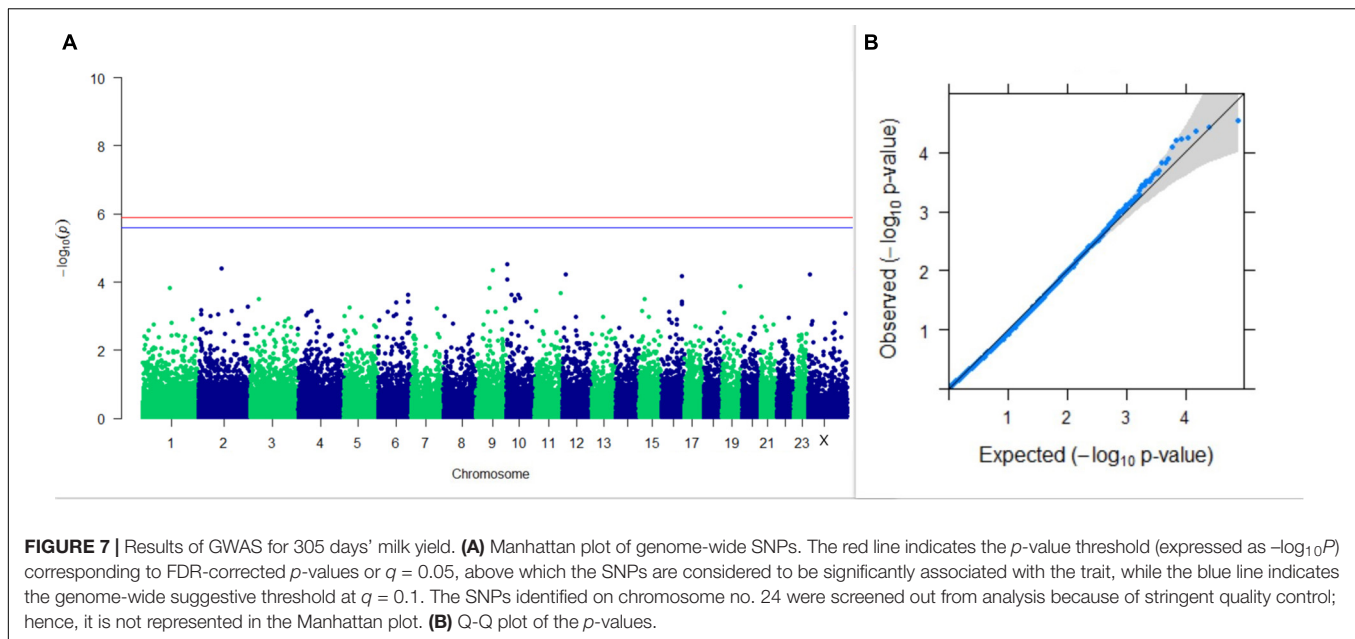


DISCUSSION

The test-day model (TDM) as repeated measurements of milk yield traits is the method of choice for predicting 305 days' milk yield. TDMs account for environmental effects on each test day and is useful to model individual lactation curves (Schaeffer et al., 2000; Bignardi et al., 2012). PCA being a powerful multivariate technique is often used to predict 305 days' milk yield from the inter-related variables such as test-day records (Taggar et al., 2012). Wara et al. (2019) had performed GWAS on test day's milk yield in Vrindavani cattle; however, reports of GWAS on PCs of test days are very few. In the present study, PCA was applied

on seven test-day records of first lactation in order to perform GWAS on the PCs explaining maximum variation rather than the test-day records themselves. The first two PCs cumulatively explained 40–78% of total variation in test-day records of traits included, indicating their potential to be used as GWAS traits to identify novel SNPs for milk yield, fat percentages, and SNF percentages.

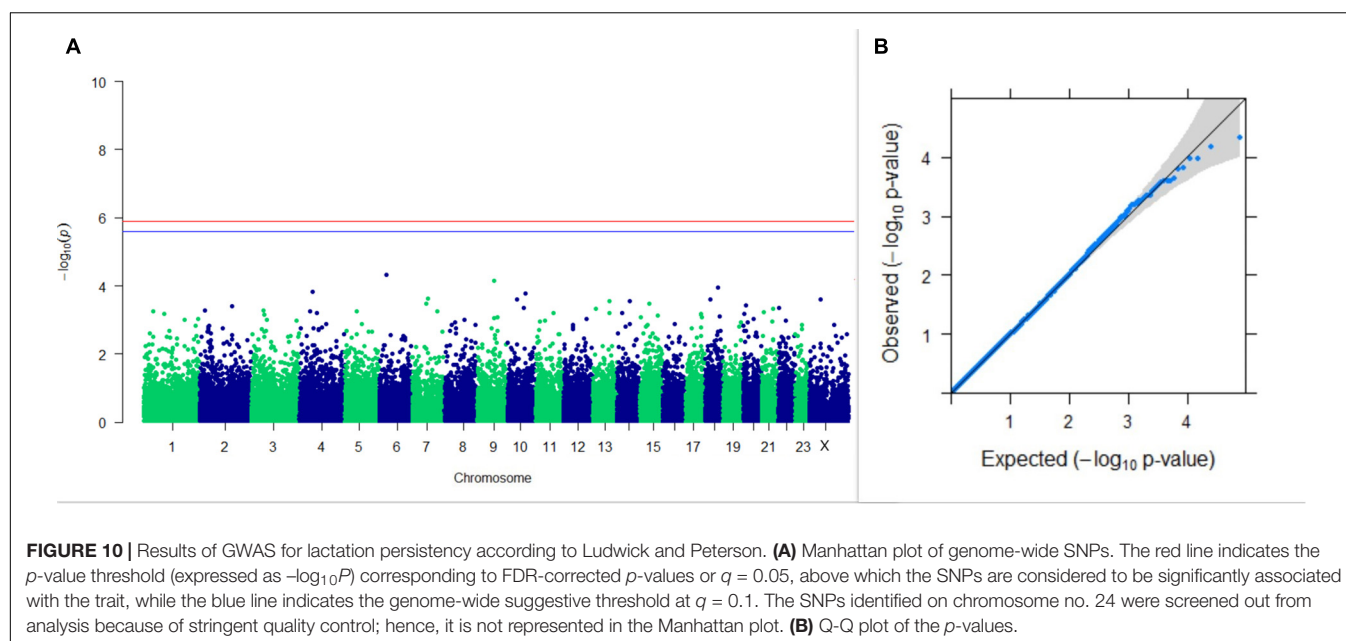
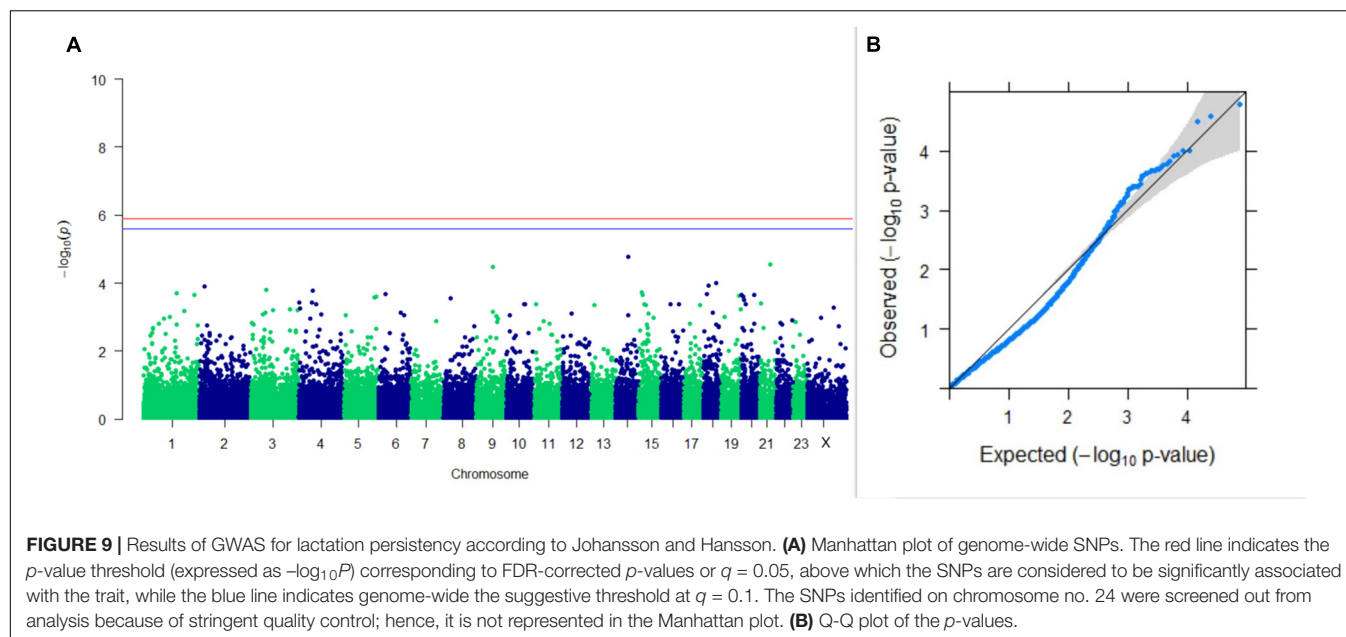
It is important to understand the reasons for considering 96 samples taken in the study as optimum, as in the buffalo farming scenario in Asia, particularly in India, the maximum number of buffaloes maintained at any large organized herd ranges from 250 to 500, with 100–200 breedable buffaloes having



complete phenotype information. The National Dairy Research Institute has the second-largest herd in the Indian Council of Agricultural Research (ICAR) with a well-managed herd of 250 breedable buffaloes. Furthermore, there is no buffalo sequencing consortium/project in operation (India). In such a case, one could afford this sample size to genotype with complete pedigree and with sufficient genetic diversity. However, the sample size is less for conducting GWAS and raises the question of whether the results are reliable enough. It was observed that the results obtained under the present study are encouraging and important, as one of the genes identified in the present study, i.e., *GRIA3*, was reported by Cole et al. (2011) in a sample of 1,654 US Holstein

cows. There are occasions where a similar sample size (96) has been used to understand body morphology traits through GWAS (Rahmatalla et al., 2018).

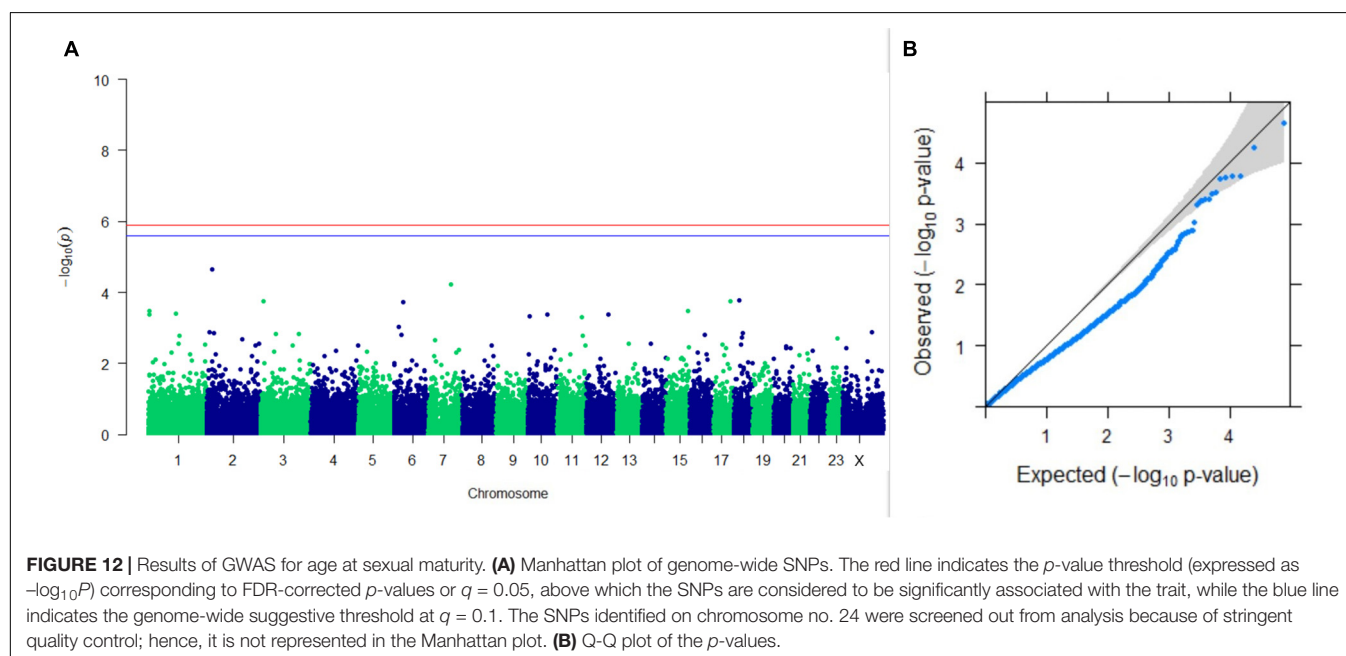
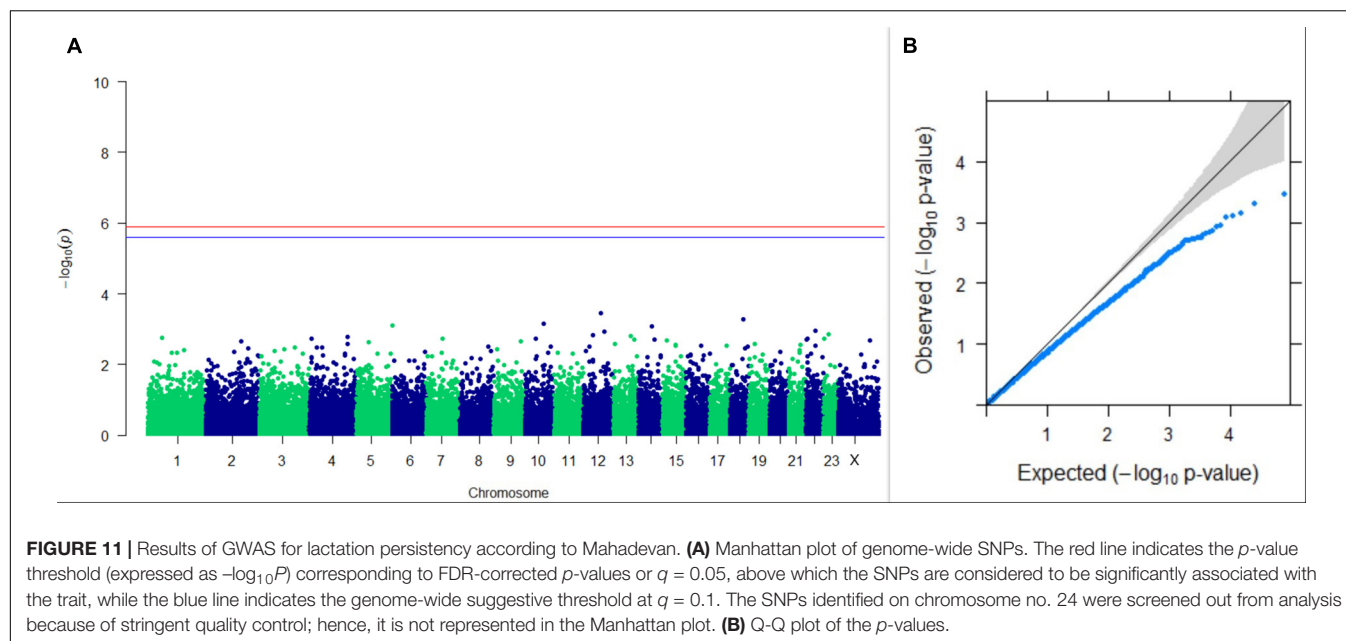
To identify novel SNPs associated with various economic traits, sequencing through the ddRAD approach was performed. Ruffalo et al. (2012) highlighted that a mapping quality of 15–40 is of the highest nature, and theoretical accuracy corresponds to ~100%, while depending on the aligner, the actual accuracy of mapping or base call may vary from 40 to 60%. Similar mapping quality averaging 30 for all samples was obtained through BWA aligner, which usually implies good overall base quality of reads and few mismatches in the best possible alignment.



An earlier bovine SNP chip was used to study the traits of buffalo. Using a bovine SNP chip (Illumina BovineSNP50 BeadChip) for a GWAS in buffalo population, Wu et al. (2013) identified seven SNPs that were significantly associated with milk yield. Later on, Affymetrix's 90K SNP chip commercialization led to several GWAS on milk production traits in buffalo. de Camargo et al. (2015) in a GWAS on buffaloes reported that *LOC100847171*, *BCL6*, *RTP2*, *SST*, *PTGS2*, *LOC100295047*, and *LOC101908004* are associated with milk production; *KCTD8*, *LOC782855*, *LOC101904777*, *ESRRG*, *TRNAY-AUA*, and *GPATCH2* are associated with fat percentage; *LOC101903483*, *SART3*, *ISCU*, *CMKLR1*, *WSCD2*, *MFNG*, *CARD10*, and *USP18*

are associated with protein percentage. Liu et al. (2018) identified several candidate genes, namely, *MFSD14A*, *SLC35A3*, *PALMD*, *RGS22*, and *VPS13B*, for milk production in the Italian Mediterranean buffalo through GWAS. In another GWAS in buffalo, Iamartino et al. (2017) reported that genes regulating the D-glucose level in the blood affect milk production in buffalo.

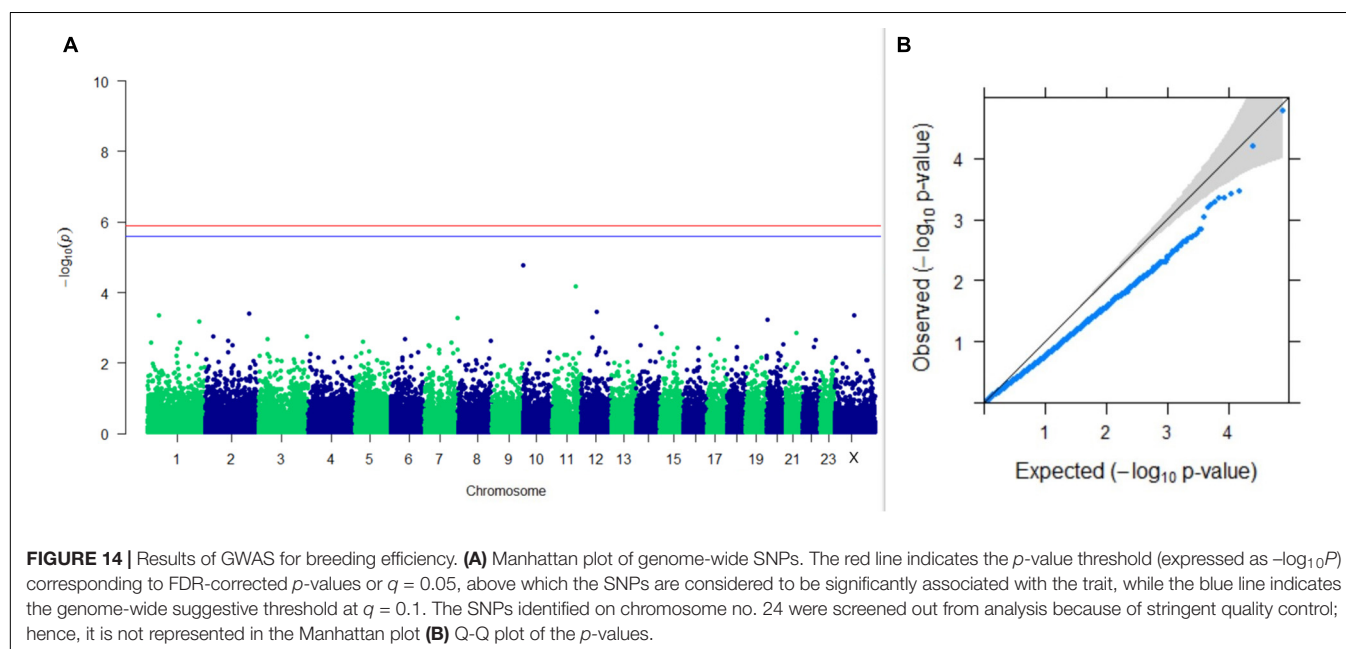
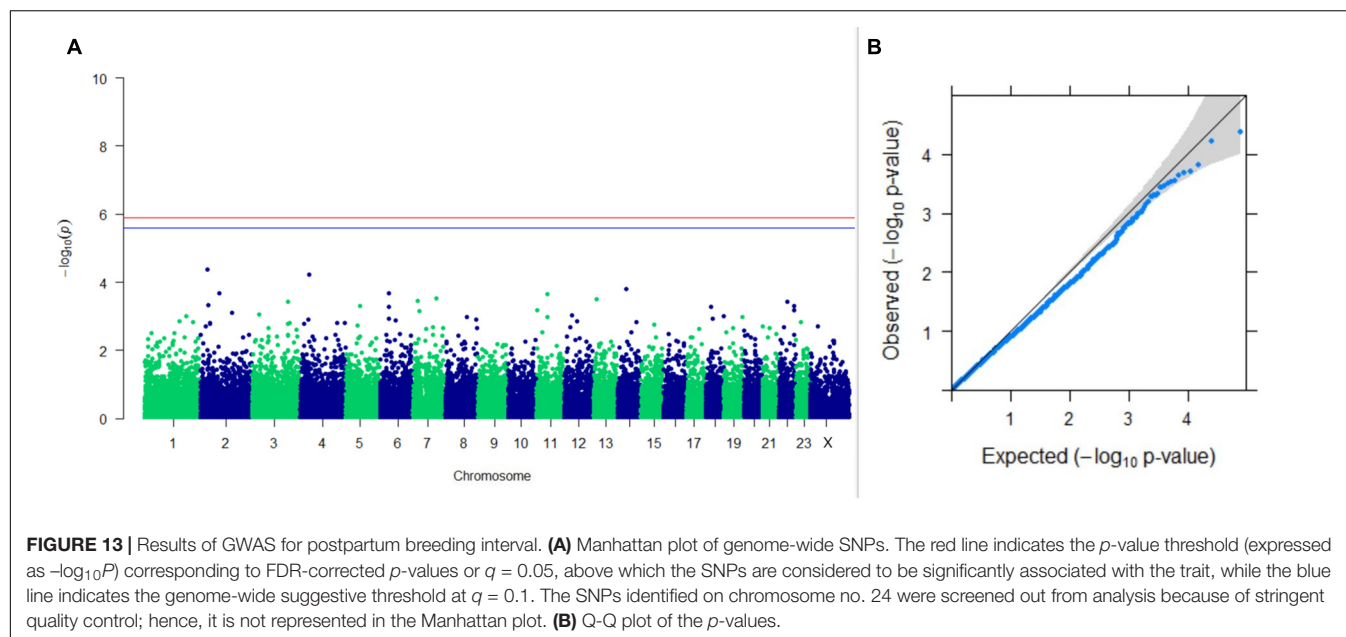
In the present study, GWAS performed on the milk yield PCs explaining maximum variation, which revealed that three SNPs were significantly associated with PC₁ of TDMY at the 5% FDR level. However, six SNPs present on chromosomes 1, 6, and 9 at 182,059,836, 35,648,660, and 48,559,942 bp, respectively, were above the suggestive



threshold of 10% FDR and are presented in **Table 1** as the top six SNPs associated with PC₁ of TDMY. However, for the rest of the traits studied, no SNPs were detected to be significantly associated (or as rejections) at 5% or even 10% FDR.

Upon scanning ± 20 kb around that SNP on the X chromosome, the *GRIA3* gene was found. Glutamate ionotropic receptor AMPA type subunit 3 (*GRIA3*) is reported to be very significantly associated with daughter pregnancy rate in US Holstein cows (Cole et al., 2011). In a study, a positive selection signal has been observed for the loci containing the *CNIH3* gene. *CNIH3* interacts with *GRIA1*, *GRIA2*, *GRIA3*, *GRIA4*, and

GRIK1 belonging to a class of α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) glutamate receptors, thus regulating trafficking of AMPA receptors. AMPA receptors are known to participate in luteinizing hormone (LH) secretion (Utsunomiya et al., 2013). The nearest genes to the top 10 SNPs from PC₁ GWAS for milk yield were *ZNF292*, *LOC112444602*, *TIGD2*, *GRIK2*, and *LRRC34*. Similarly, for PC₂ of milk yield, GWAS revealed that *TCERG1*, *ANGPT4*, *ANKRD44*, *LOC112442949*, and *AKAP6* were the nearest genes from the top 10 SNPs. It could be observed that five SNPs present in the list of top SNPs from GWAS results of PC₁ and PC₂ of the test day's milk yield were also present in the top 10 SNP list



for 305 days' milk yield (**Table 1**), highlighting the efficacy of PCA in predicting 305 days' milk yield. Additionally, *MYNN* and *ACTRT3* were found nearest to the top SNPs of GWAS for 305 days' milk yield. Cai et al. (2020) have also reported the *ANKRD44* gene to be associated with milk yield in Nordic Holstein cattle.

A genome-wide scan for novel genes for fat percentage outlined *CCSER1*, *DAPK2*, *CAMTA1*, *ROR1*, *CCDC34*, *GSTA1*, *GSTA2*, *GRIK4*, *CACNG6*, *SH3BP5L*, and *ZNF672* as the nearest genes to the top SNPs obtained from GWAS. Kolbehdari et al. (2009) in a whole-genome study in Canadian Holstein cattle mapped the gene *DAPK2* to several milk production-related

QTLs. They also mapped *DAPK2* to a milk fat percentage QTL which concurs with the findings of our study. Genes other than *DAPK2* were not reported earlier to be associated with milk fat percentage in buffalo; however, *CAMTA1* may possibly have a role in fatty acid metabolism.

TICAM2, *TXN*, *HNF4G*, *SYBU*, *LOC104974614*, *TRIQQ*, *CFAP44*, *CDH18*, *LOC782977*, and *SEC63* were found near the top SNPs associated with PCs of TDSNF. We could not find any previous reports stating the role of these genes in regulating milk SNF percentage; however, a functional enrichment analysis study by Zhou et al. (2019) revealed *HNF4G* and *SYBU* as candidate genes that regulate milk fat synthesis, transport, and

metabolism. All the genes involved with milk yield and its composition identified through the study were enriched for pathways in Cytoscape. A sub-network of the genes is depicted in **Supplementary Figure 2**. It can be observed that most of the genes identified belong to the glutamate receptor family, involved in the regulation of the neuronal system.

In Canadian Holstein cattle, Do et al. (2017) reported a significant genetic effect of *MAN1C1*, *MAP3K5*, *HCN1*, *TSPAN9*, *MRPS30*, *TEX14*, and *CCL28* genes on lactation persistency. Deng et al. (2019) reported that the *MAP3K5* gene regulates p38 MAPKs and Jun N-terminal kinases (JNKs) pathways involved in mammary gland development. In the present study, lactation persistency was calculated using four different methods, and it was observed that the top four SNPs identified by GWAS for persistency as estimated by the Johansson and Hansson (1940) method were in common with the GWAS results obtained for persistency as per the Ludwick and Petersen (1943) method. The genes present near the top SNPs were *DENND3*, *PSTPIP1*, *ADAMTS20*, *PRODH2*, *NPHS1*, *RREL2*, *FAM19A1*, *CLIC5*, *GPR12*, *DEFB125*, *STPG2*, *CFAP206*, *RINT1*, *EFCAB10*, *TCERG1*, and *C18H16orf78*. Upon network enrichment, it was observed that *TCERG1*, *RINT1*, *CLIC5*, and *PSTPIP1* are co-expressed with several other genes in the breast mammary tissue of human, indicating their possible role in persistency of lactation in dairy animals.

GWAS for fertility traits were performed on breeding efficiency, age at sexual maturity, and postpartum breeding interval. Genes present near the top SNPs of breeding efficiency were *APC*, *LOC112448352*, *NDFIP2*, *LOC107131404*, and *PDP1*. Mohamed et al. (2019) reported that *APC* gene in the canonical WNT signaling pathway plays a critical role in the regulation of ovarian development. Mis-regulation of this key pathway in the adult ovary is associated with subfertility in mice. The *APC* gene has also been mapped to the QTL region for conception rate in Holstein cattle (Kolbehdari et al., 2009). Genes identified near the top SNPs for age at sexual maturity are *PDE11A*, *SLAMF6*, *LOC104974658*, *GRID2*, *LOC104970778*, *EIF5A2*, and *RPL22L1*. Coyral-Castel et al. (2018) mapped the *SLAMF6* gene to the QTL region affecting female fertility located on the bovine chromosome three (QTL-F-Fert-BTA3). The *GRID2* gene is however reported to be associated with growth traits in Simmental cattle and may have role in regulating age at sexual maturity (Braz et al., 2020). Christensen et al. (2005) have reported *EIF5A2* as a candidate gene for infertility in human. Genes identified for postpartum breeding interval were *ERICH2*, *MALSU1*, *IGF2BP3*, *TCF24*, *PPP1R42*, *CCDC93*, *GRID2*, *EDIL3*, *LOC112449367*, *LOC101904981*, and *LOC112447353*. However, none of these genes was identified to be involved significantly in any pathway through the gProfiler.

Although we have already emphasized that this a preliminary GWAS on buffaloes covering a large set of economic traits (lactation, lactation persistency, and fertility), the results obtained in the present study are bias free, as indicated by the FDR. The findings of the present study in the Murrah population will

encourage researchers to come forward for GWAS in buffalo in a holistic manner.

CONCLUSION

Optimum production and reproduction in buffaloes are long-standing questions in the Indian subcontinent. We present the analysis and identification of genomic regions that play role in shaping the selection and breeding decisions of Murrah buffalo for persistency of production and fertility. The dataset information presented in the paper is also submitted so that it can be used to compare and evaluate other breeds of buffalo based on the genomic information generated. The putative identified regions have a potential to improve the existing breeding decisions in this important dairy germplasm.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the European Variation Archive repository, accession number PRJEB47270 (<https://www.ebi.ac.uk/eva/?eva-study=PRJEB47270>).

ETHICS STATEMENT

The animal study was reviewed and approved by the ICAR-National Dairy Research Institute IAEC.

AUTHOR CONTRIBUTIONS

VV conceptualized the work and interpreted the data. SC, GG, RA, and AM performed the collection of materials, and performed the data analysis. AV and SD provided the data and resources. VV and SC did writing of the manuscript and its critical evaluation. All authors contributed to the article and approved the submitted version.

FUNDING

The authors thank the Director of ICAR-NDRI, Karnal for providing funds and support to carry out the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.696109/full#supplementary-material>

Supplementary Figure 1 | MDS plot based on the first and second MDS components showing population stratification.

Supplementary Figure 2 | Sub-network of enriched pathways of genes identified responsible for milk yield and its composition.

REFERENCES

- Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (accessed August 21, 2020).
- Barbosa, A. M. (2020). FuzzySim: applying fuzzy logic to binary similarity indices in ecology. *Methods Ecol. Evol.* 6, 853–858. doi: 10.1111/2041-210X.12372
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Bignardi, A. B., El Faro, L., Rosa, G. J. M., Cardoso, V. L., Machado, P. F., and Albuquerque, L. G. D. (2012). Principal components and factor analytic models for test-day milk yield in Brazilian Holstein cattle. *J. Dairy Sci.* 95, 2157–2164. doi: 10.3168/jds.2011-4494
- Braz, C. U., Rowan, T. N., Schnabel, R. D., and Decker, J. E. (2020). Extensive genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. *Biorxiv*. [Preprint]. doi: 10.1101/2020.01.09.900902
- Brian, B. (2014). “BBMap: a fast, accurate, splice-aware aligner,” in *Proceedings of the 9th Annual Genomics of Energy & Environment Meeting*, (Berkeley, CA: Lawrence Berkeley National Lab).
- Bush, W. S., and Moore, J. H. (2012). Genome-wide association studies. *PLoS Comput. Biol.* 8:e1002822. doi: 10.1371/journal.pcbi.1002822
- Cai, Z., Dusza, M., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2020). Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genet. Sel. Evol.* 52:19. doi: 10.1186/s12711-020-00538-6
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8
- Christensen, G. L., Ivanov, I. P., Atkins, J. F., Mielnik, A., Schlegel, P. N., and Carrell, D. T. (2005). Screening the SPO11 and EIF5A2 genes in a population of infertile men. *Fertil. Steril.* 84, 758–760. doi: 10.1016/j.fertnstert.2005.03.053
- Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., et al. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC Genomics* 12:408. doi: 10.1186/1471-2164-12-408
- Coyral-Castel, S., Ramé, C., Cogne, J., Lecardonnell, J., Marthey, S., Esquerré, D., et al. (2018). KIRREL is differentially expressed in adipose tissue from ‘fertil’ and ‘fertil’-cows: in vitro role in ovary? *Reproduction* 155, 181–196. doi: 10.1530/REP-17-0649
- da Costa Barros, C., de Abreu Santos, D. J., Aspilcueta-Borquis, R. R., De Camargo, G. M. F., de Araújo Neto, F. R., and Tonhati, H. (2018). Use of single-step genome-wide association studies for prospecting genomic regions related to milk production and milk quality of buffalo. *J. Dairy Res.* 85, 402–406. doi: 10.1017/S0022029918000766
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012
- de Camargo, G. M. F., Aspilcueta-Borquis, R. R., Fortes, M. R. S., Porto-Neto, R., Cardoso, D. F., Santos, D. J. A., et al. (2015). Prospecting major genes in dairy buffaloes. *BMC Genomics* 16:872. doi: 10.1186/s12864-015-1986-2
- Deng, T., Liang, A., Liang, S., Ma, X., Lu, X., Duan, A., et al. (2019). Integrative analysis of transcriptome and GWAS data to identify the hub genes associated with milk yield trait in buffalo. *Front. Genet.* 10:36. doi: 10.3389/fgene.2019.00036
- Do, D. N., Bissonnette, N., Lacasse, P., Miglior, F., Sargolzaei, M., Zhao, X., et al. (2017). Genome-wide association analysis and pathways enrichment for lactation persistency in Canadian Holstein cattle. *J. Dairy Sci.* 100, 1955–1970. doi: 10.3168/jds.2016-11910
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379
- Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354
- FAO (2014). *Food and Agriculture Organization the United Nations Statistics Division*. Available online at: <http://faostat3.fao.org/compare/E> (accessed November 12, 2020).
- Ganguli, N. C. (1981). Buffalo as a candidate for milk production. *Int. Dairy Fed. Bull.* 137, 48–56.
- García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28, 2678–2679. doi: 10.1093/bioinformatics/bts503
- Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* 67, 850–857. doi: 10.1016/j.jclinepi.2014.03.012
- Gordon, I. (1996). *Controlled Reproduction in Cattle and Buffaloes*, Vol. 1. Wallingford: Centre for Agriculture and Bioscience International.
- Griffiths, M. W. (2010). *Improving the Safety and Quality of Milk: Improving Buffalo Milk*. Amsterdam: Elsevier. doi: 10.1533/9781845699437
- Iamartino, D., Nicolazzi, E. L., Van Tassel, C. P., Reecy, J. M., Fritz-Waters, E. R., Koltes, J. E., et al. (2017). Design and validation of a 90K SNP genotyping assay for the water buffalo (*Bubalus bubalis*). *PLoS One* 12:e0185220. doi: 10.1371/journal.pone.0185220
- Johansson, I., and Hansson, A. (1940). Causes of variation in milk and butterfat yield of dairy cows. *Kungliga Lantbruksakademiens Handlingar* 79:127.
- Khedkar, C., Kalyankar, S., and Deosarkar, S. (2016). “Buffalo milk,” in *Encyclopedia of Food and Health*, eds B. Caballero, M. Finglas, and F. Toldrá (New York, NY: Academic Press), 522–528. doi: 10.1016/B978-0-12-384947-2.0093-3
- Kolbehari, D., Wang, Z., Grant, J. R., Murdoch, B., Prasad, A., Xiu, Z., et al. (2009). A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. *J. Animal Breed. Genet.* 126, 216–227. doi: 10.1111/j.1439-0388.2008.00793.x
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698
- Liu, J. J., Liang, A. X., Campanile, G., Plastow, G., Zhang, C., Wang, Z., et al. (2018). Genome-wide association studies to identify quantitative trait loci affecting milk production traits in water buffalo. *J. Dairy Sci.* 101, 433–444. doi: 10.3168/jds.2017-13246
- Ludwick, T. M., and Petersen, W. E. (1943). A measure of persistency of lactation in dairy cattle. *J. Dairy Sci.* 26, 439–445. doi: 10.3168/jds.S0022-0302(43)92739-0
- Macciotta, N. P. P., Vicario, D., and Cappio-Borlino, A. (2005). Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. *J. Dairy Sci.* 88, 1178–1191. doi: 10.3168/jds.S0022-0302(05)72784-3
- Mahadevan, P. (1951). The effect of environment and heredity on lactation. II. persistency of lactation. *J. Agric. Sci.* 41, 89–93. doi: 10.1017/S0021859600058573
- Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27:e1608. doi: 10.1002/mpr.1608
- Mohamed, N. E., Hay, T., Reed, K. R., Smalley, M. J., and Clarke, A. R. (2019). APC2 is critical for ovarian WNT signalling control, fertility and tumour suppression. *BMC Cancer* 19:677. doi: 10.1186/s12885-019-5867-y
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135. doi: 10.1371/journal.pone.0037135
- Rahmatalla, S. A., Arends, D., Reissmann, M., Wimmers, K., Reyer, H., and Brockmann, G. A. (2018). Genome-wide association study of body morphological traits in Sudanese goats. *Anim. Genet.* 49, 478–482. doi: 10.1111/age.12686
- Raina, V., Narang, R., Malhotra, P., Kaur, S., Dubey, P. P., Tekam, S., et al. (2016). Breeding efficiency of crossbred cattle and Murrah buffaloes at organized dairy farm. *Indian J. Anim. Res.* 50, 867–871. doi: 10.18805/ijar.v0iOF.6663
- Ruffalo, M., Koyutürk, M., Ray, S., and LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics* 28, i349–i355. doi: 10.1093/bioinformatics/bts408

- Sambrook, J., and Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol: chloroform. *Cold Spring Harb. Protoc.* 2006:4455. doi: 10.1101/pdb.prot4455
- Schaeffer, L. R., Jamrozik, J., Van Dorp, R., Kelton, D. F., and Lazenby, D. W. (2000). Estimating daily yields of cows from different milking schemes. *Livest. Prod. Sci.* 65, 219–227. doi: 10.1016/S0301-6226(00)00153-6
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Taggar, R. K., Das, A. K., Kumar, D., and Mahajan, V. (2012). Prediction of milk yield in Jersey cows using principal component analysis. *Progr. Res.* 7, 272–274.
- Tomar, N. S. (1965). A note on the method of working out breeding efficiency in Zebu cows and buffaloes. *Indian Dairymen* 17, 389–390.
- Utsunomiya, Y. T., O'Brien, A. M. P., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., et al. (2013). Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PloS One* 8:e64280. doi: 10.1371/journal.pone.0064280
- Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5, 247–252. doi: 10.1038/nmeth.1185
- Venturini, G. C., Cardoso, D. F., Baldi, F., Freitas, A. C., Aspilcueta-Borquis, R. R., Santos, D. J. A., et al. (2014). Association between single-nucleotide polymorphisms and milk production traits in buffalo. *Genet. Mol. Res.* 13, 10256–10268. doi: 10.4238/2014.December.4.20
- Wara, A. B., Kumar, A., Singh, A., Arthikeyan, A. K., Dutt, T., and Mishra, B. P. (2019). Genome wide association study of test day's and 305 days milk yield in crossbred cattle. *Indian J. Anim. Sci.* 89, 861–865.
- Warriach, H. M., McGill, D. M., Bush, R. D., Wynn, P. C., and Chohan, K. R. (2015). A review of recent developments in buffalo reproduction—a review. *Asian-Australas. J. Anim. Sci.* 28:451. doi: 10.5713/ajas.14.0259
- Wood, P. D. P. (1967). Algebraic model of the lactation curve in cattle. *Nature* 216, 164–165. doi: 10.1038/216164a0
- Wu, J. J., Song, L. J., Wu, F. J., Liang, X. W., Yang, B. Z., Wathes, D. C., et al. (2013). Investigation of transferability of BovineSNP50 BeadChip from cattle to water buffalo for genome wide association study. *Mol. Biol. Rep.* 40, 743–750. doi: 10.1007/s11033-012-1932-1
- Zhou, C., Shen, D., Li, C., Cai, W., Liu, S., Yin, H., et al. (2019). Comparative transcriptomic and proteomic analyses identify key genes associated with milk fat traits in Chinese Holstein cows. *Front. Genet.* 10:672. doi: 10.3389/fgene.2019.00672
- Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., et al. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 10:R42. doi: 10.1186/gb-2009-10-4-r42

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Vohra, Chhotaray, Gowane, Alex, Mukherjee, Verma and Deb. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Insights Into mRNA and Long Non-coding RNA Profiling RNA Sequencing in Uterus of Chickens With Pink and Blue Eggshell Colors

Siyu Chen¹, Kecheng Chen¹, Jiaming Xu¹, Fangwei Li², Jinlong Ding², Zheng Ma¹, Gen Li¹ and Hua Li^{1*}

¹ Guangdong Provincial Key Laboratory of Animal Molecular Design and Precise Breeding, Key Laboratory of Animal Molecular Design and Precise Breeding of Guangdong Higher Education Institutes, School of Life Science and Engineering, Foshan University, Foshan, China, ² Guizhou Changshun Tiannong Green Shell Laying Hen Industrial Co. Ltd, Chang Shun City, China

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey
(ITESM), Mexico

Reviewed by:

Qingyou Liu,
Guangxi University, China
Tatsuhiko Goto,
Obihiro University of Agriculture and
Veterinary Medicine, Japan

*Correspondence:

Hua Li
okhuali@fosu.edu.cn

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Veterinary Science

Received: 05 July 2021

Accepted: 30 August 2021

Published: 07 October 2021

Citation:

Chen S, Chen K, Xu J, Li F, Ding J,
Ma Z, Li G and Li H (2021) Insights
Into mRNA and Long Non-coding
RNA Profiling RNA Sequencing in
Uterus of Chickens With Pink and Blue
Eggshell Colors.
Front. Vet. Sci. 8:736387.
doi: 10.3389/fvets.2021.736387

The blue egg is both of biological interest and economic importance for consumers, egg retailers, and scientists. To date, the genetic mechanisms underlying pigment have mainly focused on protein-coding genes. However, the underpinning mechanism of non-coding RNAs on the pigment deposition among different eggshell colors remains unknown. In this study, RNA sequencing was employed to profile the uterine gland transcriptome (lncRNA and mRNA) of 15 Changshun blue eggshell layers, to better understand the genetic mechanisms of deposition of blue eggshell color. Results showed that differentially expressed mRNAs, GO terms, and KEGG pathways among pink-eggshell and blue-eggshell chickens were mainly targeting immune- and transporter-related terms with the SLC family, *IgJ*, CD family, and *MTMR* genes. Furthermore, the progesterone-mediated oocyte maturation and cortisol synthesis and secretion pathway with targeted gene *PGR* and *Pbx1* were significantly enriched between blue- and pink-eggshell chickens. Integrating analysis of lncRNA and mRNA profiles predicted 4 and 25 lncRNA–gene pairs by antisense and cis analysis. They were relative to immune, nerve, and lipids and amino acid metabolisms, porphyrin, and chlorophyll metabolism with targeted gene *FECH* and oxidative phosphorylation and cardiac muscle contraction pathways with targeted gene *COX6A1*. Within blue-eggshell chickens, the GO terms hindbrain tangential cell migration and phosphatidylinositol monophosphate phosphatase activity with targeted gene *Plxna2* and *MTRM1* were identified. Integrating analysis of lncRNA and mRNA profiles predicted 8 and 22 lncRNA–gene pairs. Most pathways were mainly enriched on lipid-related metabolisms as found in mRNA sequencing. The lncRNAs did exert similar functions in color formation by modulating pigment disposition and immune- and lipid-related metabolisms. Our results provide a catalog of chicken uterine lncRNAs and genes worthy of further studies to understand their roles in the selection for blue eggshell color layers.

Keywords: color deposition, lncRNAs, mRNA, uterus, chicken

INTRODUCTION

The blue egg is both of biological interest and economic importance for consumers, egg retailers, and scientists. A series of studies have been conducted on the nature of the shell pigments, the biochemical and physiological processes in various avian species involved in pigment formation, and its deposition in and on the shell (1, 2). It has been long known that the primary avian eggshell pigments are protoporphyrin IX, biliverdin IX, and biliverdin IX zinc chelate in both wild birds and poultry (3, 4). The pink, light red, and brown eggshell colors are involved with the deposition of protoporphyrin IX, while blue and green-blue eggshell colors are associated with that of biliverdin IX and biliverdin IX zinc chelate.

To date, previous studies on the genetic mechanisms underlying pigment deposition have mainly focused on protein-coding genes. The gene solute carrier organic anion transporter family member 1B3 (*SLCO1B3*) is known to be responsible for a causative mutation for the blue eggshell phenotype and is specifically expressed in the uterus, not in the other organs in chickens (5). However, the underpinning mechanism of non-coding RNAs on the pigment deposition remains unknown. A major reason is that the functional annotation of long non-coding RNAs (lncRNAs) is largely missing. lncRNAs comprise a heterogeneous subset of RNAs that are longer than 200 nucleotides (nt) and transcribed regions without protein-coding potential. Increased advances have shown that many lncRNAs not only are transcriptional “noise,” but also play an important role in numerous biological processes including transcriptional regulation (6, 7), cell cycle and apoptosis (8), and pluripotency and differentiation control (9, 10). Thus, extensive research is required to fully define and integrate lncRNAs into genome biology. Figuring out the role of lncRNAs would better understand the underlying genetic aspects of non-coding RNA on the deposition of blue eggshells.

China has a wide variety of indigenous poultry, with 108 native chicken breeds. The intensive selection for layers producing blue eggshells has been undergoing for a few decades due to demands by consumers. Nevertheless, blue shell eggs show dark blue, light blue, and median color brown-greenish blue, of which brown-greenish shell eggs are especially not found by consumers. Notably, Changshun blue-eggshell chicken is one of the native breeds mainly producing blue eggshell, but a few of them (unselected individuals) also produce brown and pink color eggs. Thus, it is urgent to figure out the underlying genetic mechanism from non-coding RNA and its targeted genes in breeding selection for blue eggshell layers. In the present study, a high-throughput RNA sequencing (RNA-seq) was employed to profile the uterus transcriptome of Changshun blue-eggshell chickens with different eggshell colors (different proportions between protoporphyrin and biliverdin). The aims were to discover and characterize lncRNAs in chicken uterus tissue and identify key genes, lncRNAs, and pathways that are associated with blue eggshell deposition of chickens.

MATERIALS AND METHODS

Animals and Treatments

The study was approved by the Animal Care Committee of Foshan University (Approval ID: FOSU#080). The experiment was carried out at a breeding farm, Changshun blue eggshell layer, Tiannong Corporation, Guizhou province. A total of 331 layers in a house at 210 days old were observed for 30 days. During these days, three hens stably producing dark blue (DB), four hens producing light blue (LB), and dark brown and greenish (between blue and pink, DP), respectively, were selected for this study (see **Figure 1**). Besides, four hens producing pink (PK) shell eggs, as a control group, were also involved in the study. Uterine glands from each bird were collected and immediately stored in dry ice and then at -80°C until further processing.

RNA-seq

Fifteen cDNA libraries were constructed using total uterus RNA. Total RNA was extracted using the RNeasy Mini-Extraction kit according to the manufacturer's instructions (Aidlab, RN2802, China). The 2100 Bioanalyser (Agilent) was used to determine the RNA quality and was further quantified using the ND-2000 (NanoDrop Technologies). Only high-quality RNA sample was used to construct a sequencing library. The relative expression of color deposition-related genes including solute carrier organic anion transporter family member 1C1 (*SLCO1C1*), solute carrier family 16 member 7 (*SLC16A7*), *CD4*, and *ST6* N-acetylgalactosaminide α -2,6-sialyltransferase 4 (*ST6GALNAC4*) was measured. Primer sequence 5.0 was used to design the primer sequence and synthesized by Shanghai Bioengineering limited company (**Supplementary Table 1**). The glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) was selected as a reference gene (11). The purity of RNA was the ratio of OD 260/OD 280 (a measure of protein contamination) in 2.0–2.2, of which 2.2 represented high-quality RNA. The RNA integrity was analyzed by the method of agarose gel electrophoresis, of which the quantity of 28S rRNA was twice than that of 18S rRNA. Then, each qualified RNA sample has reversed the transcript to cDNA using TRUEScript RT MasterMIX in a 20- μl volume containing 1,000 ng RNA and RNase free to 16 μl , 4 μl 5 \times TRUE RT MasterMix under the following conditions: 42°C for 10 min (Aidlab, PC5801, China) for RT-qPCR analyses. RT-qPCR was conducted on qTOWER 2.2 touch (Analytik Jena, Germany) in a 20- μl volume containing 10 μl SYBR \times Premix Ex Taq (Aidlab, PC3302, China), 0.5 μl of each forward and reverse primer (10 μM), 1 μl of cDNA, and 8 μl ddH $_2$ O under the following conditions: 95°C for 15 min; 95°C for 10 s, annealing (see **Table 1**) for 20 s and 72°C for 20 s for 40 cycles. Each amplification was performed for three control replicates and three case replicates. The amplification efficiencies were close to 100%, using the $2^{-\Delta\Delta\text{Ct}}$ method for calculating the relative gene expression levels of a sample.

The total RNA of uterine glands from each was further RNA-seq for mRNA and lncRNA. Strand-specific RNA-seq libraries were generated by TruSeq Stranded Total RNA with RiboZero Gold kit (Illumina, CA, USA) following the manufacturer's recommendations. Sequencing was performed on an Illumina



FIGURE 1 | The description of four different eggshell colors of Changshun blue-eggshell chicken. DB, DP, LB, and PK mean chickens producing dark blue, dark brown and greenish, light blue, and pink eggshell eggs, respectively.

Hiseq 2500 instrument using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina, CA, USA) to generate 150-bp paired-end reads. Quality control and reads statistics were determined by FastQC (0.11.2) (12). Reads containing adapter or poly-N and low-quality reads were discarded, while the remaining clean reads were aligned to the reference chicken genome (*Gallus_gallus*-5.0) using Hisat (2.0.1) (13). Stringtie (1.2.4) was used to assemble mapped transcripts individually (14), and reference gene annotation was supplied to guide the assembly process. Transcripts from all samples were then merged together with Stringtie merge mode to build a consensus set of transcripts across samples to identify lncRNAs and their nearest-neighbor genes. To reduce the false-positive rates, assembled transcripts were obtained as follows to receive candidate lncRNAs: (1) transcripts with two and above two exons and longer than 200 bp; (2) the reads coverage of transcript more than three was calculated using Stringtie (1.2.4); (3) protein coding potency of transcripts were calculated by three software including coding-non-coding-index (score < 0), coding potential calculator (15) (score < 0), and SwissProt. Transcripts were filtered according to the abovementioned requirements and were considered as candidate lncRNAs and were then blasted to chicken lncRNAs in the ALDB v1.0 database. Reference gene annotation was used

to search the nearest-neighboring genes of lncRNAs, and 100 kb was set as the threshold. Differentially expressed lncRNAs were calculated for further prediction, of which those located within the 100-kb distance of the differentially expressed lncRNAs were selected as potential target genes to reduce false positives. The cis role of lncRNAs was on neighboring target genes (16, 17). The quantification of lncRNAs and mRNAs in each sample was calculated by Stringtie. Differentially expressed mRNAs and lncRNAs of uterine glands among different eggshell color chickens were analyzed using the ballgown (2.6.0) R package (18). p -value < 0.05 and |fold-change| > 2 were considered as significance threshold. Gene ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of differentially expressed mRNAs and lncRNAs were carried out by Goatoools (<https://github.com/tanghaibao/Goatoools>) and KOBAS (<http://kobas.cbi.pku.edu.cn/home.do>).

RESULTS

Overview of Sequencing and Identification of lncRNA in Chicken Uterus

After quality control, more than 98.2% of the total clean reads with high quality were mapped to Galgal 5.0, and 61,369

TABLE 1 | Differently enriched KEGG pathways with targeted genes of comparisons between chickens with different eggshell colors.

Comparison	Pathways	<i>p</i>	Targeted genes
DB vs. DP	Transporters	0.008	<i>SLC34A1, SLC13A3, SLC35F1, SLC22A6-A, SLC26A1</i>
	Exosome	0.011	<i>SLC34A1, GSN, CUBN, DSP, HSPG2, FN1, FLNA</i>
	Cytoskeleton proteins	0.041	<i>GSN, DSP, Tpm3, FLNA</i>
	Cortisol synthesis and secretion	0.028	<i>Pbx1</i>
	Apoptosis	0.041	<i>ITPR2, CASP6</i>
	Glycerophospholipid metabolism	0.045	<i>Etnk1</i>
	Lipid metabolism	0.044	<i>LPPR4</i>
DB vs. LB	Transporters	<0.001	<i>SLC13A3, SLC35F1, SLC25A47, SLC22A6-A, SLC26A1</i>
	Exosome	0.023	<i>SLC34A1, CUBN, DSP, ANXA11, COL18A1</i>
	Cortisol synthesis and secretion	0.025	<i>Pbx1</i>
	NOD-like receptor signaling pathway	0.028	<i>ITPR2, TRAF5</i>
	Progesterone-mediated oocyte maturation	0.034	<i>PGR, CPEB2</i>
	Glycerophospholipid metabolism	0.035	<i>Etnk1</i>
DP vs. LB	Inositol phosphate metabolism	0.010	<i>MTMR1, IPMK</i>
	Cysteine and methionine metabolism	0.022	<i>SRM</i>
	Phosphatidylinositol signaling system	0.024	<i>MTMR1, IPMK</i>
	Protein phosphatase and associated proteins	0.031	<i>MTMR1, IPMK, SSH2, SH3RF1</i>
PK vs. DB	Transporters	<0.001	<i>SLCO1C1, SLC16A7, SLC34A1, SLC13A3, SLC25A47, SLC22A4, SLC26A1, OCLN, CNNM2, ABCA4</i>
	Progesterone-mediated oocyte maturation	0.002	<i>PGR, BRAF, ANAPC1, CPEB2</i>
	Ascorbate and aldarate metabolism	0.012	<i>Ugt1a9</i>
	Lysosome	0.034	<i>Cd164</i>
	General function prediction only	0.039	<i>ABHD2</i>
	Glycosyltransferases	0.041	<i>STT3B, Ganlt16, Ugt1a9, ST8SIA5</i>
	Butanoate metabolism	0.048	.-
PK vs. DP	Transcription factors	0.004	<i>PGR, FOXP1, Nr6a1, IRF4, Rfx2, Pbx1</i>
	Transporters	0.006	<i>SLCO1C1, SLC35F1, SLC4A4, SLC25A47, SLC35F5, SLC22A4</i>
	Hedgehog signaling pathway	0.007	<i>GSK3B, PTCH1</i>
	Other glycan degradation	0.011	<i>MANBA</i>
	Lysosome	0.016	<i>MANBA</i>
	Nuclear receptors	0.019	<i>PGR, Nr6a1</i>
	Cortisol synthesis and secretion	0.027	<i>Pbx1</i>
PK vs. LB	Phosphatidylinositol signaling system	0.002	<i>PI4KA</i>
	Inositol phosphate metabolism	0.005	<i>PI4KA</i>
	Cortisol synthesis and secretion	0.026	<i>Pbx1</i>
	Membrane trafficking	0.026	<i>PI4KA</i>
	DNA replication proteins	0.039	<i>FOXP1, Pbx1</i>

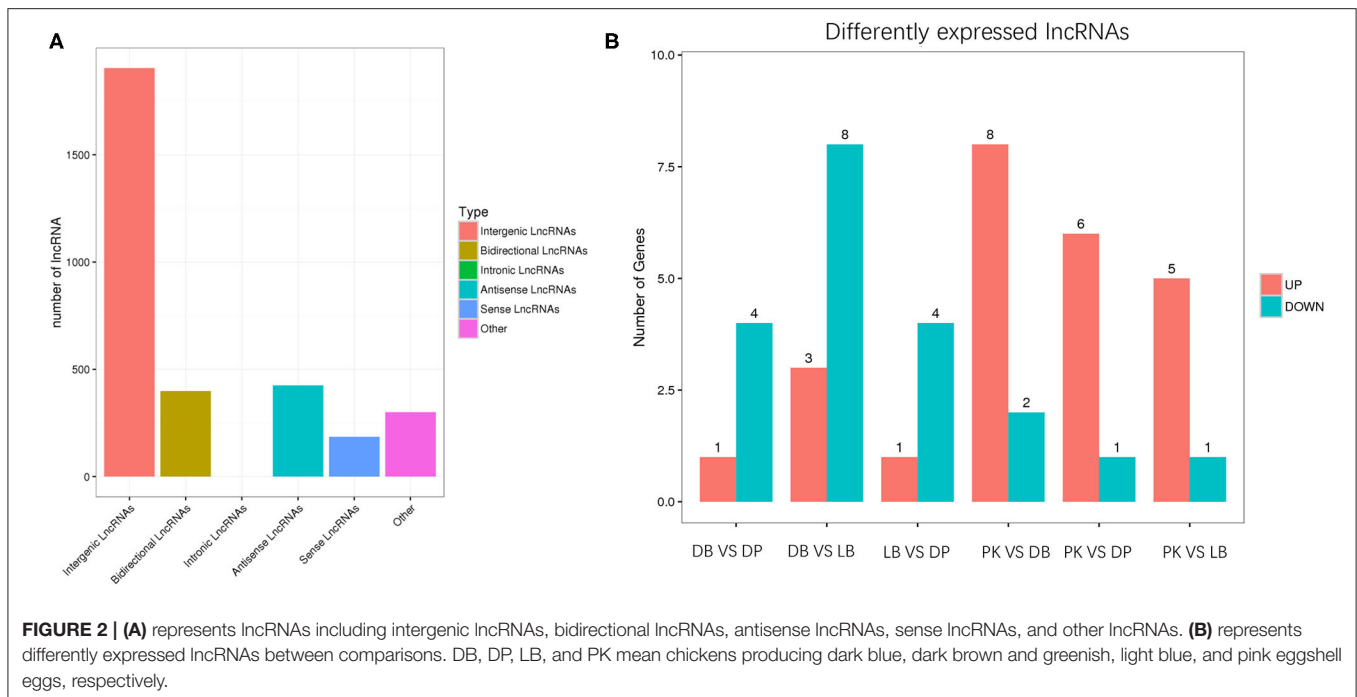
DB, DP, LB, and PK mean chickens producing dark blue, dark brown and greenish, light blue, and pink eggshell eggs, respectively.

assembled transcripts were produced. Detailed information on data quality and mapping statistics is presented in **Supplementary Table 2**. As a result, 6,275 candidate lncRNAs were captured, with 3,060 known by blasting against the known chicken lncRNAs in ALDB database and 3,215 new lncRNAs. Of the new lncRNAs, there were 1,904 intergenic lncRNAs, 399 bidirectional lncRNAs, 426 antisense lncRNAs, 186 sense lncRNAs, and 300 others (**Figure 2**).

Genomic Feature of lncRNAs

The comparison of differently expressed lncRNAs is presented in **Supplementary Figure 1B**. Between the DB and DP, there

were one upregulated and four downregulated genes in the DP group compared to the DB group ($p < 0.05$). Between the DB and LB, there were three upregulated and eight downregulated genes in the LB group compared to the DB group ($p < 0.05$). Between the DP and LB, there were one upregulated and four downregulated genes in the LB group compared to the DP group ($p < 0.05$). We found that as compared to pink-shell eggs, there were eight upregulated and two downregulated genes in dark green-shell eggs, six upregulated and one downregulated gene in green-blue-shell eggs, and five upregulated and one downregulated gene in light green-shell eggs ($p < 0.05$).



There are more known (16,910 on average) and novel mRNAs (14,207 on average) than known (1,430 on average) and novel lncRNAs (2,083 on average) in all the four groups. The heatmaps displayed differentially expressed lncRNAs (**Supplementary Figure 1**) and mRNAs (**Supplementary Figure 2**).

Genomic Feature of mRNAs

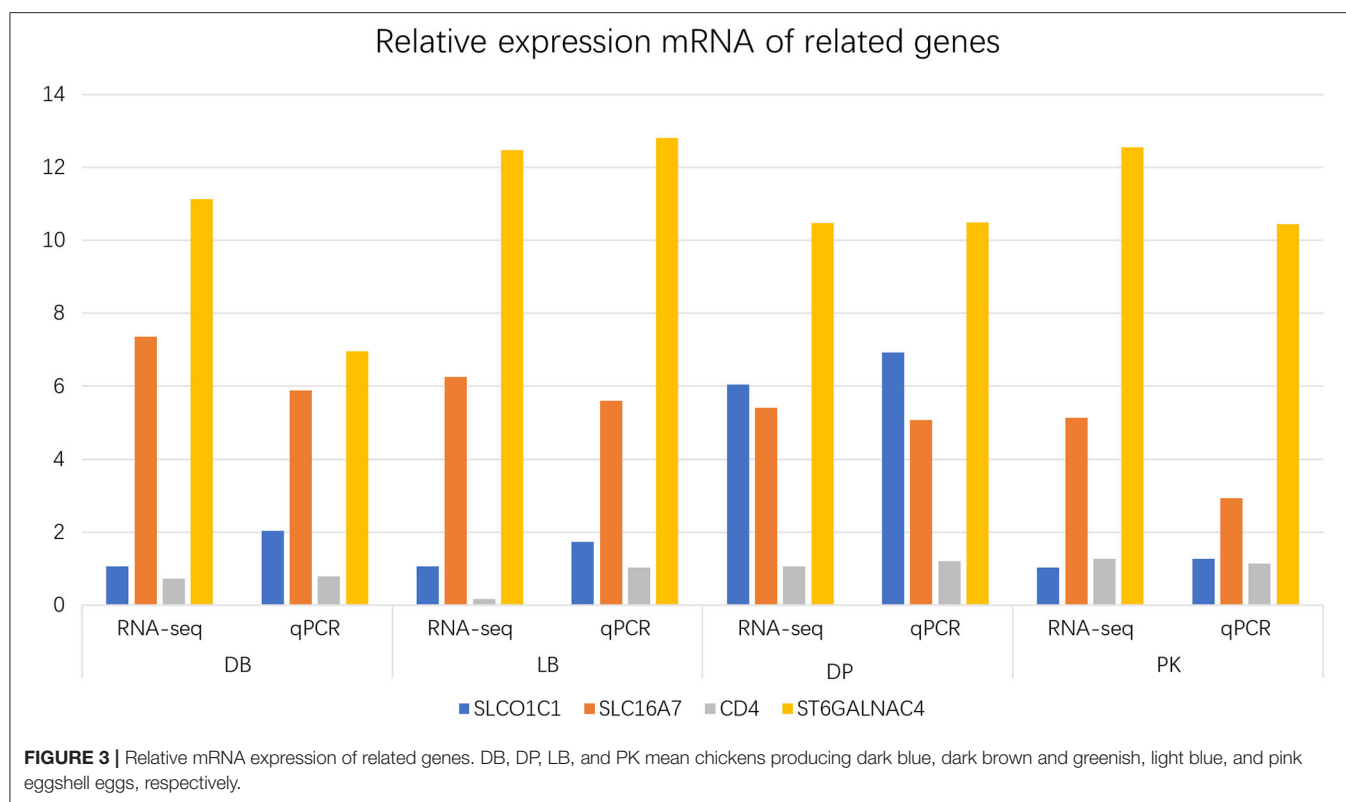
In relation to the mRNA, 24,102 (79.67%) known transcripts and 199,963 new transcripts were obtained. The differentially expressed genes between comparisons are shown in **Supplementary Figure 3**. Furthermore, the relative foldchange of those selected genes in qPCR was consistent with RNA-seq results, suggesting that the transcript identification and abundance estimation were highly reliable (**Figure 3**).

There were no significant GO terms between DB and DP. Furthermore, a total of seven significant KEGG pathways including cortisol synthesis and secretion, lipid metabolism, and glycerophospholipid metabolism were obtained ($p < 0.05$, **Table 1**). SLC family genes were the main targeted genes ($p < 0.05$, **Table 1**). Besides, gelsolin (*GSN*), cubilin (*CUBN*), desmoplakin (*DSP*), heparan sulfate proteoglycan 2 (*HSPG2*), fibronectin 1 (*FN1*), filamin A (*FLNA*), pre-B-cell leukemia transcription factor 1 (*Pbx1*), inositol 1,4,5-trisphosphate receptor type 2 (*ITPR2*), caspase 6 (*CASP6*), phospholipid phosphatase related 4 (*PLPPR4*), and ethanolamine kinase 1 (*Etnk1*) were also differently enriched between groups, of which the relative expression of *GSN*, *CUBN*, *Pbx1*, and *PLPPR4* were upregulated, whereas *DSP*, *HSPG2*, *FN1*, *FLNA*, *ITPR2*, *CASP6*, *Etnk1*, and SLC family genes were downregulated in the former group as compared to the latter group ($p < 0.01$).

Between the comparison of DB vs. LB, the GO term hindbrain tangential cell migration of biological process was significantly enriched, with targeting on gene *PLXNA2* ($p < 0.05$, **Supplementary Table 3**). Cortisol synthesis and secretion and five other pathways were significantly enriched between the two groups ($p < 0.05$, **Table 1**). *CUBN*, *DSP*, Annexin A11 (*ANXA11*), Collagen type XVIII alpha 1 chain (*COL18A1*), *Pbx1*, *ITPR2*, TNF receptor associated factor (*TRAF5*), progesterone receptor (*PGR*), cytoplasmic polyadenylation element binding protein 2 (*CPEB2*), *Etnk1*, and SLC family genes were targeted genes between the two groups, of which the relative expression of SLC family genes, *CUBN*, *COL18A1*, *Pbx1*, *ITPR2*, *TRAF5*, *PGR*, and *CPEB2* was upregulated, while that of *DSP*, *ANXA11*, and *Etnk1* was downregulated in the DB group as compared to the LB group ($p < 0.05$, **Table 1**).

The GO term of phosphatidylinositol monophosphate phosphatase activity with targeted gene *MTMR1* was found to be enriched between DP and DL ($p < 0.05$, **Supplementary Table 3**). Inositol phosphate metabolism, cysteine and methionine metabolism, phosphatidylinositol signaling system, and protein phosphatase and associated proteins were identified to be enriched between the two groups ($p < 0.05$, **Table 1**). Targeted genes myotubularin related protein 1 (*Mtmr1*), inositol polyphosphate multikinase (*IPMK*), spermidine synthase (*SRM*), slingshot protein phosphatase 2 (*SSH2*), and SH3 domain containing ring finger 1 (*SH3RF1*) were identified, and the relative expression of all these genes was upregulated in the DP group compared to the DL group ($p < 0.05$, **Table 1**).

There were no significant GO terms between the comparison of PK vs. DB. However, seven pathways including transporters, progesterone-mediated oocyte maturation, ascorbate and



aldarate metabolism, lysosome, glycosyltransferases, and butanoate metabolism were found to be significantly enriched ($p < 0.05$, **Table 1**). These pathways were related to targeted genes including SLC family genes, occludin (*Ocln*), adaptor related protein complex 1 gamma 1 subunit (*APIG1*), ATP binding cassette subfamily A member 4 (*ABCA4*), *PGR*, heat shock protein 90 (*Hsp90*), UDP glycosyltransferase (*UGT*), L-gulonogamma-lactone oxidase (*GULO*), raf proto-oncogene, serine/threonine kinase (*RAF*), legumain (*LGMN*), anaphase promoting complex subunit 1 (*ANAPC1*), *CPEB*, *CD164*, abhydrolase domain containing 2 (*ABHD2*), polypeptide N-acetylgalactosaminyltransferase 16 (*Galnt16*), and ST8 alpha-N-acetylneuraminide alpha-2,8-sialyltransferase 5 (*ST8SIA5*), of which the relative expression of gene *Ocln*, *PGR*, *CPEB*, *Hsp90*, *UGT*, *GULO*, *Galnt16*, and *ST8SIA5* was upregulated, while *ABCA4*, *RAF*, *CD164*, *LGMN*, *CPEB*, and SLC family genes were downregulated in the former group as compared to the latter group ($p < 0.01$).

There were eight GO terms and one GO term significantly enriched for cellular components and molecular function between PK and DP ($p < 0.05$, **Supplementary Table 2**). These GO terms were mainly involved with immune activities including targeted genes ENSGALT00000018840 and TCONS_00059917. Furthermore, transcription factors, transporters, hedgehog signaling pathway, lysosome, nuclear receptors, and cortisol synthesis and secretion were identified to be enriched between the two groups ($p < 0.05$, **Table 1**). Targeted genes, such as SLC family genes, *PGR*, forkhead box P1 (*FOXP1*), nuclear receptor

subfamily 6 group A member 1 (*Nr6a1*), interferon regulatory factor 4 (*IFR4*), *Pbx1*, glycogen synthase kinase 3 alpha (*GSK3B*), patched 1 (*PTCH1*), and mannosidase beta (*MANBA*), were located, of which the relative expression of *Nr6a1*, *GSK3B*, *Pbx1*, and *PTCH1* was upregulated, while *PGR*, *FOXP1*, *IFR4*, *MANBA*, and SLC family genes were downregulated in the former group as compared to the latter group ($p < 0.01$).

Between the PK and LB groups, four and three GO terms for molecular function and biological process ($p < 0.05$, **Supplementary Table 2**) and pathways including the metabolism of protein family and cortisol synthesis and secretion pathways with genes phosphatidylinositol 4-kinase alpha (*PI4KA*), *Pbx1*, and *FOXP1* were significantly enriched ($p < 0.05$, **Table 1**), of which the relative expression of *Pbx1* was regulated, while *PI4KA* and *FOXP1* were downregulated in the former group as compared to the latter group ($p < 0.01$).

Interaction Analyses of lncRNAs and mRNA

The RNAplex was used to find the interaction between two long-chain RNA, to predict the complementary binding between antisense/cis lncRNA and mRNA. The program includes the Vienna RNA package and calculates the minimum free energy according to its thermodynamic structure to predict the best base-pairing relationship.

For the comparison of DB vs. DP, there were 153 differentially expressed lncRNAs. ECM-receptor interaction with candidate gene *CD4* binding to lncRNA TCONS_00009117 was identified

TABLE 2 | Differently enriched KEGG pathways with targeted genes of lncRNA and mRNA of comparisons between chickens with different eggshell colors.

Comparison	Pathway	p	lncRNA	mRNA	Symbol
DB vs. DP	ECM–receptor interaction	0.040	TCONS_00009117	ENSGALT00000037104	<i>CD4</i>
DB vs. LB	Cell adhesion molecules	0.042	TCONS_00009117	ENSGALT00000037104	<i>CD4</i>
	RNA transport	0.049	TCONS_00067464	ENSGALT00000008544	-
DP vs. LB	Focal adhesion	0.011	TCONS_00002665	ENSGALT00000082154	<i>VWF</i>
PK vs. DB	Cytokine–cytokine receptor interaction	0.005	TCONS_00044519	ENSGALT00000000328	<i>GH</i>
	Jak–STAT signaling pathway	0.031	TCONS_00044523	ENSGALT00000000328	<i>GH</i>
PK vs. DP	Cytokine–cytokine receptor interaction	0.018	TCONS_00034554	ENSGALT00000070014	<i>CX3CR1</i>
	SNARE interactions in vesicular transport	0.037	TCONS_00075755	ENSGALT00000082865	-
	Neuroactive ligand–receptor interaction	0.038	TCONS_00044519	ENSGALT00000000328	<i>GH</i>
	Tryptophan metabolism	0.046	TCONS_00047705	ENSGALT00000016138	<i>HAAO</i>
PK vs. LB	ECM–receptor interaction	0.040	TCONS_00000515	ENSGALT00000012832	<i>LAMB1</i>
	Spliceosome	0.050	ENSGALT00000029552	ENSGALT00000040616	<i>ENSGALT00000040616</i>

DB, DP, LB, and PK mean chickens producing dark blue, dark brown and greenish, light blue, and pink eggshell eggs, respectively.

to be significantly different between groups ($p < 0.05$, **Table 2**). With respect to cis analysis, 70 targeted lncRNAs were found to be involved in the adjacent protein coding of mRNA. These related mRNAs were significantly enriched on toll-like receptor signaling pathway, retinol metabolism, phagosome, spliceosome, glycosphingolipid biosynthesis-ganglio series, glycerolipid metabolism, glycerophospholipid metabolism, and ECM–receptor interaction different pathways ($p < 0.05$, **Table 3**). These pathways were mainly enriched on genes toll-like receptor 2 family member A (*TLR2A*), retinol saturase (*RETSAT*), Catenin beta like 1 (*CTNBL1*), ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 4 (*ST6GALNAC4*), lysocardiolipin acyltransferase 1 (*LCLAT1*), and diacylglycerol kinase zeta (*DGKZ*).

Between the comparison of DB vs. LB, 153 differentially expressed lncRNAs were identified, of which 9 lncRNAs were associated with the binding to sense chain of mRNA by antisense analysis. Cell adhesion molecules with targeted gene *CD4* binding to lncRNA TCONS_00009117 and RNA transport pathways were differently identified ($p < 0.05$, **Table 2**). For cis analysis, 54 targeted lncRNAs were found to be involved in the adjacent protein coding of mRNA. TGF-beta signaling pathway, glycosphingolipid biosynthesis-ganglio series, pyruvate metabolism, propanoate metabolism, biosynthesis of secondary metabolites, selenocompound metabolism, and metabolic pathways were identified to be different between groups ($p < 0.05$, **Table 3**). These pathways were found to relate with differentially expressed genes including Sp1 transcription factor (*Sp1*), *ST6GALNAC4*, *LCLAT1*, and thioredoxin reductase 1 (*TXNRD1*).

Between the comparison of DP vs. LB, there were 224 differentially expressed lncRNAs, of which 18 lncRNAs were associated with the binding to sense chain of mRNA by antisense analysis. Focal adhesion pathway with a candidate gene von Willebrand factor (*VWF*) binding to lncRNA TCONS_00002665 was identified between groups ($p < 0.05$, **Table 2**). For cis analysis, 115 targeted lncRNAs were found to be involved in the adjacent protein coding of mRNA.

These related mRNAs were significantly enriched on 19, 41, and 223 GO terms for cellular components, molecular function, and biological process including pigment metabolic process, respectively ($p < 0.05$). Retinol metabolism, other types of O-glycan biosynthesis, drug metabolism—other enzymes, cell cycle, N-glycan biosynthesis, Wnt signaling pathway, biosynthesis of secondary metabolites, p53 signaling pathway, and neurotrophin signaling pathways were identified ($p < 0.05$, **Table 3**). *RETSAT*, glucoside xylosyltransferase 1 (*GXYLT1*), hypoxanthine phosphoribosyltransferase 1 (*HPRT1*), MDM2 proto-oncogene (*MDM2*), tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon (*YWHAE*), mannosidase, alpha, class 1C, member 1 (*MAN1C1*), chitobiosyldiphosphodolichol beta-mannosyltransferase (*ALG1*), *APC*, mannose phosphate isomerase (*MPI*), and caspase 14, apoptosis-related cysteine peptidase (*CASP14*) as candidate genes were identified to be involved with these pathways.

Between the comparison of PK vs. DB, 275 differentially expressed lncRNAs were identified. Furthermore, 23 targeted lncRNAs of antisense with functional annotation information were identified. Environmental information processing of cytokine–cytokine receptor interaction and Jak–STAT signaling pathways with candidate gene *GH* binding to TCONS_00044519 and TCONS_00044523 lncRNAs was identified ($p < 0.05$, **Table 2**). For cis analysis, 99 targeted genes with functional annotation information were obtained ($p < 0.05$). Besides, proteasome, TGF-beta signaling, vitamin B6 metabolism, selenocompound metabolism, retinol metabolism, sulfur metabolism, glycosphingolipid biosynthesis-ganglio series, spliceosome, and cell cycle pathways with candidate genes including proteasome subunit beta 1/3 (*PSMB1/3*), *Sp1*, 3-phosphoadenosine 5-phosphosulfate synthase 1 (*PAPSS1*), *TXNRD1*, *RETSAT*, RNA binding motif protein, X-linked like 1 (*RBMXL1*), phosphatase, orphan 2 (*PHOSPHO2*), and retinoblastoma 1 (*RB1*) were obtained ($p < 0.05$, **Table 3**).

Between the comparison of PK vs. DP, there were 368 differentially expressed lncRNAs, of which 34 targeted lncRNAs

TABLE 3 | Differently enriched KEGG pathways with targeted genes of lncRNA of comparisons between chickens with different eggshell colors.

Comparison	Pathway	<i>p</i>	Targeted genes	lncRNA
DB vs. DP	Toll-like receptor signaling pathway	< 0.001	<i>TLR2A</i>	TCONS_00055576
	Retinol metabolism	< 0.001	<i>RETSAT</i>	ENSGALT00000081853
	Phagosome	< 0.001	<i>TLR2A</i>	TCONS_00055576
	Spliceosome	< 0.001	<i>CTNBL1</i>	TCONS_00038047
	Glycosphingolipid biosynthesis-ganglio series	0.002	<i>ST6GALNAC4</i>	ENSGALT00000072290
	Glycerophospholipid metabolism	0.045	<i>LCLAT1, DGKZ</i>	ENSGALT00000048295
DB vs. LB	TGF-beta signaling pathway	< 0.001	<i>Sp1</i>	TCONS_00054721
	Glycosphingolipid biosynthesis-ganglio series	< 0.001	<i>ST6GALNAC4</i>	ENSGALT00000072290
	Pyruvate metabolism	0.002	<i>ENSGALT00000074835</i>	ENSGALT00000051137
	Propanoate metabolism	0.003	<i>ENSGALT00000074835</i>	ENSGALT00000051137
	Biosynthesis of secondary metabolites	0.016	<i>LCLAT1</i>	TCONS_00050772
	Selenocompound metabolism	0.034	<i>TXNRD1</i>	TCONS_00008143
	Metabolic pathways	0.039	<i>ST6GALNAC4, DPM2, LCLAT1</i>	ENSGALT00000072290
DP vs. LB	Retinol metabolism	< 0.001	<i>RETSAT</i>	ENSGALT00000081853
	Other types of O-glycan biosynthesis	0.011	<i>GXYLT1</i>	ENSGALT00000069802
	Drug metabolism-other enzymes	0.016	<i>HPRT1, MDM2</i>	TCONS_00058295
	Cell cycle	0.029	<i>YWHA, MDM2</i>	ENSGALT00000071577
	N-Glycan biosynthesis	0.033	<i>MAN1C1, ALG1</i>	TCONS_00040933
	Wnt signaling pathway	0.038	<i>APC</i>	TCONS_00040933
	Biosynthesis of secondary metabolites	0.045	<i>HPRT1, MPI</i>	ENSGALT00000009843
	p53 signaling pathway	0.046	<i>MINM2, CASP14</i>	ENSGALT00000071577
PK vs. DB	Neurotrophin signaling pathway	0.046	-	ENSGALT00000077105
	Proteasome	< 0.001	<i>PSMB1/3</i>	ENSGALT00000088176
	TGF-beta signaling pathway	< 0.001	<i>Sp1</i>	TCONS_00040436
	Selenocompound metabolism	< 0.001	<i>PAPSS1, TXNRD1</i>	TCONS_00056173
	Retinol metabolism	< 0.001	<i>TXNRD1, RETSAT</i>	ENSGALT00000081853
	Sulfur metabolism	< 0.001	<i>RETSAT, TXNRD1</i>	TCONS_00056173
	Glycosphingolipid biosynthesis-ganglio series	0.003	<i>RETSAT</i>	ENSGALT00000072290
	Spliceosome	0.024	<i>RBML1</i>	ENSGALT00000070508
	Vitamin B6 metabolism	0.025	<i>PHOSPHO2</i>	ENSGALT00000085067
	Cell cycle	0.025	<i>RB1</i>	ENSGALT00000051325
PK vs. DP	beta-Alanine metabolism	< 0.001	<i>HIBCH</i>	TCONS_00071353
	p53 signaling pathway	< 0.001	<i>SESNI, CASP14</i>	ENSGALT00000078895
	Propanoate metabolism	< 0.001	<i>HIBCH</i>	TCONS_00071353
	RNA degradation	0.001	<i>XRNI</i>	ENSGALT00000084047
	Porphyrin and chlorophyll metabolism	0.002	<i>FECH</i>	ENSGALT00000087341
	Valine, leucine and isoleucine degradation	0.002	<i>HIBCH</i>	TCONS_00071353
	Sulfur metabolism	0.003	<i>PAPSS1</i>	TCONS_00056173
	Retinol metabolism	< 0.001	<i>RETSAT</i>	ENSGALT00000081853
PK vs. LB	N-Glycan biosynthesis	0.003	<i>MAN1C1, ALG1</i>	TCONS_00040933
	Cardiac muscle contraction	0.008	<i>COX6A1</i>	TCONS_00023198
	Metabolic pathways	0.013	<i>MAN1C1, ALG1</i>	TCONS_00040933
	Oxidative phosphorylation	0.014	<i>COX6A1</i>	TCONS_00023198
	Protein processing in endoplasmic reticulum	0.032	<i>MAN1C1, ALG1</i>	TCONS_00040933
	SNARE interactions in vesicular transport	0.048	<i>VAMP3</i>	TCONS_00039175

DB, DP, LB, and PK mean chickens producing dark blue, dark brown and greenish, light blue, and pink eggshell eggs, respectively.

associated with the binding to sense chain of mRNA by antisense analysis. Cytokine–cytokine receptor interaction, SNARE interactions in vesicular transport, neuroactive ligand–receptor interaction, and tryptophan metabolism pathways with

candidate genes C-X3-C motif chemokine receptor 1 (*CX3CR1*), *GH*, and 3-hydroxyanthranilate 3,4-dioxygenase (*HAAO*) were significantly identified ($p < 0.05$, Table 2). In relation to cis analysis, 168 targeted lncRNAs were located, having potential

roles involving the adjacent protein coding of mRNA. These related mRNAs were significantly enriched on 44, 35, and 57 GO terms for cellular components, molecular function, and biological process, respectively ($p < 0.05$). Beta-alanine metabolism, p53 signaling pathway, propanoate metabolism, RNA degradation, porphyrin and chlorophyll metabolism, valine, leucine, and isoleucine degradation, and sulfur metabolism pathways with 3-hydroxyisobutyryl-CoA hydrolase (*HIBCH*), sestrin 1 (*SESN1*), *CASP14*, 5,-3, exoribonuclease 1 (*XRN1*), ferrochelatase (*FECH*), and *PAPSS1* as candidate genes were significantly identified between groups ($p < 0.05$, **Table 3**).

Between the comparison of PK vs. LB, 189 differentially expressed lncRNAs were identified, of which 10 lncRNAs were associated with the binding to sense chain of mRNA by antisense analysis. These targeted mRNAs were enriched on 15, 3, and 48 GO for cellular components, molecular function, and biological process, respectively ($p < 0.05$). ECM-receptor interaction and spliceosome different pathways were located ($p < 0.05$, **Table 2**). For cis analysis, 67 targeted lncRNAs were found to be involved in the adjacent protein coding of mRNA. These related mRNAs were significantly enriched on 44, 35, and 57 GO terms for cellular components, molecular function, and biological process, respectively ($p < 0.05$). Retinol metabolism, N-Glycan biosynthesis, cardiac muscle contraction, oxidative phosphorylation, protein processing in endoplasmic reticulum, fatty acid elongation, biosynthesis of unsaturated fatty acids, and SNARE interactions in vesicular transport were significantly identified between groups ($p < 0.05$, **Table 3**). Besides, candidate genes *RETSAT*, mannosidase, alpha, class 1C, member 1 (*MAN1C1*), *ALG1*, cytochrome c oxidase subunit 6A1 (*COX6A1*), and vesicle-associated membrane protein 3 (*VAMP3*) were identified between groups.

DISCUSSION

Profiling of the mRNA Sequencing

With respect to the coding proteins, we found that different GO terms between pink-eggshell and blue-eggshell chickens were mainly targeting immune- and transporter-related terms with SLC family, *Igf*, CD family, topoisomerase (DNA) III beta (*Top3b*), and *MTMR* genes. Our results again provide evidence of the candidate gene *SLCO1B3* for a causative mutation on pigment deposition of the blue eggshell phenotype (5) and also an immune relative implication. Within those blue-eggshell chickens, the GO terms hindbrain tangential cell migration and phosphatidylinositol monophosphate phosphatase activity with targeted gene *Plxna2* and *Mtmr1* were specifically identified in dark blue-eggshell chickens. Both the *Plxna2* and *Mtmr1* are known to refer to nerve systems (19) (Pasterkamp), which indicates divergence of nerve function among different degrees of blue-eggshell chickens.

In relation to the KEGG pathway, the lysosome pathway targeted with immune-related genes *CD164* and *IRF4* was significantly enriched in the DB as compared to the PK group. Protoporphyrin IX, as a precursor of heme, is ubiquitously present in all living cells in small amounts, which is involved

with inflammation (20). The zinc protoporphyrin disodium (Zn-PP-2Na) is demonstrated to have anti-inflammatory properties by inhibiting type II collagen-induced arthritis in mice (21). In an overview of the differently enriched pathways, we found that several pathways of the comparison of PK vs. DP coincided with both the comparison of PK vs. DB and PK vs. LB. Besides, each comparison has its pathways varied among other comparisons. These indicate that different genetic mechanisms of chickens with different eggshell colors and the median color between pink and blue did, to some extent, overlap both the two biological functions of pink- and dark blue-eggshell chickens. This finding is also evidenced by the current results that showed that the relative expression of the SLC family is higher in the DB and the DP groups than in the PK group, and in the DB group than in the DP and the LB groups. Notably, targeted SLC family-related genes were not different between the comparison of chickens producing light blue eggshell eggs and pink shell eggs, as well as chickens with light blue shell and brown-blue shell eggs. These imply that the mechanism of deposition of blue pigment is different between the light and dark blue eggshell, of which the deposition of dark and brown-greenish blue is involved with the SLC family gene, but not the light blue pigment.

Clearly, the transporter pathway is mainly targeting SLC family genes and is involved with chickens producing pink and light/brown eggs. The progesterone-mediated oocyte maturation pathway with targeted gene *PGR* was significantly enriched in dark eggshell chickens as compared to light and pink-eggshell chickens. This implies that the progesterone may be involved with the pigment deposition when oviposition occurs. The cuticle is related to the deposition of, or contains to some extent, the pigment deposition of the brown eggshell pigment (22). Indeed, the progesterone, as a factor controlling cuticle deposition, is also related to pigment deposition (23). This leads to an explanation that shell strength associated with cuticle deposition increased as the darkness of the shell increased (24). Besides, the cortisol synthesis and secretion pathway and the relative expression of targeted gene *Pbx1* upregulated with the darkness of the shell decreased. It is generally known that the cuticle and pigment deposition is affected by a mild environmental stressor that further causes temporary inhibition of the reproductive axis and an increase in circulating corticosteroids (25). A previous study demonstrated that environmental contamination (one of the environmental stressors) negatively correlated with the blue-green chroma (26). These indicate that the increased stress may be associated with the decreased coloration of eggshell, but more work is needed to investigate what induces its release.

Profiling of the lncRNA Sequencing

lncRNAs in the chicken genome exhibit similar features to those reported in other species, for instance, a significant expression correlation with adjacent protein-coding genes and a high level of tissue specificity. Enrichment analyses of lncRNA-adjacent protein-coding genes also show that chicken lncRNAs likely regulate transcription, cell proliferation, apoptosis, and development (27). In this study, we consider that the potential of lncRNAs would advance our understanding of the genetic mechanisms underlying the trait of pigment deposition of layers.

Our lncRNA results showed low expression, shorter transcript length, and fewer exons among species as compared to mRNA, which agree with the results of previous studies and indicate that the lncRNAs identified here were reliable (28, 29). In an overview, we found some enriched pathways such as cell cycle related to lncRNAs.

Significantly different lncRNAs were associated with the binding to sense chain of mRNA by antisense analysis, of which those mRNAs were further used for GO term and KEGG analysis. As compared to the profiling of mRNA sequencing, there are relatively fewer GO terms and pathways relating to targeted lncRNAs associating with mRNAs. It is likely that the significant pathways between comparisons of PK and different blue shell chickens were mainly on immune- and nerve-related pathways, which is consistent with the above mRNA profiling. The pathway cytokine–cytokine receptor interaction is known to have effects on the activation of interleukin-3 (30), while the Jak–STAT signaling pathway is involved with innate immune response in invertebrates (31) and regulates the gene expression of IL-6 and IL-10 (32). Besides, SNARE proteins are known to be linked to exocytosis of insulin granules in β cells (33). The tryptophan metabolism is associated with the body discoloration of crustacea (34) and the regulation of pigment synthesis in *Malassezia furfur* (35). Notably, the dysfunction of tryptophan metabolism is relative to neurodegenerative diseases, such as tuberculosis (36) and schizophrenia (37). This agrees with the role of the transcript profiling of coding protein on nerve-related function between chickens producing pink and blue shell eggs. These pathways were targeted on *GH*, *CX3CR1*, *HAAO*, *LAMB1*, and ENSGALT00000040616 binding to several lncRNAs including ENSGALT00000000328, ENSGALT00000070014, ENSGALT00000082865, ENSGALT00000016138, ENSGALT00000012832, and ENSGALT00000040616 by antisense analysis, which indicates that these lncRNAs may be involved with the base pairing, gene silencing, and transcription and stability of those binding mRNA. Interestingly, except for some pathways relative to lipids and amino acid metabolisms, porphyrin and chlorophyll metabolism with targeted gene *FECH* and oxidative phosphorylation and cardiac muscle contraction pathways with targeted gene *COX6A1* were found between pink-eggshell and blue-eggshell chickens. These pathways and genes are known to be linked to the color pigments, which implies that lncRNAs also play an important role in affecting adjacent protein coding of mRNAs.

lncRNAs including TCONS_00009117, TCONS_00067464, and TCONS_00002665 were identified to be involved with the eggshell color deposition. These genes and lncRNAs are relative to ECM–receptor interaction, cell adhesion molecules, RNA transport, and focal adhesion pathways. Among chickens producing different blue color eggs, targeted genes *CD4* and *VWF* are located. The immune-related gene found between dark blue-eggshell chickens and brown greenish eggshell chickens are also supported by the cis analysis that shows that the toll-like receptor signaling pathway with the TLR family gene that serves as an immune-related indicator (38) was enriched between the two groups. Among the comparisons of the three blue shell chickens, most pathways were mainly enriched on lipid-related metabolisms, such as retinol metabolism, pyruvate

metabolism, propanoate metabolism, O-glycan biosynthesis, N-Glycan biosynthesis, glycerophospholipid metabolism, and glycosphingolipid biosynthesis-ganglio series pathways. The glycerophospholipid metabolism in both the protein-coding and non-coding genes related to pathways was found between chickens producing different degrees of blue eggshell color. The glycerophospholipid metabolism has predictive value on acute graft vs. host disease (1). The disturbance of hippocampal glycerophospholipid metabolism may result in an imbalance of hippocampal sphingolipid and glycerophospholipid metabolism (39). These literatures somehow indicate that the dark blue color may play a protective role in bacterial infection to animals, but further investigation is needed to explain the interaction between lipid-related metabolisms and blue pigment deposition. The glycerophospholipid-related pathways are also found in mRNA sequencing, suggesting that the lncRNAs (ENSGALT00000072290) exerted similar functions in color formation by modulating the adjacent proteins. Accordingly, *Plxna2*, *Mtmr1*, *Pbx1*, *PGR*, *FECH*, *COX6A1*, *CD4*, and *VWF* may be key candidate genes, and ENSGALT00000000328, ENSGALT00000070014, ENSGALT00000082865, ENSGALT00000016138, ENSGALT00000012832, ENSGALT00000040616, TCONS_00009117, TCONS_00067464, and TCONS_00002665 are key lncRNAs related to the blue color deposition.

CONCLUSION

In conclusion, transcriptome sequencing was first used in this study to generate the expression profile of lncRNAs and mRNAs in the chicken uterus. Integrating analysis of lncRNA and mRNA profiles of most pathways was mainly enriched on lipid-related metabolisms as found in mRNA sequencing. The lncRNAs exerted similar functions in color formation by modulating, including pigment disposition and immune- and lipid-related metabolisms. Our results provide a catalog of chicken uterine lncRNAs and genes worthy of further studies to understand their roles in selection for blue eggshell color layers.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: NCBI [accession: PRJNA746050].

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care Committee of Foshan University (Approval ID: FOSU#080).

AUTHOR CONTRIBUTIONS

HL obtained the funding and designed this project. SC, KC, FL, JD, ZM, JX, GL, and HL performed the experiment. SC, KC, and ZM analyzed and interpreted the data. SC, KC, and HL drafted

and revised the manuscript. All authors came to an agreement for publication.

FUNDING

This work was supported by the Guangdong Provincial Key Laboratory of Animal Molecular Design and Precise Breeding (2019B030301010), the Key Laboratory of Animal Molecular Design and Precise Breeding of Guangdong Higher Education Institutes (2019KSYS011), and the Germplasm Improvement Talent Base in Guizhou Changshun Tiannong Green Shell Laying Hen Industrial Co. Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

REFERENCES

- Liu Y, Huang A, Chen Q, Chen X, Fei Y, Zhao X, et al. A distinct glycerophospholipid metabolism signature of acute graft versus host disease with predictive value. *JCI Insight*. (2019) 4:e129494. doi: 10.1172/jci.insight.129494
- Lang MR, Wells JW. A review of eggshell pigmentation. *Worlds Poult Sci J*. (1987) 43:238–46. doi: 10.1079/WPS19870016
- Ito S. Celadon: an eggshell color mutation in Japanese quail. *J Hered*. (1993) 84:145–7. doi: 10.1093/oxfordjournals.jhered.a111301
- Kennedy GY, Vevers HG. A survey of avian eggshell pigments. *Comp. Biochem. Physiol. B, Biochem.* (1976) 55:117–23. doi: 10.1016/0305-0491(76)90183-8
- Wang Z, Qu L, Yao J, Yang X, Li G, Zhang Y, et al. An EAV-HP insertion in 5' flanking region of SLC01B3 causes blue eggshell in the chicken. *PLoS Genet*. (2013) 9:e1003183. doi: 10.1371/journal.pgen.1003183
- Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*. (2013) 20:300–7. doi: 10.1038/nsmb.2480
- Li T, Wang S, Wu R, Zhou X, Zhu D, Yong Z. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. *Genomics*. (2012) 99:292–8. doi: 10.1016/j.ygeno.2012.02.003
- Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, et al. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet*. (2013) 9:e1003368. doi: 10.1371/journal.pgen.1003368
- Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*. (2011) 477:295–U60. doi: 10.1038/nature10398
- Ng JH, Ng HH. LincRNAs join the pluripotency alliance. *Nat Genet*. (2010) 42:1035–36. doi: 10.1038/ng1210-1035
- Borowska D, Rothwell L, Bailey RA, Watson K, Kaiser P. Identification of stable reference genes for quantitative PCR in cells derived from chicken lymphoid organs. *Vet Immunol Immunopathol*. (2016) 170:20–4. doi: 10.1016/j.vetimm.2016.01.001
- Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. (2011) 27:863–4. doi: 10.1093/bioinformatics/btr026
- Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. (2015) 12:357–60. doi: 10.1038/nmeth.3317
- Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg S. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. (2015) 33:290–95. doi: 10.1038/nbt.3122
- Kong L, Yong Z, Ye ZQ, Liu XQ, Ge G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. (2007) 35:W345–9. doi: 10.1093/nar/gkm391
- UA Ø, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti, G. Long noncoding RNAs with enhancer-like function in human cells. *Cell*. (2010) 143:46–58. doi: 10.1016/j.cell.2010.09.001
- Harrow J, Denoeud F, Frankish A, Reymond A. GENCODE: producing a reference annotation for ENCODE. *Genome Biology*. (2006) 7:1–9. doi: 10.1186/gb-2006-7-s1-s4
- Frazee AC, Perteau G, Jaffe AE, Langmead B, Salzberg SL, Leek, JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol*. (2015) 33:243. doi: 10.1038/nbt.3172
- Pasterkamp RJ, Verhaagen J. Emerging roles for semaphorins in neural regeneration. *Brain Res Brain Res Rev*. (2001) 35:36–54. doi: 10.1016/S0165-0173(00)00050-3
- Metz R, Duhadaway JB, Rust S, Munn DH, Muller AJ. Zinc protoporphyrin IX stimulates tumor immunity by disrupting the immunosuppressive enzyme indoleamine 2,3-Dioxygenase. *Mol Cancer Ther*. (2010) 9:1864. doi: 10.1158/1535-7163.MCT-10-0185
- Nagai H, Kitagaki K, Kuwabara K, Koda A. Anti-inflammatory properties of zinc protoporphyrin disodium (ZnPP2Na). *Agents Actions*. (1992) 37:273–83. doi: 10.1007/BF02028120
- Samiullah S, Roberts JR. The location of protoporphyrin in the eggshell of brown-shelled eggs. *Poult Sci*. (2013) 92:2783–88. doi: 10.3382/ps.2013-03051
- Wilson PW, Suther CS, Bain MM, Icken W, Jones A, Quinlan-Pluck F, et al. Understanding avian egg cuticle formation in the oviduct: a study of its origin and deposition. *Biol Reprod*. (2017) 97:39–49. doi: 10.1093/biolre/iox070
- Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: A domestic-animal long noncoding RNA Database. *PLoS ONE*. (2015) 10:e0124003. doi: 10.1371/journal.pone.0124003
- Dunn IC, Wilson PW, D'Eath RB, Boswell T. Hypothalamic agouti-related peptide mRNA is elevated during natural and stress-induced anorexia. *J Neuroendocrinol*. (2015) 27:681–95. doi: 10.1111/jne.12295
- Hanley D, Doucet SM. Does environmental contamination influence egg coloration? a long-term study in herring gulls. *J Appl Ecol*. (2012) 49:1055–63. doi: 10.1111/j.1365-2664.2012.02184.x
- Kim HJ, Chung JK, Hwang DW, Dong SL, Kim S. *In vivo* imaging of miR-221 biogenesis in papillary thyroid carcinoma. *Mol Imaging Biol*. (2009) 11:71–8. doi: 10.1007/s11307-008-0188-6
- He Y, Ding Y, Zhan F, Zhang H, Han B, Hu G, et al. The conservation and signatures of lincRNAs in Marek's disease of chicken. *Sci Rep*. (2015) 5:15184. doi: 10.1038/srep15184
- Derrien T, Johnson R, Bussotti G, Tanzer A, Guigó R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. (2012) 22:1775–89. doi: 10.1101/gr.132159.111
- Dey R, Ji K, Liu Z, Chen L. A cytokine–cytokine interaction in the assembly of higher-order structure and activation of the interleukine-3:receptor complex. *PLoS ONE*. (2009) 4:e5188. doi: 10.1371/journal.pone.0005188
- Chen C, Eldein S, Zhou X, Sun Y, Gao J, Sun Y, et al. Immune function of a Rab-related protein by modulating the JAK-STAT signaling pathway in

ACKNOWLEDGMENTS

We express our sincere gratitude to the staff of Guizhou Nayong Yuanshengmuye Ltd., Bijie, 553300, Guizhou, China, for the support during the study and for providing the Weining chick. We sincerely appreciate Mr. Rong He and all people in the Qianlong organic farm for their help to the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fvets.2021.736387/full#supplementary-material>

- the silkworm, *Bombyx mori*. *Arch Insect Biochem Physiol*. (2018) 97:e21434. doi: 10.1002/arch.21434
32. Murray PJ. The JAK-STAT signaling pathway: input and output integration. *J Immunol*. (2007) 178:2623–29. doi: 10.4049/jimmunol.178.5.2623
 33. Gerber SH, Südhof T. Molecular determinants of regulated exocytosis. *Diabetes*. (2002) 51(Suppl.1):S3–11. doi: 10.2337/diabetes.51.2007.S3
 34. Negishi S, Hasegawa Y, Naito J, Nagamura Y, Ishiguro I. Involvement of tryptophan metabolism in the body color of crustacea. *Adv Exp Med Biol*. (1999) 467:649–52. doi: 10.1007/978-1-4615-4709-9_83
 35. Barchmann T, Hort W, Kramer HJ, Mayser P. Glycine as a regulator of tryptophan-dependent pigment synthesis in *Malassezia furfur*. *Mycoses*. (2011) 54:17–22. doi: 10.1111/j.1439-0507.2009.01758.x
 36. Nair S, Baron H. Tryptophan metabolism in tuberculosis. *Am Rev Respir Dis*. (1973) 108:977–9.
 37. Benassi CA, Benassi P, Allegri G, Ballarin P. Tryptophan metabolism in schizophrenic patients. *J Neurochem*. (1961) 7:264–70. doi: 10.1111/j.1471-4159.1961.tb13512.x
 38. Snyder-Mackler N, Sanz J, Kohn JN, Brinkworth JF, Tung J. Social status alters immune regulation and response to infection in macaques. *Science*. (2016) 291:1041–5. doi: 10.1126/science.aah3580
 39. Zheng P, Wu J, Zhang H, Perry SW, Yin B, Tan X, et al. The gut microbiome modulates gut-brain axis glycerophospholipid metabolism in a region-specific manner in a nonhuman primate model of depression. *Mol Psychiatry*. (2020) 26:2380–92. doi: 10.1038/s41380-020-0744-2

Conflict of Interest: FL and JD are employed by Guizhou Changshun Tiannong Green Shell Laying Hen Industrial Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chen, Chen, Xu, Li, Ding, Ma, Li and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome-Wide Association Studies for Growth Curves in Meat Rabbits Through the Single-Step Nonlinear Mixed Model

Yonglan Liao¹, Zhicheng Wang¹, Leonardo S. Glória², Kai Zhang³, Cuixia Zhang⁴, Rui Yang⁴, Xinmao Luo¹, Xianbo Jia¹, Song-Jia Lai^{1*} and Shi-Yi Chen^{1*}

¹Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, China, ²Laboratory of Animal Science, State University of Northern of Rio de Janeiro, Campos dos Goytacazes, Brazil, ³Sichuan Academy of Grassland Sciences, Chengdu, China, ⁴Animal Breeding and Genetics Key Laboratory of Sichuan Province, Sichuan Animal Science Academy, Chengdu, China

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Lei Zhou,
China Agricultural University, China
Guillermo Giovambattista,
CONICET Institute of Veterinary
Genetics (IGEVE), Argentina

*Correspondence:

Song-Jia Lai
laisj5794@163.com
Shi-Yi Chen
sychensau@gmail.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 31 July 2021

Accepted: 15 September 2021

Published: 08 October 2021

Citation:

Liao Y, Wang Z, Glória LS, Zhang K,
Zhang C, Yang R, Luo X, Jia X, Lai S-J
and Chen S-Y (2021) Genome-Wide
Association Studies for Growth Curves
in Meat Rabbits Through the Single-
Step Nonlinear Mixed Model.
Front. Genet. 12:750939.
doi: 10.3389/fgene.2021.750939

Growth is a complex trait with moderate to high heritability in livestock and must be described by the longitudinal data measured over multiple time points. Therefore, the used phenotype in genome-wide association studies (GWAS) of growth traits could be either the measures at the preselected time point or the fitted parameters of whole growth trajectory. A promising alternative approach was recently proposed that combined the fitting of growth curves and estimation of single-nucleotide polymorphism (SNP) effects into single-step nonlinear mixed model (NMM). In this study, we collected the body weights at 35, 42, 49, 56, 63, 70, and 84 days of age for 401 animals in a crossbred population of meat rabbits and compared five fitting models of growth curves (Logistic, Gompertz, Brody, Von Bertalanffy, and Richards). The logistic model was preferably selected and subjected to GWAS using the approach of single-step NMM, which was based on 87,704 genome-wide SNPs. A total of 45 significant SNPs distributed on five chromosomes were found to simultaneously affect the two growth parameters of mature weight (A) and maturity rate (K). However, no SNP was found to be independently associated with either A or K. Seven positional genes, including *KCNIP4*, *GBA3*, *PPARGC1A*, *LDB2*, *SHISA3*, *GNA13*, and *FGF10*, were suggested to be candidates affecting growth performances in meat rabbits. To the best of our knowledge, this is the first report of GWAS based on single-step NMM for longitudinal traits in rabbits, which also revealed the genetic architecture of growth traits that are helpful in implementing genome selection.

Keywords: GWAS, longitudinal data, genomic analysis, NMM, body weight

INTRODUCTION

The domestic rabbit (*Oryctolagus cuniculus*) is an important livestock species in China and has been intensively raised for producing meat, wool, and fur. The most commonly raised type is meat rabbits in China, and the rabbit meat production reached 849,150 tons in 2016, which almost accounted for about 60% of global production (Li et al., 2018). However, progresses on genetic selection and improvement in rabbits have obviously lagged behind in comparison with other livestock species; therefore, the Chinese meat rabbit industry is still largely depending on these imported breeds, such

as the Hyla, Hycole, and Hyplus rabbits from France (Qin, 2019). One of the important reasons is the serious lack of relevant studies conducted in rabbits, such as the genome-wide association studies (GWAS) and genomic selection (GS) for the economically important traits. Recently, some pioneer studies were published about the GWAS (Sosa-Madrid et al., 2020; Yang et al., 2020; Bovo et al., 2021) and GS (Chen et al., 2021; Helal et al., 2021; Mancin et al., 2021) in rabbits.

Growth is a complex and economically important trait with moderate to high estimates of heritability in rabbits (Akanno and Ibe, 2005; Dige et al., 2012; Soliman et al., 2014). In contrast to traits that are collected at a single time point (such as litter size and carcass performances), growth must be described by the longitudinal data repeatedly measured over multiple time points. Therefore, the relevant genetic studies on growth in livestock can be implemented through different approaches. The first approach is to select one or a few representative time points and subject them to separate analysis; for instance, the GWAS of growth traits were separately performed among multiple time points of age in meat rabbits (Yang et al., 2020). The second approach is to fit growth curves using nonlinear regression models and obtain the growth curve parameters (such as mature weight and maturity rate), and subsequently, these derived parameters are used as the pseudo-phenotypes for association analysis. This is the classical two-step method and has been commonly found in literature, such as the studies in beef cattle (Crispim et al., 2015; Duan et al., 2021). Furthermore, the two-step method could be followed by an additional step of multi-trait meta-analysis to indirectly combine the multiple parameters together (Duan et al., 2021). Recently, Silva et al. (2017) proposed an alternative modeling framework to integrate the fitting of growth curves and estimation of single-nucleotide polymorphism (SNP) effects simultaneously under nonlinear mixed model (NMM), which was applied to pigs and revealed to have the advantages of higher statistical power and joint modeling of residual effects in comparison with the two-step method. To our best knowledge, this single-step method has not yet been applied to GWAS of growth curves in rabbits.

In this context, we collected the individual growth records from weaning at 35 days of age (DOA) to finishing at 84 DOA in a commercial crossbred population of meat rabbits. Subsequently, the fitting of growth curves and GWAS were simultaneously analyzed using a single-step NMM to identify the prospective candidate variants, genes, and biological processes associated with growth trajectory. These results could be helpful in understanding the biological mechanisms underlying growth and implementing GS of growth traits in rabbits.

MATERIALS AND METHODS

Animals and Phenotypes

One commercial crossbred population of meat rabbits, by crossing 22 Kangda5 rabbits (♂) with 53 Californian rabbits (♀), was subjected to collection of phenotypic records, which was described in our previous study (Yang et al., 2020). In brief, individual body weight (BW) was initially measured for 461 rabbits at seven time points, including 35, 42, 49, 56, 63, 70,

and 84 DOA, respectively. At each time point, the phenotypic records were set to missing values if they deviated by more than three standard deviations (SD) from the population mean. As the short time intervals were measured, the individual BW was allowed to be slightly decreased (<5%) between two consecutive time points; otherwise, the latter record was set to missing value. The individuals that have more than two missing values at the seven time points were also removed, after which 405 individuals remained. No pedigree information is available for this population.

Genotypes and Quality Controls

For the initial SNP set that was generated from specific-locus amplified fragment sequencing approach (Yang et al., 2020), we reapplied more strict criterion of quality control (QC) using the filtering expression of “QualByDepth (QD) < 2.0 || FisherStrand (FS) > 60.0 || RMSMappingQuality (MQ) < 40.0” intrinsically implemented in GATK software v4.2 (McKenna et al., 2010). A total of 6,721,762 SNPs were obtained and subjected to additional QC steps using PLINK software v1.9 (Chang et al., 2015), which required the genotype missing rate lower than 0.1, individual missing rate lower than 0.2, minor allele frequency (MAF) higher than 0.05, and no extreme deviation from Hardy-Weinberg equilibrium (i.e., only retained SNPs with $p > 1.0E-08$). Furthermore, the missing genotypes were imputed using Beagle software v5.1 with default parameters (Browning et al., 2018). Using PLINK software v1.9 (Chang et al., 2015), the tightly linked SNPs were further discarded if the linkage disequilibrium (LD) values were higher than 0.9. Finally, 87,704 SNPs were used for GWAS among 401 individuals (215 males and 186 females), and these SNPs were distributed among all 21 rabbit autosomes (OCU). To investigate population structure, principal component analysis (PCA) was performed based on the finally included genotypes using PLINK software v1.9 (Chang et al., 2015).

Modeling of Growth Curves

Five nonlinear regression models were evaluated for fitting the growth curves (Koya and Goshu, 2013), including the logistic of $w_t = A[1 + b \exp(-Kt)]^{-1}$, Gompertz of $w_t = A \exp[-b \exp(-Kt)]$, Brody of $w_t = A[1 - b \exp(-Kt)]$, Von Bertalanffy of $w_t = A[1 - b \exp(-Kt)]^3$, and Richards of $w_t = A[1 \pm b \exp(-Kt)]^m$. Among them, w_t is the individual BW at time t ; and the parameter A , K , and b are the mature weight, maturity rate, and time-scale parameter, respectively. Furthermore, m is the shape parameter in Richards model. The fitting of growth curves was performed using the *nlme* package of R (Heisterkamp et al., 2017), and the model with the best goodness of fit was selected according to the Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978).

Genome-Wide Association Studies

The logistic was selected as best model (see *Results* section) and therefore used for the GWAS of growth curves through single-step NMM following Silva et al. (2017). This method fitted the two biological meaningful parameters of growth curves (i.e., A

TABLE 1 | The descriptive statistics of body weight at the seven time points.

Days of age	Number of records	Body weight (g)			
		Min	Max	Mean	SD
35	399	456	1,120	788.05	122.65
42	398	741	1,327	1,012.73	112.42
49	401	874	1,657	1,244.87	136.78
56	401	972	2,005	1,474.96	179.96
63	381	974	2,354	1,706.96	235.45
70	371	1,050	2,726	1,948.76	285.08
84	363	1,487	2,888	2,238.13	285.29

Note. SD, standard deviation.

and K) through the NMM. Therefore, the null model (M0) without considering SNP effect was defined as follows:

$$w_{it} = \frac{\mu_A + Sex + PC + \varepsilon_{A_i}}{1 + \mu_b \exp[-(\mu_K + Sex + PC + \varepsilon_{K_i})t]} + e_{it},$$

where w_{it} is the BW of the individual i at time t ; μ_A , μ_K , and μ_b are the general means for parameter A, K, and b, respectively; Sex and PC are the fixed effects of sex and five principal components (PC) of genotype matrix (PC_1 , PC_2 , PC_3 , PC_4 , and PC_5), respectively; ε_{A_i} and ε_{K_i} are the specific residuals for parameter A and K of individual i ; and e_{it} is a general residual of individual i at time t , assumed with $e_{it} \sim N(0, \sigma_e^2)$. The assumed (co)variance structures of ε_{A_i} and ε_{K_i} were as follows:

$$\begin{bmatrix} \varepsilon_{A_i} \\ \varepsilon_{K_i} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_A^2 & \sigma_{A,K} \\ \sigma_{A,K} & \sigma_K^2 \end{bmatrix}\right),$$

where σ_A^2 , σ_K^2 , and $\sigma_{A,K}$ are the specific residual variances and covariance for parameter A and K.

According to Silva et al. (2017), the SNP effects could be further integrated into the null model through three different ways. First, the SNP effects are assumed to simultaneously affect both A and K parameters, and this full model (M1) was given as follows:

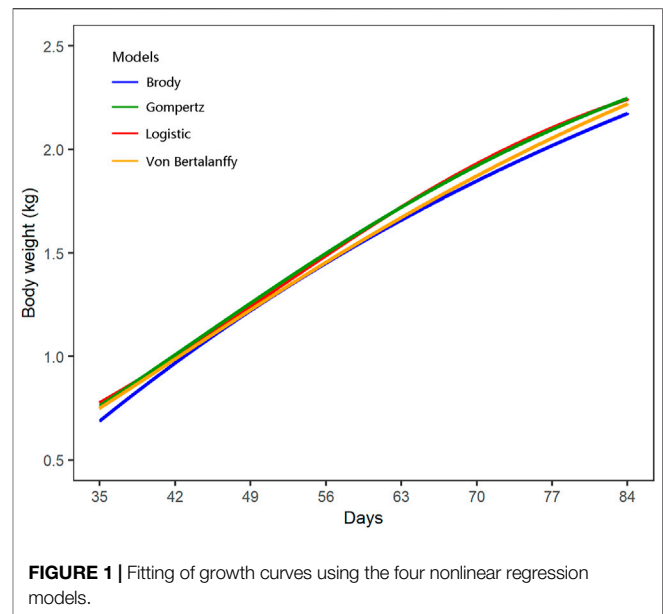
$$w_{it} = \frac{\mu_A + Sex + PC + SNP + \varepsilon_{A_i}}{1 + \mu_b \exp[-(\mu_K + Sex + PC + SNP + \varepsilon_{K_i})t]} + e_{it},$$

where SNP is fixed effects. Alternatively, the SNP effects independently affect either A (M2) or K (M3), and their models were, respectively, given as follows:

$$w_{it} = \frac{\mu_A + Sex + PC + SNP + \varepsilon_{A_i}}{1 + \mu_b \exp[-(\mu_K + Sex + PC + \varepsilon_{K_i})t]} + e_{it},$$

$$w_{it} = \frac{\mu_A + Sex + PC + \varepsilon_{A_i}}{1 + \mu_b \exp[-(\mu_K + Sex + PC + SNP + \varepsilon_{K_i})t]} + e_{it}.$$

The fitting of these four NMM was performed using the *nlme* package of R (Heisterkamp et al., 2017). Based on the likelihood ratio test (LRT), the statistical significance of SNP effects could be deduced by comparing the specific alternative hypotheses (i.e., the model of M1, M2, or M3) with the null hypothesis of M0, respectively. The derived LRT statistics are assumed to follow χ^2 distribution with n degrees of freedom, where n is the difference of the number of parameters between the two

**FIGURE 1** | Fitting of growth curves using the four nonlinear regression models.

models compared. To address the multiple comparison problem, the false discovery rate (FDR) method was employed for computing the adjusted p -values using the *qvalue* package of R (Storey et al., 2004). As a result, SNP was statistically significant with FDR < 0.05.

Functional Analysis

In this study, the QTLs were empirically defined as chromosomal regions of ± 100 kb around the significant SNPs (i.e., a total of 200-kb genomic region was selected). The candidate genes within QTL, including protein encoding and long non-coding RNAs (lncRNA), were retrieved using the biomaRt R package (Smedley et al., 2015). The OryCun2.0 assembly was used as the reference genome (<https://www.ncbi.nlm.nih.gov/genome/?term=rabbit>). For all the candidate genes, the functional enrichments were conducted using the DAVID tool (Huang et al., 2009), including the Gene Ontology (GO) terms (The Gene Ontology Consortium, 2019) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa et al., 2019). The default parameters and method of multiple testing correction were used for computing p -values, and the threshold of 0.05 was set.

RESULTS

Descriptive Statistics

For the 401 finally included individuals, the descriptive statistics of BW at the seven time points are shown in Table 1, and their normal distributions were visually checked at every time points (Supplementary Figure S1). The 87,704 SNPs were distributed among 21 autosomes with the mean (\pm SD) of $24,099 \pm 59,103$ bp for their pairwise physical distances and 0.256 ± 0.133 for MAF, respectively (Supplementary Figure S2).

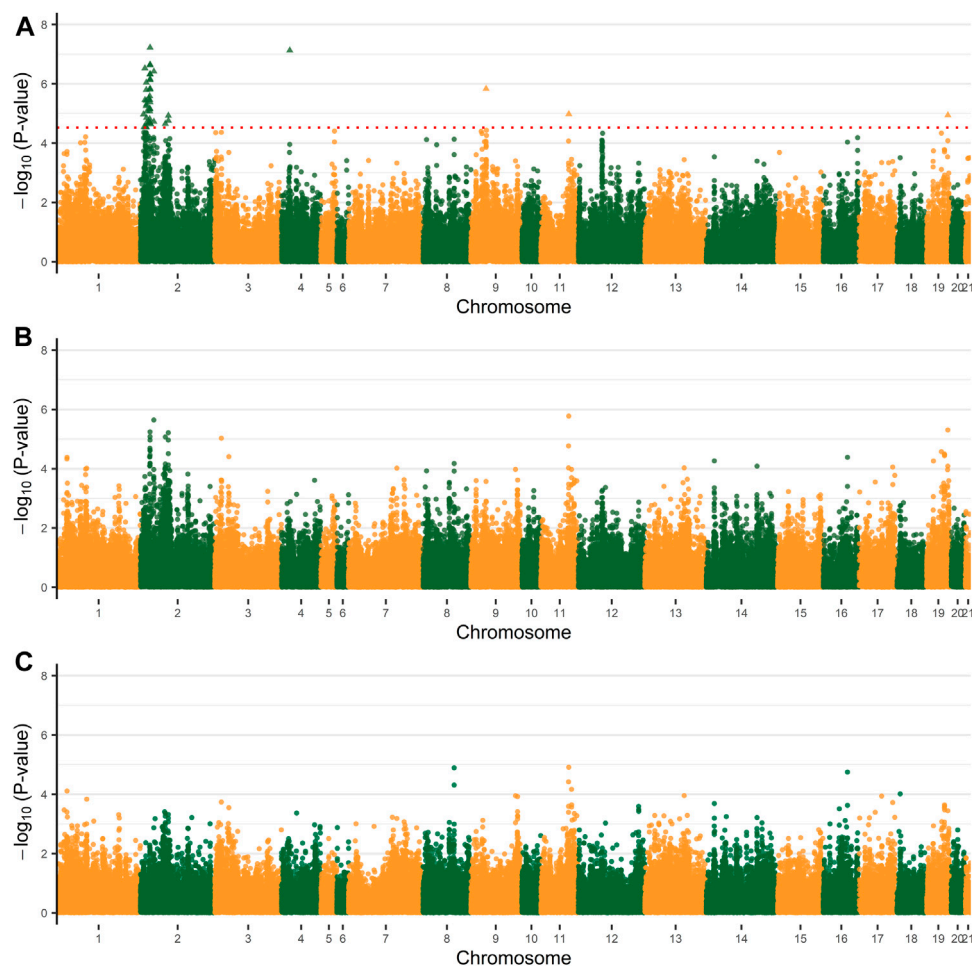


FIGURE 2 | Manhattan plots of genome-wide association analysis (GWAS) for mature weight (A), maturity rate (K). **(A)** The Manhattan plot of both the parameter A and K. **(B)** The Manhattan plot of the parameter A. **(C)** The Manhattan plot of the parameter K. The dashed line of red indicates a 5% FDR-corrected threshold and the significant single-nucleotide polymorphisms (SNPs) are represented by triangles.

Fitted Growth Curves

The growth curves of all individuals were successfully fitted using the four candidate models of logistic (AIC = 35,667.62 and BIC = 35,697.32), Gompertz (AIC = 35,699.36 and BIC = 35,729.06), Brody (AIC = 35,830.33 and BIC = 35,860.03), and Von Bertalanffy (AIC = 37,737.74 and BIC = 37,767.44), whereas the model of Richards did not converge and was therefore excluded for comparison (**Figure 1**; **Supplementary Table S1**). Among the four NMM successfully fitted, the logistic model showed the best goodness of fit with the lowest values of AIC and BIC, and was selected for the following GWAS. The estimates of parameter A and K of the logistic growth curves were 2,615.45 and 0.054, respectively. Furthermore, the growth curves of females and males were separately fitted, which also supported the logistic model having the best goodness of fit and similar growth parameters (**Supplementary Table S1**). There were only small differences for the A and K parameters estimated between males and females.

Association Analyses

Based on the PCA results (**Supplementary Figure S3**), no obvious population stratification was observed in this population studied. The first five PCs explained about 58.9% of total variability, which were included in the NMM as fixed effects with alleviated convergence problems. A total of 45 significant SNPs were revealed to simultaneously affect both parameter A and K, which were distributed among five chromosomes, OCU2, OCU4, OCU9, OCU11, and OCU19 (**Figure 2**; **Table 2**). Among them, the highest numbers of significant SNPs were observed on OCU2 (N = 41), and the three most significant SNPs were located on OCU2 ($p = 5.98\text{E-}08$), OCU4 ($p = 7.51\text{E-}08$), and OCU2 ($p = 2.22\text{E-}07$), respectively. All the 45 significant SNPs were clustered into 24 QTLs, and three of their QTLs (OCU2: 13.67–13.99 Mb, OCU2: 21.86–22.25 Mb, OCU2: 22.45–22.77 Mb) were identified based on three or more SNPs (**Table 2**). When considering the SNP effect separately for either parameter A or K, no significant SNP was identified at the

TABLE 2 | Significant SNPs, QTLs, and candidate genes simultaneously affect both parameter A and K of the logistic growth curve in rabbits.

Chromosomes	SNP position (bp)	p	Locations	QTL region (bp)	Candidate genes
OCU2	7,392,553	1.06E-05	Intron	7,292,553–7,492,553	<i>LDB2</i>
	8,791,893	2.88E-05	Intergenic	8,691,893–8,891,893	ENSOCUG00000035404 ^a
	10,501,993	3.52E-06	Intergenic	10,401,993–10,691,957	None
	10,591,957	2.98E-07	Intergenic		
	11,378,860	2.94E-05	Intron	11,278,860–11,478,860	<i>PACRGL</i> and <i>KCNIP4</i>
	13,386,997	5.63E-06	Intergenic	13,286,997–13,486,997	<i>GBA3</i>
	13,488,329	3.69E-06	Intergenic	13,388,329–13,588,329	ENSOCUG00000031640 ^a
	13,770,022	2.93E-05	Intergenic	13,670,022–13,985,731	ENSOCUG00000031081 ^a and ENSOCUG00000034770 ^a
	13,775,984	9.13E-07	Intergenic		
	13,847,789	1.60E-06	Intergenic		
	13,885,731	2.86E-05	Intergenic		
	14,218,502	2.09E-05	Intergenic	14,118,502–14,318,502	None
	14,381,825	1.39E-05	Intergenic	14,281,825–14,488,318	<i>PPARGC1A</i>
	14,388,318	4.97E-06	Intergenic		
	15,540,704	8.39E-06	Intergenic	15,440,704–15,640,704	<i>CCDC149</i> , <i>LG12</i> , ENSOCUG00000029972 ^a , ENSOCUG00000037097 ^a , and ENSOCUG00000037279 ^a
	19,348,590	2.04E-05	Intergenic	19,248,590–19,448,590	None
	19,519,932	3.07E-06	Intergenic	19,419,932–19,619,932	None
	21,766,582	1.53E-05	Intergenic	21,666,582–21,866,582	None
	21,961,157	2.03E-05	Intergenic	21,861,157–22,251,345	ENSOCUG00000039621 ^a
	21,961,341	2.61E-06	Intergenic		
	21,991,030	8.43E-06	Intergenic		
	22,076,402	7.17E-06	Intergenic		
	22,080,992	2.06E-05	Intergenic		
	22,082,402	6.76E-06	Intergenic		
	22,151,345	3.50E-06	Intergenic		
	22,287,659	1.54E-06	Intergenic	22,187,659–22,387,659	None
	22,552,417	5.98E-08	Intergenic	22,452,417–22,774,072	ENSOCUG00000032798 ^a
	22,559,081	4.80E-07	Intergenic		
	22,559,709	2.22E-07	Intergenic		
	22,559,791	4.18E-06	Intergenic		
	22,585,579	2.37E-07	Intergenic		
	22,597,775	7.40E-07	Intergenic		
	22,628,803	6.92E-07	Intergenic		
	22,632,153	1.45E-06	Intergenic		
	22,674,072	4.67E-07	Intergenic		
	24,405,104	2.30E-05	Intergenic	24,305,104–24,505,104	None
	31,443,212	1.89E-05	Intergenic	31,343,212–31,566,952	<i>ATP8A1</i> and <i>SHISA3</i>
	31,466,952	3.82E-07	3'-UTR		
	58,261,249	2.21E-05	Intergenic	58,161,249–58,361,249	None
	65,306,234	1.18E-05	Intergenic	65,206,234–65,496,063	<i>LOC100358067</i>
	65,396,063	1.74E-05	Intergenic		
OCU4	19,121,968	7.51E-08	Intergenic	19,021,968–19,221,968	ENSOCUG00000038375 ^a
OCU9	35,523,218	1.48E-06	Intron	35,423,218–35,623,218	<i>TAF1</i>
OCU11	64,951,640	1.05E-05	Intergenic	64,851,640–65,051,640	ENSOCUG00000037935 ^a , ENSOCUG00000029125, ENSOCUG00000037904 ^a , and <i>FGF10</i>
OCU19	52,159,278	1.15E-05	Intron	52,059,278–52,259,278	<i>RGS9</i> , ENSOCUG00000036189, <i>GNA13</i> , <i>AMZ2</i> , <i>SLC16A6</i> , and <i>ARSG</i>

^alncRNA; 3'-UTR, 3'-untranslated region; SNP, single-nucleotide polymorphism.

predefined threshold (Figure 2). However, some suggestive associations were also observed (with *p* lower or close to 1.0E-05), such as the SNPs on OCU2, OCU3, OCU11, and OCU19 for parameter A, and OCU8 and OCU11 for parameter K.

Candidate Genes and Functional Analyses

Within the 24 candidate QTLs regarding the significant SNP effects on both parameter A and K, a total of 19 protein-coding and 12 lncRNA positional candidate genes were identified (Table 2). Of these, four [LIM domain-binding 2 (*LDB2*),

potassium voltage-gated channel interacting protein 4 (*KCNIP4*), TAF1 chemokine like family member 1 (*TAF1*), and G protein subunit alpha 13 (*GNA13*)] and one [ATPase phospholipid transporting 8A1 (*ATP8A1*)] candidate genes were found to have the significant SNPs located on intron and 3'-untranslated region (3'-UTR), respectively. Furthermore, three protein-coding genes located on OCU2 were supported by more than one significant SNPs, including the peroxisome proliferator-activated receptor gamma coactivator-1 alpha (*PPARGC1A*), ATPase phospholipid transporting 8A1 (*ATP8A1*), and shisa

family member 3 (*SHISA3*) gene. Two lncRNA genes (*ENSOCUG00000039621* and *ENSOCUG00000032798*) had seven and nine significant SNPs that were located on intergenic regions, respectively. The detailed information of these positional candidate genes is shown in **Supplementary Table S2**.

For these positional candidate genes, 19 biological processes of GO terms were significantly enriched ($p < 0.05$, **Supplementary Table S3**). However, no significant KEGG pathway was found. Four genes of *GNAI3*, *ATP8A1*, *LDB2*, and fibroblast growth factor 10 (*FGF10*) were observed in eight GO terms that were mainly involved in the cell development, such as the biological processes of “regulation of cell migration” and “regulation of cell motility.” Furthermore, both *LDB2* and *FGF10* were enriched in the five GO terms that have the functional implications into growth, such as the biological processes of “somatic stem cell population maintenance” and “maintenance of cell number.”

DISCUSSION

Growth traits have considerable economic implications in meat rabbit industry. For Gabali rabbits in Egypt, the heritability estimates were 0.19, 0.23, 0.16, and 0.14 for BW at 4, 8, 12, and 16 weeks of age (Soliman et al., 2014), which suggested a moderate heritability for these growth traits. The estimated heritability of individual BW ranged from 0.11 at 9 weeks of age to 0.43 at 6 weeks of age in New Zealand White and Dutch breeds of rabbits (Akanno and Ibe, 2005). The moderate to high heritability (from 0.266 to 0.540) was similarly estimated using both Sire Model and Animal Model in New Zealand White rabbits (Dige et al., 2012). Abou Khadiga et al. (2008) conducted the genetic evaluation in crossbred population of Spanish synthetic maternal line V and Egyptian Baladi Black, and found that growth traits were significantly affected by direct genetic effects. Furthermore, the genotype \times environment interaction was also observed for affecting growth performances in growing rabbits (Zeferino et al., 2011). Together, these studies indicated that the improvement of growth traits by genetic selection is much feasible in rabbits. However, the relevant studies in rabbits, such as genomic evaluation and GWAS, have largely lagged behind in comparison with other livestock species (Jonas and Koning, 2015). Therefore, in this study, we performed the association analyses for individual BW at different growth time points using the genome-wide variants. As a relatively limited number of rabbits were included in the present study, however, the increased detection power of GWAS would be expected using larger datasets in future studies.

Like milk production traits in dairy livestock, the individual growth has been preferably described by longitudinal records measured over multiple time points. In practices, the phenotypic records at one or a few time points could be representatively selected and analyzed. However, an alternative approach is to fit the whole growth trajectory using nonlinear regression models and then use the derived model parameters for describing individual growth performance. In an early study (Ptak et al., 1994), three nonlinear models of Von Bertalanffy, Gompertz, and logistic were compared

for fitting the growth in purebred and crossbred rabbits, and found that the Von Bertalanffy gave the best fit. The Gompertz growth curves were fitted and used in analyzing the effect of selection for growth rate on growth curves in rabbits (Blasco et al., 2003). Recently, Ding et al. (2019) fitted the growth curves using the logistic, Gompertz, and Von Bertalanffy models for crossbred population of California rabbit \times New Zealand white rabbit and suggested that the most accurate model was logistic. In this study, the logistic was chosen as the best model to describe the growth trajectory of our crossbred population that was generated by crossing Kangda5 rabbits with Californian rabbits, which was consistent with the results of Ding et al. (2019). Therefore, the selection of the best model to fit the growth curve in rabbits would be breed or population dependent, which should be specifically compared in each study.

In livestock and poultry, mature weight and maturity rate are the two important parameters for describing growth performance; some individuals have higher maturity rate but smaller mature weight, and vice versa. Therefore, to identify genes or causal mutations independently affecting the mature weight and maturity rate is essential for implementing precision improvement of genetic selection. Using the estimated growth parameters as pseudo-phenotypes in GWAS of growth traits in Brahman cattle, a large number of significant SNPs were identified to be associated with mature weight and maturity rate, respectively (Crispim et al., 2015). A similar GWAS was recently reported for growth traits of Chinese Simmental beef cattle, which also revealed different SNPs for the two parameters (Duan et al., 2021). Silva et al. (2017) proposed an alternative method to combine the fitting of growth curves and estimation of SNP effects into one single-step NMM, and to apply to growth traits in pigs with an improved statistical power observed. In this study, we also employed the single-step NMM approach for GWAS of growth traits in a crossbred population of rabbits and found that all the significant SNPs simultaneously affected the two parameters of mature weight and maturity rate. The absence of SNPs independently associated with either mature weight or maturity rate would indicate the specific genetic architecture of growth performance for this studied population. Also, the number of significant SNPs identified in this study was also higher than our former observation (Yang et al., 2020) that was alternatively performed through the separate association analysis with BW at different time points. However, no growth curve parameter was estimated and used as pseudo-phenotype of association analysis by Yang et al. (2020), which would disable the direct comparison.

Around the significant SNPs identified in this study, we found some candidate genes that have the functional implications on growth traits in literature. Among them, the *KCNIP4*, a member of the family of voltage-gated potassium channel-interacting proteins, was found to be located within the QTL between the 21 and 67 cM regions of chromosome 6 that was associated with birth weight in Zandi sheep (Esmailizadeh, 2010; Mohammadi et al., 2020). Pasandideh et al. (2018) also suggested that the *KCNIP4* gene was involved in the regulation of muscle growth and fat deposition in sheep. In chicken, Jin et al. (2015) found that *KCNIP4* was located nearest to a significant SNP associated with the BW of 10 and 14 weeks of age. A nearby gene of *GBA3* (glucosylceramidase beta 3) is related to hydrolyze beta-galactose

and beta-glucose (Dekker et al., 2011), and was further found to be significantly associated with BW traits in sheep (Al-Mamun et al., 2015). In this study, one significant SNP was detected in the intron of *LDB2*, which is a transcriptional regulator (Johnsen et al., 2009) and was found to affect the growth traits of BW and average daily gain in chicken (Gu et al., 2011; Wang et al., 2019) and of BW in Nanjiang Yellow sheep (Guo et al., 2018).

Another important candidate gene is *PPARGC1A*, which was found in pigs to regulate the lipid deposition (Li et al., 2014b), composition of muscle fiber (Lee et al., 2012), and abdominal fat content (Stachowiak et al., 2007). Several SNPs have been identified in *PPARGC1A* gene to be associated with adult BW and average daily gain in Nanyang cattle (Li et al., 2014a), yearling weight in Nelore cattle (Fonseca et al., 2015), and birth weight and calf birth weight in Iranian Holstein cattle (Pasandideh, 2020). Furthermore, Chen et al. (2020) found a 17-bp InDel mutation within the 11th intron of *PPARGC1A* gene in sheep, which was associated with the BW. Both *GNA13* and *SHISA3* could affect individual growth as they are involved in the biological regulation of osteoclastogenesis and bone development (Wu et al., 2017; Murakami et al., 2019). Furthermore, we observed that the *FGF10* gene located on OCU11 was significantly associated with both mature weight and maturity rate. Previous studies revealed that *FGF10* could promote the proliferation and differentiation of adipocyte through the Ras/MAKP pathway (Konishi et al., 2006), and regulate adipogenesis in muscle tissue of goats (Xu et al., 2018) and Tibetan chickens (Zhang et al., 2018). We did not find the relevant publication in literature about functional implications for the 12 candidate lncRNA genes found in this study.

The post-GWAS functional studies are necessary for fine mapping the causal genetic variants and dissecting the underlying biological mechanism (Gallagher and Chen-Plotkin, 2018). Therefore, these candidate genes found in this study could be preferably selected in future studies to investigate their functional mechanisms affecting the individual growth in rabbits. On the other hand, these significant SNPs and genomic regions could be incorporated into the genomic prediction models with an improved accuracy, by using the weighted genomic best linear unbiased prediction (Zhang et al., 2016) or Bayesian (van den Berg et al., 2020) approaches.

CONCLUSION

In the crossbred population of meat rabbits, we employed the nonlinear mixed model to simultaneously fit growth curves and estimate SNP effects at the genome-wide level. The significant

SNPs on five chromosomes (OCU2, OCU4, OCU9, OCU11, and OCU19) were found to simultaneously affect the mature weight and maturity rate, which further revealed some suggestive candidate genes, including the *KCNIP4*, *GBA3*, *PPARGC1A*, *LDB2*, *SHISA3*, *GNA13*, and *FGF10*. These obtained results are useful to increase our knowledge about growth mechanisms in rabbits, and could be used for improving the accuracy of genomic selection in this population.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because all data used in this study was obtained from the previous study.

AUTHOR CONTRIBUTIONS

YL, S-YC, and S-JL conceived, designed, and coordinated this research. YL and ZW performed the data analysis with technical assistance from LG and S-YC. YL wrote the initial version of the manuscript. KZ, CZ, RY, XL, XJ, and S-JL provided all the datasets. All authors interpreted the results and edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was financially supported by National Natural Science Foundation of China (32072684), Earmarked Fund for China Agriculture Research System (Grant No. CARS-44-A-2), and Science and Technology Department of Sichuan Province (2021YFYZ0033).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.750939/full#supplementary-material>

REFERENCES

- Abou Khadiga, G., Saleh, K., Nofal, R., and Baselga, M. (2008). "Genetic Evaluation of Growth Traits in a Crossbreeding experiment Involving Line V and Baladi Black Rabbits in Egypt," in Proceedings of the 9th World Rabbit Congress (WRC), Verona, Italy, June 10–13, 2008.
- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi:10.1109/TAC.1974.1100705
- Akanno, E. C., and Ibe, S. N. (2005). Estimates of Genetic Parameters for Growth Traits of Domestic Rabbits in the Humid Tropics. *Livestock Res. Rural Dev.* 17, 86.
- Al-Mamun, H. A., Kwan, P., Clark, S. A., Ferdosi, M. H., Tellam, R., and Gondro, C. (2015). Genome-wide Association Study of Body Weight in Australian Merino Sheep Reveals an Orthologous Region on OAR6 to Human and Bovine

- Genomic Regions Affecting Height and Weight. *Genet. Sel. Evol.* 47, 66. doi:10.1186/s12711-015-0142-4
- Blasco, A., Piles, M., and Varona, L. (2003). A Bayesian Analysis of the Effect of Selection for Growth Rate on Growth Curves in Rabbits. *Genet. Sel. Evol.* 35, 21–41. doi:10.1186/1297-9686-35-1-21
- Bovo, S., Schiavo, G., Utzeri, V. J., Ribani, A., Schiavitto, M., Buttazzoni, L., et al. (2021). A Genome-wide Association Study for the Number of Teats in European Rabbits (*Oryctolagus cuniculus*) Identifies Several Candidate Genes Affecting This Trait. *Anim. Genet.* 52, 237–243. doi:10.1111/age.13036
- Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103, 338–348. doi:10.1016/j.ajhg.2018.07.015
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Raising to the challenge of Larger and Richer Datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8
- Chen, P., Zhao, H., Wu, M., He, S., Yuan, T., Yi, X., et al. (2020). A Novel 17 Bp InDel Polymorphism within the *PPARGC1A* Gene Is Significantly Associated with Growth Traits in Sheep. *Anim. Biotechnol.* 8, 1–9. doi:10.1080/10495398.2020.1796697
- Chen, S., Jia, X., and Lai, S. (2021). Evaluation of the Accuracy of Genome Selection in Rabbits. *Chin. J. Rabbit Farming* 1, 8–11. (in Chinese): CNKI:SUN:ZGYT.0.2021-01-003.
- Crispim, A. C., Kelly, M. J., Guimarães, S. E. F., e Silva, F. F., Fortes, M. R. S., Wenceslau, R. R., et al. (2015). Multi-trait GWAS and New Candidate Genes Annotation for Growth Curve Parameters in Brahman Cattle. *PLoS One* 10, e0139906. doi:10.1371/journal.pone.0139906
- Dekker, N., Voorn-Brouwer, T., Verhoek, M., Wennekes, T., Narayan, R. S., Speijer, D., et al. (2011). The Cytosolic β -glucosidase GBA3 Does Not Influence Type 1 Gaucher Disease Manifestation. *Blood Cell Mol. Dis.* 46, 19–26. doi:10.1016/j.bcmd.2010.07.009
- Dige, M. S., Kumar, A., Kumar, P., Dubey, P. P., and Bhushan, B. (2012). Estimation of Variance Components and Genetic Parameters for Growth Traits in New Zealand white Rabbit (*Oryctolagus cuniculus*). *J. Appl. Anim. Res.* 40, 167–172. doi:10.1080/09712119.2011.645037
- Ding, P., Jia, X., Chen, S., Gu, S., Hu, S., Wang, J., et al. (2019). Comparison and Analysis of California Rabbit \times New Zealand white Rabbit (F1), F1 Random Mating Population (F2) Growth Curve Model Fitting. *Chin. J. Rabbit Farming* 6, 7–11. (in Chinese): CNKI:SUN:ZGYT.0.2019-06-003.
- Duan, X., An, B., Du, L., Chang, T., Liang, M., Yang, B.-G., et al. (2021). Genome-wide Association Analysis of Growth Curve Parameters in Chinese Simmental Beef Cattle. *Animals* 11, 192. doi:10.3390/ani11010192
- Esmailzadeh, A. K. (2010). A Partial Genome Scan to Identify Quantitative Trait Loci Affecting Birthweight in Kermani Sheep. *Small Ruminant Res.* 94, 73–78. doi:10.1016/j.smallrumres.2010.07.003
- Fonseca, P. D. D. S., de Souza, F. R. P., de Camargo, G. M. F., Gil, F. M. M., Cardoso, D. F., Zetouni, L., et al. (2015). Association of ADIPOQ, OLR1 and PPARGC1A Gene Polymorphisms with Growth and Carcass Traits in Nelore Cattle. *Meta Gene* 4, 1–7. doi:10.1016/j.mgene.2015.02.001
- Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS Era: from Association to Function. *Am. J. Hum. Genet.* 102, 717–730. doi:10.1016/j.ajhg.2018.04.002
- Gu, X., Feng, C., Ma, L., Song, C., Wang, Y., Da, Y., et al. (2011). Genome-wide Association Study of Body Weight in Chicken F2 Resource Population. *PLoS One* 6, e21872. doi:10.1371/journal.pone.0021872
- Guo, J., Tao, H., Li, P., Li, L., Zhong, T., Wang, L., et al. (2018). Whole-genome Sequencing Reveals Selection Signatures Associated with Important Traits in Six Goat Breeds. *Sci. Rep.* 8, 10405. doi:10.1038/s41598-018-28719-w
- Heisterkamp, S. H., Willigen, E. v., Diderichsen, P.-M., and Maringa, J. (2017). Update of the Nlme Package to Allow a Fixed Standard Deviation of the Residual Error. *R Journal* 9, 239–251. doi:10.32614/RJ-2017-010
- Helal, M., Hany, N., Maged, N., Abdelaziz, M., Osama, N., Younan, Y. W., et al. (2021). Candidate Genes for Marker-Assisted Selection for Growth, Carcass and Meat Quality Traits in Rabbits. *Anim. Biotechnol.* 4, 1–20. doi:10.1080/10495398.2021.1908315
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211
- Jin, C. F., Chen, Y. J., Yang, Z. Q., Shi, K., and Chen, C. K. (2015). A Genome-wide Association Study of Growth Trait-Related Single Nucleotide Polymorphisms in Chinese Yancheng Chickens. *Genet. Mol. Res.* 14, 15783–15792. doi:10.4238/2015.December.1.30
- Johnsen, S. A., Güngör, C., Prenzel, T., Riethdorf, S., Riethdorf, L., Taniguchi-Ishigaki, N., et al. (2009). Regulation of Estrogen-dependent Transcription by the LIM Cofactors CLIM and RLIM in Breast Cancer. *Cancer Res.* 69, 128–136. doi:10.1158/0008-5472.CAN-08-1630
- Jonas, E., and Koning, D.-J. d. (2015). Genomic Selection Needs to Be Carefully Assessed to Meet Specific Requirements in Livestock Breeding Programs. *Front. Genet.* 6, 49. doi:10.3389/fgene.2015.00049
- Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New Approach for Understanding Genome Variations in KEGG. *Nucleic Acids Res.* 47, D590–D595. doi:10.1093/nar/gky962
- Konishi, M., Asaki, T., Koike, N., Miwa, H., Miyake, A., and Itoh, N. (2006). Role of Fgf10 in Cell Proliferation in white Adipose Tissue. *Mol. Cell Endocrinol.* 249, 71–77. doi:10.1016/j.mce.2006.01.010
- Koya, P. R., and Goshu, A. T. (2013). Generalized Mathematical Model for Biological Growths. *Open Journal of Modelling and Simulation* 01, 42–53. doi:10.4236/ojmsi.2013.14008
- Lee, J.-S., Kim, J.-M., Hong, J.-S., Lim, K.-S., Hong, K.-C., and Lee, Y. S. (2012). Effects of Polymorphisms in the 3' Untranslated Region of the Porcine PPARGC1A Gene on Muscle Fiber Characteristics and Meat Quality Traits. *Mol. Biol. Rep.* 39, 3943–3950. doi:10.1007/s11033-011-1173-8
- Li, Q., Wang, Z., Zhang, B., Lu, Y., Yang, Y., Ban, D., et al. (2014a). Single Nucleotide Polymorphism Scanning and Expression of the Pig PPARGC1A Gene in Different Breeds. *Lipids* 49, 1047–1055. doi:10.1007/s11745-014-3928-1
- Li, M., Liu, M., Liu, D., Lan, X., Lei, C., and Chen, H. (2014b). The Novel Coding Region SNPs of PPARGC1A Gene and Their Associations with Growth Traits in Chinese Native Cattle. *Mol. Biol. Rep.* 41, 39–44. doi:10.1007/s11033-013-2835-5
- Li, S., Zeng, W., Li, R., Hoffman, L. C., He, Z., Sun, Q., et al. (2018). Rabbit Meat Production and Processing in China. *Meat Sci.* 145, 320–328. doi:10.1016/j.meatsci.2018.06.037
- Mancin, E., Sosa-Madrid, B. S., Blasco, A., and Ibáñez-Escriche, N. (2021). Genotype Imputation to Improve the Cost-Efficiency of Genomic Selection in Rabbits. *Animals* 11, 803. doi:10.3390/ani11030803
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110
- Mohammadi, H., Rafat, S. A., Shahrabak, H. M., Shodja, J., and Moradi, M. H. (2020). Genome-wide Association Study and Gene Ontology for Growth and Wool Characteristics in Zandi Sheep. *J. Livestock Sci. Technol.* 8, 45–55. doi:10.22103/JLST.2020.15795.1317
- Murakami, K., Zhifeng, H., Suzuki, T., Kobayashi, Y., and Nakamura, Y. (2019). The *Shisa3* Knockout Mouse Exhibits normal Bone Phenotype. *J. Bone Miner. Metab.* 37, 967–975. doi:10.1007/s00774-019-01014-y
- Pasandideh, M., Rahimi-Mianji, G., and Gholizadeh, M. (2018). A Genome Scan for Quantitative Trait Loci Affecting Average Daily Gain and Kleiber Ratio in Baluchi Sheep. *J. Genet.* 97, 493–503. doi:10.1007/s12041-018-0941-9
- Pasandideh, M. (2020). Two SNPs in the Bovine PPARGC1A Gene Are Associated with the Birth Weight of Holstein Calves. *Meta Gene* 25, 100732. doi:10.1016/j.mgene.2020.100732
- Ptak, E., Bieniek, J., and Jagusiak, W. (1994). “Comparison of Growth Curves of Purebred and Crossbred Rabbits,” in Proceedings of the 5th World Congress of Genetic Applied to Livestock Production (WCGALP), Guelph, Canada, August 7–12, 1994.
- Qin, Y. H. (2019). Rabbit Breeding Technology and Seed Industry Development. *Feed and Husbandry* 2, 53–57. (in Chinese): CNKI:SUN:SNCM.0.2019-02-012.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* 6, 461–464. doi:10.1214/aos/1176344136
- Silva, F. F. E., Zambrano, M. F. B., Varona, L., Glória, L. S., Lopes, P. S., Silva, M. V. G. B., et al. (2017). Genome Association Study through Nonlinear Mixed Models Revealed New Candidate Genes for Pig Growth Curves. *Sci. Agric. (Piracicaba, Braz.* 74, 1–7. doi:10.1590/1678-992x-2016-0023

- Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart Community portal: an Innovative Alternative to Large, Centralized Data Repositories. *Nucleic Acids Res.* 43, W589–W598. doi:10.1093/nar/gkv350
- Soliman, H., Gad, S., and Esmail, I. (2014). Genetic Parameters for Post-weaning Growth Traits of Gabali Rabbits in Egypt. *Egyptian Poultry Science Journal* 34, 655–664. doi:10.21608/epsj.2014.32587
- Sosa-Madrid, B. S., Santacreu, M. A., Blasco, A., Fontanesi, L., Pena, R. N., and Ibáñez-Escriche, N. (2020). A Genomewide Association Study in Divergently Selected Lines in Rabbits Reveals Novel Genomic Regions Associated with Litter Size Traits. *J. Anim. Breed. Genet.* 137, 123–138. doi:10.1111/jbg.12451
- Stachowiak, M., Szydlowski, M., Cieslak, J., and Switonski, M. (2007). SNPs in the Porcine *PPARGC1a* Gene: Interbreed Differences and Their Phenotypic Effects. *Cell. Mol. Biol. Lett.* 12, 231–239. doi:10.2478/s11658-006-0066-7
- Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong Control, Conservative point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: a Unified Approach. *J. R. Stat. Soc. B* 66, 187–205. doi:10.1111/j.1467-9868.2004.00439.x
- The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 Years and Still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi:10.1093/nar/gky1055
- van den Berg, I., MacLeod, I. M., Reich, C. M., Breen, E. J., and Pryce, J. E. (2020). Optimizing Genomic Prediction for Australian Red Dairy Cattle. *J. Dairy Sci.* 103, 6276–6298. doi:10.3168/jds.2019-17914
- Wang, W. H., Wang, J. Y., Zhang, T., Wang, Y., Zhang, Y., and Han, K. (2019). Genome-wide Association Study of Growth Traits in Jinghai Yellow Chicken Hens Using SLAF-seq Technology. *Anim. Genet.* 50, 175–176. doi:10.1111/age.12346
- Wu, M., Chen, W., Lu, Y., Zhu, G., Hao, L., and Li, Y.-P. (2017). Gα13 Negatively Controls Osteoclastogenesis through Inhibition of the Akt-Gsk3β-NFATc1 Signalling Pathway. *Nat. Commun.* 8, 13700. doi:10.1038/ncomms13700
- Xu, Q., Lin, S., Wang, Y., Zhu, J., and Lin, Y. (2018). Fibroblast Growth Factor 10 (FGF10) Promotes the Adipogenesis of Intramuscular Preadipocytes in Goat. *Mol. Biol. Rep.* 45, 1881–1888. doi:10.1007/s11033-018-4334-1
- Yang, X., Deng, F., Wu, Z., Chen, S.-Y., Shi, Y., Jia, X., et al. (2020). A Genome-wide Association Study Identifying Genetic Variants Associated with Growth, Carcass and Meat Quality Traits in Rabbits. *Animals* 10, 1068. doi:10.3390/ani10061068
- Zeferino, C. P., Moura, A. S. A. M. T., Fernandes, S., Kanayama, J. S., Scapinello, C., and Sartori, J. R. (2011). Genetic Group × ambient Temperature Interaction Effects on Physiological Responses and Growth Performance of Rabbits. *Livestock Sci.* 140, 177–183. doi:10.1016/j.livsci.2011.03.027
- Zhang, R., Li, R., Feng, Q., Zhi, L., Li, Z., Xu, Y.-O., et al. (2018). Expression Profiles and Associations of *FGF1* and *FGF10* with Intramuscular Fat in Tibetan Chicken. *Br. Poult. Sci.* 59, 613–617. doi:10.1080/00071668.2018.1507018
- Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., and Misztal, I. (2016). Weighting Strategies for Single-step Genomic BLUP: an Iterative Approach for Accurate Calculation of GEBV and GWAS. *Front. Genet.* 7, 151. doi:10.3389/fgene.2016.00151

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Liao, Wang, Glória, Zhang, Zhang, Yang, Luo, Jia, Lai and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Single-cell RNA Sequencing Reveals Heterogeneity of Cultured Bovine Satellite Cells

Pengcheng Lyu¹, Yumin Qi², Zhijian J. Tu² and Honglin Jiang^{1*}

¹Department of Animal and Poultry Sciences, Virginia Tech, Blacksburg, VA, United States, ²Department of Biochemistry, Virginia Tech, Blacksburg, VA, United States

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Guillermo Giovambattista,
CONICET Institute of Veterinary
Genetics (IGEVE), Argentina
Tara G McDanel,
U.S. Meat Animal Research Center,
United States

*Correspondence:

Honglin Jiang
hojiang@vt.edu

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 15 July 2021

Accepted: 11 October 2021

Published: 28 October 2021

Citation:

Lyu P, Qi Y, Tu ZJ and Jiang H (2021)
Single-cell RNA Sequencing Reveals
Heterogeneity of Cultured Bovine
Satellite Cells.
Front. Genet. 12:742077.
doi: 10.3389/fgene.2021.742077

Skeletal muscle from meat-producing livestock such as cattle is a major source of food for humans. To improve skeletal muscle growth efficiency or quality in cattle, it is necessary to understand the genetic and physiological mechanisms that govern skeletal muscle composition, development, and growth. Satellite cells are the myogenic progenitor cells in postnatal skeletal muscle. In this study we analyzed the composition of bovine satellite cells with single-cell RNA sequencing (scRNA-seq). We isolated satellite cells from a 2-week-old male calf, cultured them in growth medium for a week, and performed scRNA-seq using the 10x Genomics platform. Deep sequencing of two scRNA-seq libraries constructed from cultured bovine satellite cells yielded 860 million reads. Cell calling analyses revealed that these reads were sequenced from 19,096 individual cells. Clustering analyses indicated that these reads represented 15 cell clusters that differed in gene expression profile. Based on the enriched expression of markers of satellite cells (PAX7 and PAX3), markers of myoblasts (MYOD1, MYF5), and markers of differentiated myoblasts or myocytes (MYOG), three clusters were determined to be satellite cells, two clusters myoblasts, and two clusters myocytes. Gene ontology and trajectory inference analyses indicated that cells in these myogenic clusters differed in proliferation rate and differentiation stage. Two of the remaining clusters were enriched with PDGFRA, a marker of fibro-adipogenic (FAP) cells, the progenitor cells for intramuscular fat, and are therefore considered to be FAP cells. Gene ontology analyses indicated active lipogenesis in one of these two clusters. The identity of the remaining six clusters could not be defined. Overall, the results of this study support the hypothesis that bovine satellite cells are composed of subpopulations that differ in transcriptional and myogenic state. The results of this study also support the hypothesis that intramuscular fat in cattle originates from fibro-adipogenic cells.

Keywords: FAP, fibro-adipogenic progenitors, ScRNA-seq, cattle, skeletal muscle, myoblast

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a relatively new technology that analyzes transcriptomes in individual cells by deep sequencing (Tang et al., 2009). scRNA-seq has proved to be a powerful method for assessing the heterogeneity of a cell population and for identifying rare or previously uncharacterized cell types in complex organs and tissues (Tang et al., 2009; Andrews and Hemberg, 2018).

Satellite cells are myogenic progenitor cells in skeletal muscle of postnatal animals (Mauro, 1961). Satellite cells are characterized by the expression of the transcription factor paired box protein 7

(PAX7) or PAX 3 (Maroto et al., 1997; Seale et al., 2000; Kuang and Rudnicki, 2008). Satellite cells play an essential role in skeletal muscle regeneration and growth (Dhawan and Rando, 2005). Satellite cells are normally quiescent but are activated under conditions such as muscle injury (Dhawan and Rando, 2005). Once activated, satellite cells proliferate as myoblasts, differentiate, and fuse with each other to generate muscle fibers or with existing muscle fibers to increase muscle size.

In our research involving bovine satellite cells (Ge et al., 2013; Leng et al., 2019; Leng and Jiang, 2019), we noticed that these cells in culture did not all become myofibers upon induction of myogenic differentiation. We hypothesized that bovine satellite cells are composed of subpopulations that differ in myogenic potential. The objective of this study was therefore to determine if bovine satellite cells are heterogeneous in terms of myogenic potential. We approached this objective by analyzing a population of bovine satellite cells with scRNA-seq.

MATERIALS AND METHODS

Isolation and Culture of Bovine Satellite Cells

Bovine satellite cells were isolated from the longissimus muscle of a 2-week-old Holstein bull calf following euthanasia. Satellite cells were isolated using a procedure involving pronase digestion and differential centrifugation as described before (Hathaway et al., 1991). Satellite cells were cultured in growth medium consisting of Dulbecco's modified eagle medium, 10% fetal bovine serum, 2 mM L-glutamine, and 1% antibiotics-antimycotics (Thermo Fisher Scientific, Waltham, MA, United States) for a week to remove dead cells and increase the number of viable cells prior to scRNA-seq. The animal-related procedure was approved by the Virginia Tech Institutional Animal Care and Use Committee.

scRNA-Seq Library Construction

Cells cultured above were detached from the culture plate with trypsin-EDTA (0.25%) and washed with resuspension buffer (phosphate-buffered saline, 0.04% bovine serum albumin). Cells were filtered through a 40 µm strainer to remove cell clumps. Cell viability was determined to be 90% by trypan blue staining and hemocytometer counting. Two scRNA-seq libraries were constructed from satellite cells to increase the number of cells sequenced. The scRNA-seq libraries were constructed using the Chromium Single Cell 3' Gel Bead-in-Emulsion (GEM), Library & Gel Bead Kit v3 (10x Genomics, Pleasanton, CA, United States). We adjusted cell concentration to 1,000 cells/µl with resuspension buffer, and loaded 12.8 µl of diluted cell suspension with a master mix of reverse transcription (RT) reagent, template switch oligo, and RT enzyme into a Chromium Chip B (10x Genomics). Single-cell GEM generation, barcoding, and reverse transcription were achieved by running the Chromium Chip B on the Chromium Controller (10x Genomics). Specifically, single cells, RT reagents, gel beads containing barcoded oligonucleotides, and oil were combined on a microfluidic chip to form reaction nanovesicles. Within each reaction nanovesicle, a single cell was lysed, the gel bead was dissolved to free the identically barcoded RT oligonucleotides, and polyadenylated mRNA was then reverse transcribed into cDNA. Thus,

all cDNAs from the same cell would have the same barcode, which would allow the sequencing reads to be traced back to their single cells of origin. cDNA was amplified for 11 cycles. Based on an Agilent High Sensitivity TapeStation analysis, both scRNA-seq libraries contained a single peak of DNA between 300 and 700 bp, with an average fragment size of 440 bp.

scRNA-Seq Sequencing and Data Analysis

Each scRNA-seq library was pair-end sequenced in a single cell flow lane on an Illumina HiSeq system at the Novogene-UC Davis Sequencing Center (Novogene, Sacramento, CA, United States). Sequencing reads were processed and analyzed using the 10x Genomics Cell Ranger 3.0.2 software, which was composed of different analysis pipelines (Zheng et al., 2017). Specifically, sequencing reads were de-multiplexed using the cellranger mkfastq pipeline and aligned to the bovine genome and transcriptome (Bos_taurus.ARS-UCD1.2) at default parameters using the cellranger count pipeline. Uniquely mapped sequences from each library were used for unique molecular identifiers (UMI) counting using the cellranger count pipeline. The output files from the cellranger counting of reads from two scRNA-seq libraries were combined using the cellranger aggr pipeline.

Cells were initially clustered by expression similarity using the graph-based clustering algorithm, which consisted of building a sparse nearest-neighbor graph followed by Louvain Modularity Optimization, an algorithm that sought to find highly-connected modules in the graph (Blondel et al., 2008). Cells were further clustered by an additional cluster-merging step, which included hierarchical clustering on the cluster-medoids in principal components analysis (PCA) space and merging pairs of sibling clusters if there were no genes differentially expressed between them (with B-H adjusted *p*-value below 0.05). The hierarchical clustering and merging was repeated until there were no more cluster-pairs to merge. Differential gene expression between cell clusters was identified using the quick and simple method sSeq and edgeR (Robinson and Smyth, 2007; Yu et al., 2013). All these analysis pipelines and algorithms were part of the Cell Ranger software and run using parameters recommended by 10X Genomics.

Cell clusters were annotated based on significantly enriched expression (with B-H adjusted *p*-value below 0.05) of marker genes of cell types. Cell clusters and gene expression data were visualized in the Loupe Cell Browser (10x Genomics). Gene ontology analysis was performed with the PANTHER classification system (Mi et al., 2013). Gene ontology terms with corrected *p*-values less than 0.05 were considered significantly enriched.

Trajectory inference was performed using the Monocle algorithm (version 2.20.0) (Trapnell et al., 2014; Qiu et al., 2017). Briefly, the output files of the cellranger count pipeline were read in to generate the count matrix using the Read10X function of Seurat 4 (Hao et al., 2021). A Seurat object was then built from the count matrix using the CreateSeuratObject function of Seurat. The Seurat object was converted to a CellDataSet modeled with the negative binomial distribution using the newimport function of Monocle. Genes with mean expression levels >0.1 were selected to define the state of cells. Following a dimensionality reduction using the reduceDimension

TABLE 1 | Summary of scRNA-seq sequencing, mapping, and analysis.

Item	Library 1	Library 2	Aggregation
Number of reads	431,539,783	429,256,266	860,796,049
Valid barcodes	97.6%	97.4%	97.5%
Reads mapped uniquely to genome (bos_taurus.ars-ucd1.2)	85.5%	85.2%	85.4%
Reads mapped uniquely to transcriptome (bos_taurus.ars-ucd1.2)	50.8%	51.5%	51.1%
Fraction reads in cells	90.8%	89.1%	90.0%
Estimated number of cells	9,939	9,157	19,096
Mean reads per cell	43,418	46,877	43,528
Median genes per cell	3,559	3,764	3,609
Total genes detected	16,227	16,356	16,816

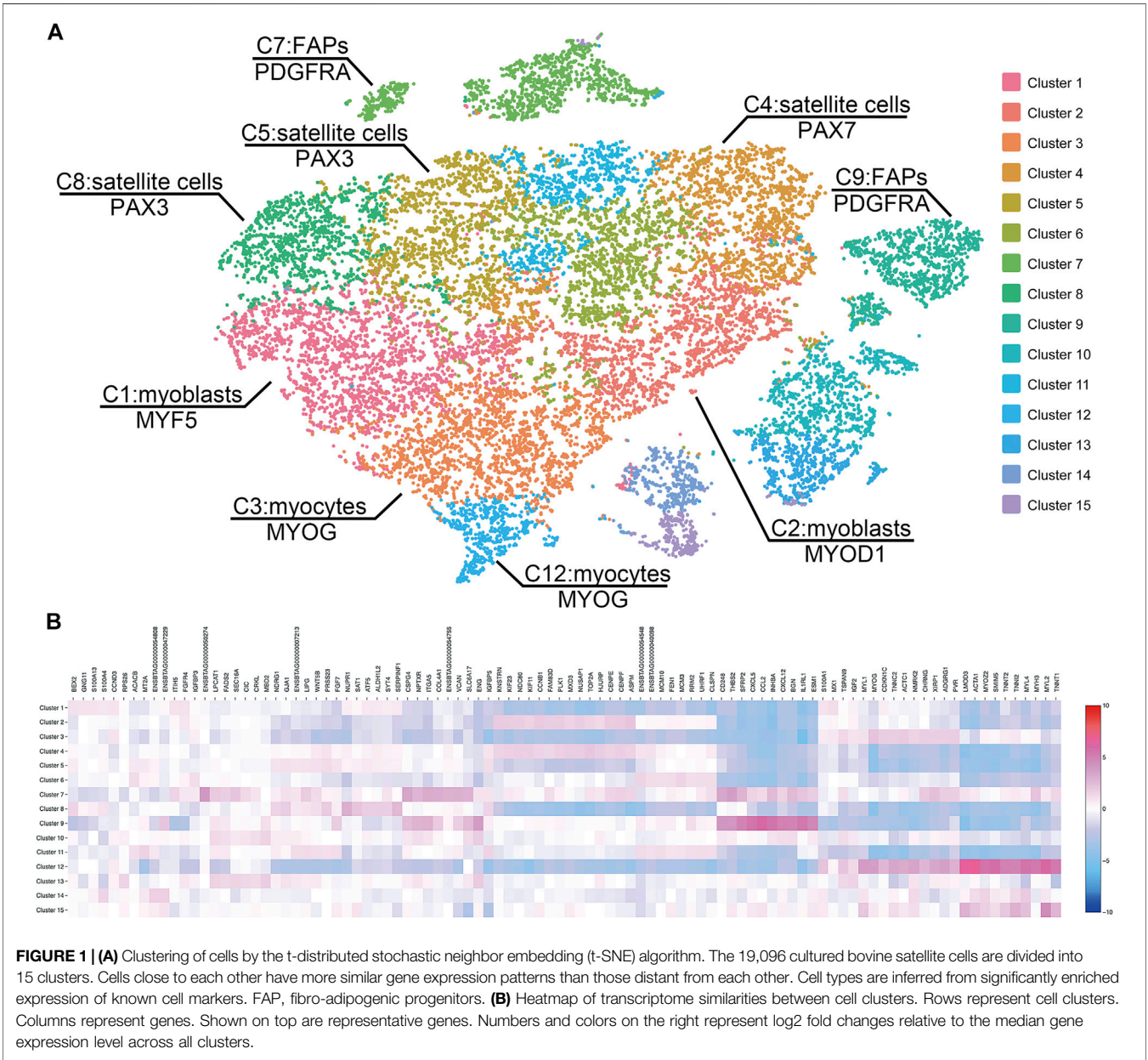


FIGURE 1 | (A) Clustering of cells by the t-distributed stochastic neighbor embedding (t-SNE) algorithm. The 19,096 cultured bovine satellite cells are divided into 15 clusters. Cells close to each other have more similar gene expression patterns than those distant from each other. Cell types are inferred from significantly enriched expression of known cell markers. FAP, fibro-adipogenic progenitors. **(B)** Heatmap of transcriptome similarities between cell clusters. Rows represent cell clusters. Columns represent genes. Shown on top are representative genes. Numbers and colors on the right represent log2 fold changes relative to the median gene expression level across all clusters.

TABLE 2 | Numbers and percentages of clustered cells.

	Number of cells	% Total cells
Cluster 1	2,493	13.1
Cluster 2	2,119	11.1
Cluster 3	2,034	10.7
Cluster 4	1,768	9.3
Cluster 5	1,641	8.6
Cluster 6	1,575	8.2
Cluster 7	1,298	6.8
Cluster 8	1,241	6.5
Cluster 9	1,080	5.7
Cluster 10	1,010	5.3
Cluster 11	845	4.4
Cluster 12	627	3.3
Cluster 13	602	3.2
Cluster 14	446	2.3
Cluster 15	317	1.7

function, cells were ordered along the trajectory using the orderCells function of Monocle.

Scripts and codes used to run the Cell Ranger, Seurat, and Monocle programs in this study were adopted from the manuals and websites for these programs and are listed in the **Supplementary File S1**.

RESULTS AND DISCUSSION

Cultured Bovine Satellite Cells Are Heterogenous in Gene Expression

Sequencing of the two scRNA-seq libraries of bovine satellite cells in culture generated approximately 430 million reads per library. More than 85 and 50% of these reads were uniquely mapped to the bovine genome and transcriptome (bos_taurus.ars-ucd1.2), respectively (**Table 1**). Based on the Cell Ranger analyses, these reads represented the transcriptomes in more than 9,000 cells for each library (**Table 1**). The mean number of reads detected per cell was more than 43,000 for both libraries, and the median number of genes detected per cell was over 3,500 for both libraries (**Table 1**). These numbers met or exceeded the minimum requirements for a quality scRNA-seq analysis (Handley et al., 2015; Haque et al., 2017; Ziegenhain et al., 2017). Because the two scRNA-seq libraries were prepared from the same cells, sequencing data from the two libraries were combined to increase the total number of cells analyzed. Clustering the transcriptomes of 19,096 cells combined from the two scRNA-libraries revealed 15 cell clusters that differed in gene expression pattern (**Figure 1**, **Supplementary File S2**). These 15 clusters contained 300–2,500 cells, or 2–13% total cells analyzed, with cluster 1 being the largest cluster and cluster 15 being the smallest cluster (**Table 2**).

Cultured Bovine Satellite Cells Contain Subpopulations That Differ in Myogenic Stage and Proliferation Rate

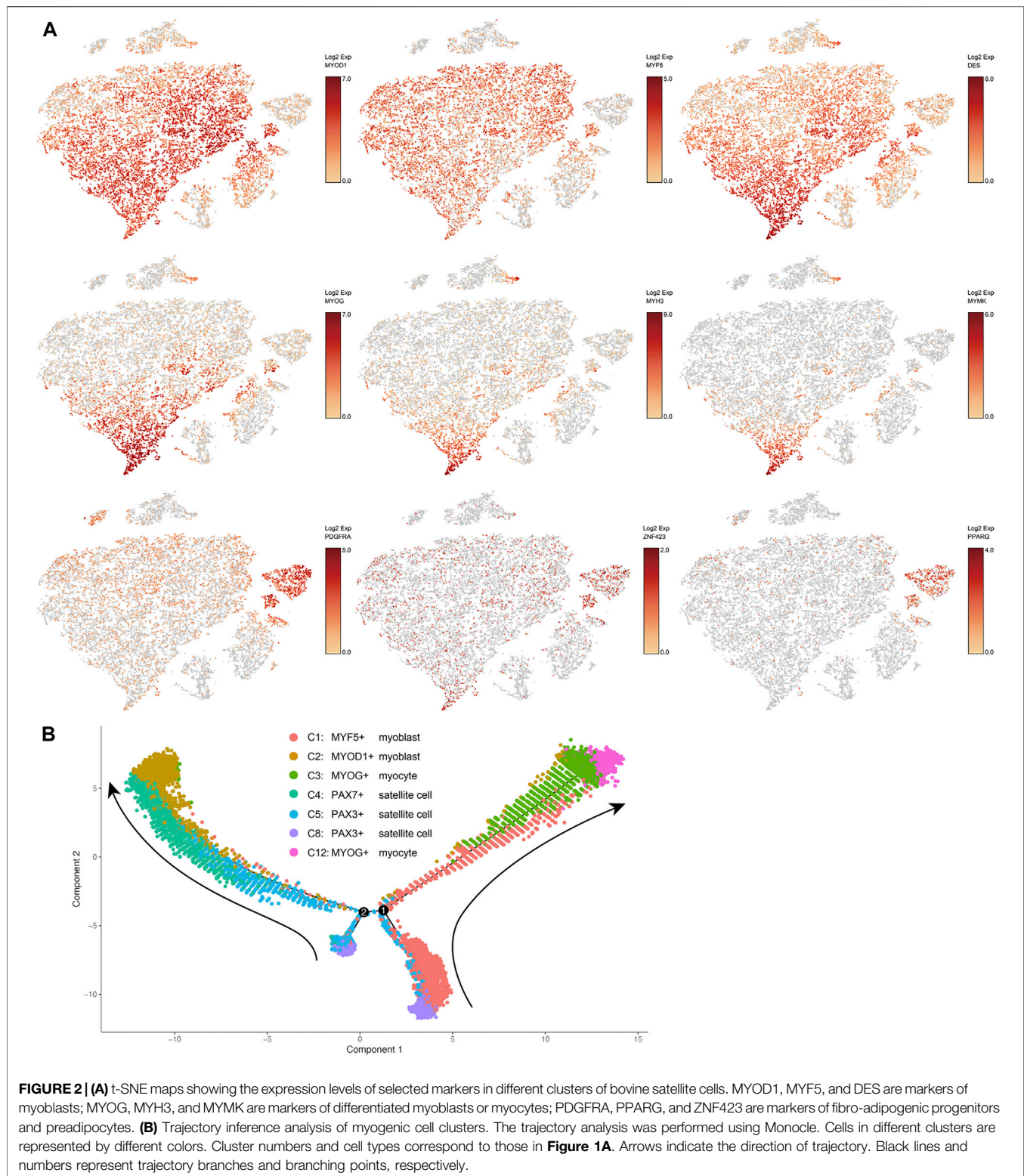
Based on the expression levels of marker genes (MYOD1, MYF5, and DES) of myoblasts (Pownall et al., 2002), clusters 1, 2, 3, and

12 were characterized as subsets of myoblasts, which are activated and proliferating satellite cells (**Figures 1A, 2A**). Expression of MYF5 mRNA in cluster 1, expression of MYOD1 mRNA in cluster 2, and expression of both MYF5 and MYOD1 mRNAs in cluster 3 were significantly enriched (**Supplementary File S2**, **Figure 2A**). PAX7 and PAX3 are markers of satellite cells (Maroto et al., 1997; Seale et al., 2000). Expression of PAX7 mRNA was significantly enriched in cluster 4, and PAX3 mRNA expression was significantly enriched in clusters 5 and 8 (**Supplementary File S2**). Because they were enriched with PAX7 or PAX3, clusters 4, 5, and 8 were determined to be subsets of satellite cells (**Figure 1A**).

MYOG is a master transcriptional regulator of myoblast differentiation (Pownall et al., 2002). MYOG mRNA expression was significantly enriched in clusters 3 and 12 (**Supplementary File S2**, **Figure 2A**). Besides MYOG, many muscle-specific genes such as MB, MYH3, MYL1, NEB, and STAC3 and several myoblast differentiation and fusion regulatory genes such as MEF2A, MEF2D and MYMK (Millay et al., 2013; Estrella et al., 2015) were upregulated in clusters 3 and 12 (**Supplementary File S2**, **Figure 2A**). These two clusters clearly contained differentiating or differentiated myoblasts, i.e., myocytes. Between clusters 3 and 12, more muscle-specific genes were upregulated in cluster 12 than in cluster 3, and the same muscle-specific genes were expressed at greater levels in cluster 12 than in cluster 3 (**Supplementary File S2**, **Figure 2A**). These differences suggest that myoblasts in cluster 12 were more terminally differentiated than those in cluster 3.

Gene ontology analyses of genes upregulated in each cluster indicated that cells in different clusters differed in function. For example, gene ontology analyses of genes upregulated in clusters 2 and 4 indicated that many of these genes were involved in the biological processes, cellular components, and molecular function related to DNA synthesis and cell cycle (**Supplementary File S3**), suggesting that cells in these clusters were undergoing active proliferation. Gene ontology analyses of genes upregulated in clusters 3 and 12 indicated that many of these genes were involved in the biological processes, cellular components, and molecular function related to mature skeletal muscle structure and contraction (**Supplementary File S3**), suggesting that cells in these two clusters were differentiating into functional muscle cells.

A trajectory inference analysis using the Monocle program (Trapnell et al., 2014) revealed the potential lineage relationships between the seven myogenic cell clusters, i.e., clusters 1–5, 8, and 12 (**Figure 2B**). This analysis suggested two trajectories along which the myogenic progenitors PAX3+ satellite cells transitioned toward myogenic differentiation. On one trajectory, PAX3+ satellite cells in cluster 8 committed to MYF5+ myoblasts in cluster 1, and these MYF5+ myoblasts then differentiated into MYOG+ myocytes in cluster 3, which then differentiated further into MYOG+ myocytes in cluster 12 (**Figure 2B**). On the other trajectory, PAX3+ satellite cells in cluster 8 first transitioned to a population of satellite cells in cluster 5 that had a different gene expression pattern from cells in cluster 8 but were still PAX3



positive; PAX3+ satellite cells in cluster 5 then became PAX7+ satellite cells in cluster 4, which were subsequently activated to become MYOD1+ myoblasts in cluster 2 (**Figure 2B**). Overall,

this trajectory analysis further supports the conclusion earlier that cultured bovine satellite cells are composed of subsets of myogenic cells that differ in transcriptional and myogenic state.

Cultured Bovine Satellite Cells Contain Subpopulations That May Be Intramuscular Preadipocytes

Platelet derived growth factor receptor alpha (PDGFRA) is established as a marker of the progenitor cells for intramuscular adipocytes, i.e., intramuscular fibro-adipogenic progenitor (FAP) cells, in mice and humans (Uezumi et al., 2010; Uezumi et al., 2014; Joe et al., 2010). The expression of PDGFRA mRNA was significantly higher in clusters 7 and 9 than in other clusters (Supplementary File S2 and Figure 2A). It is interesting to note that cells in cluster 9 were also enriched with peroxisome proliferator activated receptor gamma (PPARG) and zinc finger protein 423 (ZNF423) mRNAs, two transcriptional regulators of early adipogenesis (Tontonoz and Spiegelman, 2008; Gupta et al., 2010; Leferova et al., 2014). Gene ontology analyses of genes upregulated in clusters 7 and 9 indicated active lipogenesis in cluster 9 (Supplementary File S3). Cells in neither cluster 7 nor cluster 9 expressed markers of mature adipocytes such as leptin (LEP) and adiponectin (ADIPOQ) (Supplementary File S2). These data together indicated that clusters 7 and 9 were FAPs, or intramuscular preadipocytes, with cluster 9 appearing to be more developed preadipocytes than cluster 7. Because cells in clusters 7 and 9 were not enriched with MYOD1, MYF5, DES, or MYOG, markers of myogenic cells, it remains to be determined if these FAP cells were derived from satellite cells during culture or accidentally co-isolated with satellite cells from skeletal muscle.

Cultured Bovine Satellite Cells Contain Subpopulations Whose Identities Remain to Be Determined

Compared to other clusters (Supplementary File S2), clusters 6, 10, 11, 13, 14, and 15 did not express significantly higher levels of markers of myogenic cells such as PAX3, PAX7, MYOD1, MYF5, and MYOG (Pownall et al., 2002); thus, these clusters were not myogenic cells. Skeletal muscle contains not only myogenic cells but also nonmyogenic cells such as endothelial cells, pericytes, smooth muscle cells, fibroblasts, and glial cells. However, none of clusters 6, 10, 11, 13, 14, and 15 appeared to be these nonmyogenic cells based on the expression levels of marker genes such as CDH5 and PECAM1 for endothelial cells (Elmentaite et al., 2021), NOTCH3 and MCAM for pericytes (Elmentaite et al., 2021), ACTA2 and MYH11 for smooth muscle cells (Kumar et al., 2017), COL1A and S100A4 for fibroblasts (Kumar et al., 2017), and FOXD3 and SOX10 for glial cells (Elmentaite et al., 2021) in these clusters (Supplementary File S2). It is possible that clusters 6, 10, 11, 13, 14, and 15 represented novel cell types in bovine skeletal muscle co-isolated with satellite cells or that they were inaccurately clustered by the computational program used.

REFERENCES

Andrews, T. S., and Hemberg, M. (2018). Identifying Cell Populations with scRNASeq. *Mol. Aspects Med.* 59, 114–122. doi:10.1016/j.mam.2017.07.002

CONCLUSIONS

Results of this scRNA-seq study suggest that bovine satellite cells are possibly composed of subpopulations that differ in transcriptional status, proliferation rate, and myogenic potential. This notion is consistent with the conclusion from scRNA-seq studies of mouse satellite cells (Cho and Doles, 2017; van den Brink et al., 2017). Results of this study also suggest the presence of FAP cells in bovine skeletal muscle, although their origin remains to be determined. Because skeletal muscle growth and intramuscular fat are economically important traits in cattle, further characterization of the different subpopulations of satellite cells as well as FAPs in bovine skeletal muscle could lead to the development of new strategies to improve these traits or to identify DNA sequences and variants associated with these traits in cattle.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184128>.

ETHICS STATEMENT

The animal study was reviewed and approved by Virginia Tech Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

HJ designed the research; PL and YQ performed the experiments; PL and HJ analyzed the data; PL, ZT, and HJ wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This project was supported in part by Agriculture and Food Research Initiative competitive grant no. 2016-67015-24471 from the USDA National Institute of Food and Agriculture.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.742077/full#supplementary-material>

Cho, D. S., and Doles, J. D. (2017). Single Cell Transcriptome Analysis of Muscle Satellite Cells Reveals Widespread Transcriptional Heterogeneity. *Gene* 636, 54–63. doi:10.1016/j.gene.2017.09.014

Dhawan, J., and Rando, T. A. (2005). Stem Cells in Postnatal Myogenesis: Molecular Mechanisms of Satellite Cell Quiescence, Activation and Replenishment. *Trends Cel Biol.* 15 (12), 666–673. doi:10.1016/j.tcb.2005.10.007

- Elmentaite, R., Kumasaka, N., Roberts, K., Fleming, A., Dann, E., King, H. W., et al. (2021). Cells of the Human Intestinal Tract Mapped Across Space and Time. *Nature* 597 (7875), 250–255. doi:10.1038/s41586-021-03852-1
- Estrella, N. L., Desjardins, C. A., Nocco, S. E., Clark, A. L., Maksimenko, Y., and Naya, F. J. (2015). MEF2 Transcription Factors Regulate Distinct Gene Programs in Mammalian Skeletal Muscle Differentiation. *J. Biol. Chem.* 290 (2), 1256–1268. doi:10.1074/jbc.M114.589838
- Ge, X., Zhang, Y., and Jiang, H. (2013). Signaling Pathways Mediating the Effects of Insulin-like Growth Factor-I in Bovine Muscle Satellite Cells. *Mol. Cell Endocrinol.* 372 (1–2), 23–29. doi:10.1016/j.mce.2013.03.017
- Gupta, R. K., Arany, Z., Seale, P., Mepani, R. J., Ye, L., Conroe, H. M., et al. (2010). Transcriptional Control of Preadipocyte Determination by Zfp423. *Nature* 464 (7288), 619–623. doi:10.1038/nature08816
- Handley, A., Schauer, T., Ladurner, A. G., and Margulies, C. E. (2015). Designing Cell-type-specific Genome-wide Experiments. *Mol. Cell.* 58 (4), 621–631. doi:10.1016/j.molcel.2015.04.024
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., et al. (2021). Integrated Analysis of Multimodal Single-Cell Data. *Cell* 184 (13), 3573–3587. doi:10.1016/j.cell.2021.04.048
- Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A Practical Guide to Single-Cell RNA-Sequencing for Biomedical Research and Clinical Applications. *Genome Med.* 9 (1), 75. doi:10.1186/s13073-017-0467-4
- Hathaway, M. R., Hembree, J. R., Pampusch, M. S., and Dayton, W. R. (1991). Effect of Transforming Growth Factor Beta-1 on Ovine Satellite Cell Proliferation and Fusion. *J. Cel. Physiol.* 146 (3), 435–441. doi:10.1002/jcp.1041460314
- Joe, A. W. B., Yi, L., Natarajan, A., Le Grand, F., So, L., Wang, J., et al. (2010). Muscle Injury Activates Resident Fibro/adipogenic Progenitors that Facilitate Myogenesis. *Nat. Cel Biol.* 12 (2), 153–163. doi:10.1038/ncb2015
- Kuang, S., and Rudnicki, M. A. (2008). The Emerging Biology of Satellite Cells and Their Therapeutic Potential. *Trends Mol. Med.* 14 (2), 82–91. doi:10.1016/j.molmed.2007.12.004
- Kumar, A., D'Souza, S. S., Moskvina, O. V., Toh, H., Wang, B., Zhang, J., et al. (2017). Specification and Diversification of Pericytes and Smooth Muscle Cells from Mesenchymalangioblasts. *Cel Rep.* 19 (9), 1902–1916. doi:10.1016/j.celrep.2017.05.019
- Leferova, M. I., Haakonsson, A. K., Lazar, M. A., and Mandrup, S. (2014). PPAR γ and the Global Map of Adipogenesis and beyond. *Trends Endocrinol. Metab.* 25 (6), 293–302. doi:10.1016/j.tem.2014.04.001
- Leng, X., Ji, X., Hou, Y., Settlege, R., and Jiang, H. (2019). Roles of the Proteasome and Inhibitor of DNA Binding 1 Protein in Myoblast Differentiation. *FASEB J.* 33 (6), 7403–7416. doi:10.1096/fj.201800574rr
- Leng, X., and Jiang, H. (2019). Effects of Arachidonic Acid and its Major Prostaglandin Derivatives on Bovine Myoblast Proliferation, Differentiation, and Fusion. *Domest. Anim. Endocrinol.* 67, 28–36. doi:10.1016/j.domaniend.2018.12.006
- Maroto, M., Reshef, R., Münsterberg, A. E., Koester, S., Goulding, M., and Lassar, A. B. (1997). Ectopic Pax-3 Activates MyoD and Myf-5 Expression in Embryonic Mesoderm and Neural Tissue. *Cell* 89 (1), 139–148. doi:10.1016/s0092-8674(00)80190-7
- Mauro, A. (1961). Satellite Cell of Skeletal Muscle Fibers. *J. Biophys. Biochem. Cytol.* 9, 493–495. doi:10.1083/jcb.9.2.493
- Mi, H., Muruganujan, A., Casagrande, J. T., and Casagrande, P. D. (2013). Large-scale Gene Function Analysis with the PANTHER Classification System. *Nat. Protoc.* 8 (8), 1551–1566. doi:10.1038/nprot.2013.092
- Millay, D. P., O'Rourke, J. R., Sutherland, L. B., Bezprozvannaya, S., Shelton, J. M., Bassel-Duby, R., et al. (2013). Myomaker Is a Membrane Activator of Myoblast Fusion and Muscle Formation. *Nature* 499 (7458), 301–305. doi:10.1038/nature12343
- Pownall, M. E., Gustafsson, M. K., and Emerson, C. P., Jr (2002). Myogenic Regulatory Factors and the Specification of Muscle Progenitors in Vertebrate Embryos. *Annu. Rev. Cel Dev. Biol.* 18, 747–783. doi:10.1146/annurev.cellbio.18.012502.105758
- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nat. Methods* 14 (10), 979–982. doi:10.1038/nmeth.4402
- Robinson, M. D., and Smyth, G. K. (2007). Moderated Statistical Tests for Assessing Differences in Tag Abundance. *Bioinformatics* 23 (21), 2881–2887. doi:10.1093/bioinformatics/btm453
- Seale, P., Sabourin, L. A., Girgis-Gabardo, A., Mansouri, A., Gruss, P., and Rudnicki, M. A. (2000). Pax7 Is Required for the Specification of Myogenic Satellite Cells. *Cell* 102 (6), 777–786. doi:10.1016/s0092-8674(00)00066-0
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6 (5), 377–382. doi:10.1038/nmeth.1315
- Tontonoz, P., and Spiegelman, B. M. (2008). Fat and Beyond: The Diverse Biology of PPAR γ . *Annu. Rev. Biochem.* 77, 289–312. doi:10.1146/annurev.biochem.77.061307.091829
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells. *Nat. Biotechnol.* 32 (4), 381–386. doi:10.1038/nbt.2859
- Uezumi, A., Fukada, S.-i., Yamamoto, N., Takeda, S. i., and Tsuchida, K. (2010). Mesenchymal Progenitors Distinct from Satellite Cells Contribute to Ectopic Fat Cell Formation in Skeletal Muscle. *Nat. Cel Biol.* 12 (2), 143–152. doi:10.1038/ncb2014
- Uezumi, A., Fukada, S., Yamamoto, N., Ikemoto-Uezumi, M., Nakatani, M., Morita, M., et al. (2014). Identification and Characterization of PDGFR α + Mesenchymal Progenitors in Human Skeletal Muscle. *Cell Death Dis.* 5, e1186. doi:10.1038/cddis.2014.161
- van den Brink, S. C., Sage, F., Vértessy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., et al. (2017). Single-cell Sequencing Reveals Dissociation-Induced Gene Expression in Tissue Subpopulations. *Nat. Methods* 14 (10), 935–936. doi:10.1038/nmeth.4437
- Yu, D., Huber, W., and Vitek, O. (2013). Shrinkage Estimation of Dispersion in Negative Binomial Models for RNA-Seq Experiments with Small Sample Size. *Bioinformatics* 29 (10), 1275–1282. doi:10.1093/bioinformatics/btt143
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell.* 65 (4), 631–643. doi:10.1016/j.molcel.2017.01.023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lyu, Qi, Tu and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

ACTA2 actin alpha 2, smooth muscle

ADIPOQ adiponectin

BP biological process

CC cellular component

CDH5 cadherin 5

COL1A collagen type I alpha 1 chain

DES desmin

FAPs fibro-adipogenic progenitors

FDR false discovery rate

FE fold enrichment

FOXD3 forkhead box D3

GEM gel bead-in-emulsion

GO gene ontology

LEP leptin

MB myoglobin

MCAM melanoma cell adhesion molecule

MEF myocyte enhancer factor

MF molecular function

MYF5 myogenic factor 5

MYH3 myosin heavy chain 3

MYH11 myosin heavy chain 11

MYL1 myosin light chain 1

MYMK myomaker or myoblast fusion factor

MYOD1 myogenic differentiation 1

MYOG myogenin

NEB nebulin

NOTCH3 notch receptor 3

PAX3 paired box 3

PAX7 paired box 7

PDGFRA platelet derived growth factor receptor alpha

PECAM1 platelet and endothelial cell adhesion molecule 1

PPARG peroxisome proliferator activated receptor gamma

S100A4 S100 calcium binding protein A4

scrNA-seq single-cell RNA sequencing

SOX10 SRY-box transcription factor 10

STAC3 SH3 and cysteine rich domain 3

UMI unique molecular identifier

ZNF423 zinc finger protein 423



FMixFN: A Fast Big Data-Oriented Genomic Selection Model Based on an Iterative Conditional Expectation algorithm

Wenwu Xu, Xiaodong Liu, Mingfu Liao, Shijun Xiao, Min Zheng, Tianxiong Yao, Zuoquan Chen, Lusheng Huang and Zhiyan Zhang *

State Key Laboratory for Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University, Nanchang, China

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

Yutaka Masuda,
Rakuno Gakuen University, Japan
Alencar Xavier,
Corteva Agriscience™, United States

*Correspondence:

Zhiyan Zhang
bioducklily@hotmail.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 07 June 2021

Accepted: 22 October 2021

Published: 18 November 2021

Citation:

Xu W, Liu X, Liao M, Xiao S, Zheng M,
Yao T, Chen Z, Huang L and Zhang Z
(2021) FMixFN: A Fast Big Data-
Oriented Genomic Selection Model
Based on an Iterative Conditional
Expectation algorithm.
Front. Genet. 12:721600.
doi: 10.3389/fgene.2021.721600

Genomic selection is an approach to select elite breeding stock based on the use of dense genetic markers and that has led to the development of various models to derive a predictive equation. However, the current genomic selection software faces several issues such as low prediction accuracy, low computational efficiency, or an inability to handle large-scale sample data. We report the development of a genomic prediction model named FMixFN with four zero-mean normal distributions as the prior distributions to optimize the predictive ability and computing efficiency. The variance of the prior distributions in our model is precisely determined based on an F2 population, and genomic estimated breeding values (GEBV) can be obtained accurately and quickly in combination with an iterative conditional expectation algorithm. We demonstrated that FMixFN improves computational efficiency and predictive ability compared to other methods, such as GBLUP, SSgblup, MIX, BayesR, BayesA, and BayesB. Most importantly, FMixFN may handle large-scale sample data, and thus should be able to meet the needs of large breeding companies or combined breeding schedules. Our study developed a Bayes genomic selection model called FMixFN, which combines stable predictive ability and high computational efficiency, and is a big data-oriented genomic selection model that has potential in the future. The FMixFN method can be freely accessed at <https://zenodo.org/record/5560913> (DOI: 10.5281/zenodo.5560913).

Keywords: genomic selection, model, big data-oriented, GEBV, FMixFN

INTRODUCTION

Based on the use of genomic information and prediction of the genetic merit of animals, genomic selection is changing breeding strategies and approaches in livestock (Goddard and Hayes, 2009). Among many agricultural animals and plants, estimated breeding values (EBV) predicted from genomic information are now widely used (Duchemin et al., 2012; Pollak et al., 2012; Preisinger, 2012; Ibáñez-Escriche et al., 2014; Samorè and Fontanesi, 2016; Mrode et al., 2018). Comparative studies on both simulated and real data have shown that genomic EBV (GEBV) tends to have higher accuracy than breeding values estimated using pedigree relationships. The accuracy of GEBV is mainly impacted by the nature of the single nucleotide polymorphism (SNP) panel used, the size of the training data, the population structure, the relationships between individuals in the training and validation population, and the genetic architecture of the trait, in particular, the number of loci

affecting the trait and the distribution of their effects (Daetwyler et al., 2008; Goddard, 2009; Meuwissen, 2009). Usually, the most accurate method to predict genetic value or phenotype based on the SNP genotypes is to fit all SNPs simultaneously, treating the SNP effects as they are drawn from a prior distribution that matches as closely as possible the true distribution of SNP effects (Goddard, 2009; Chatterjee et al., 2013). The assumption that, in the SNP-best linear unbiased prediction (BLUP) or genomic BLUP (GBLUP) model, each of the SNPs explains equal variance, i.e., that the more complex traits are controlled by very many quantitative trait loci (QTL), each with a tiny effect, could be imprecise if a trait is affected by a small number of QTL, each with a large effect (Meuwissen et al., 2001; VanRaden, 2008). In other models, the distribution of SNP effects is allowed to depart from a pseudo-infinitesimal distribution. BayesA extended the SNP-BLUP model by estimating the variance of each marker separately, and an inverse chi-square prior was used to estimate these variances (Meuwissen et al., 2001). In BayesB, it was assumed that most of the markers have a zero effect on the targeted trait, and the prior distribution of the variances is a mixture of a distribution with zero variance and an inverse chi-squared distribution, with some SNPs having a zero effect, and some SNPs having a large effect on the trait (Meuwissen et al., 2001). The true distribution of the effect sizes is not known, but a mixture of normal distributions can approximate a wide variety of distributions by varying the mixing proportions (Mclachlan Basford et al., 1988; Silverman, 1996; Luan et al., 2009; Moser et al., 2015; Goddard et al., 2016). Kemper et al. and Erbe et al. presented and extended a model named “BayesR,” which used a mixture of four normal distributions as prior, each with a zero mean but with variances of 0, $0.0001 \delta_g^2$, $0.001 \delta_g^2$, and $0.01 \delta_g^2$ for genomic prediction (Erbe et al., 2012; Kemper et al., 2015). In the applications of this model, it has been assumed that the mixing proportions are drawn from a Dirichlet distribution with parameters (1, 1, 1, 1). In a simulation study in which the genetic model included a finite number of loci with exponentially distributed effects, the Bayes-based model provided more accurate prediction of genetic value than GBLUP.

Although Bayes-based models have the potential for the development of more faithful genetic models and seem to be the best choice for estimating GEBV, they require long computing times since they use computer-intensive MCMC techniques (Meuwissen et al., 2001; Xu, 2003; Verbyla et al., 2010; Habier et al., 2011; Cheng et al., 2015). For practical applications and for computer simulations of genomic selection breeding schemes, which need many selection rounds and replications, it would be useful to have a much faster algorithm for the calculation of Bayes-based GEBV. Several non-MCMC algorithms have been proposed to improve computational efficiency for linear models with differential shrinkage of SNP effects or with variable selection. Methods BL and BhGLM were developed by Xu et al. and Yi et al., respectively, which used Expectation-Maximization (EM) algorithms (Yi and Banerjee, 2009; Xu, 2010). VanRaden et al. (2009) presented two non-linear predictions A and B that are analogous to the BayesA and BayesB, respectively. Meuwissen et al. (2009) presented a fast heuristic iterative conditional expectation (Zhao et al.) algorithm,

where the posterior expectation of SNP effects was calculated analytically, assuming a fixed known double exponential (DE) parameter and dispersion parameters. Dong et al. (2017) formulated an algorithm based on the same model as the ICE algorithm, which uses a product of univariate densities instead of the multivariate normal density to estimate SNP effects, but the a priori hypothesis on the size of the SNP effects is based on the Pareto principle, which was proposed by the economist Vilfredo Pareto at the beginning of the 20th century. This principle states that approximately 20% of the population possesses 80% of the wealth in a country. Similar theories have been further applied in various fields, such as in genomic prediction by Yu and Meuwissen (2011). In their study, the a priori distribution of the genomic prediction model was a mixture of two normal distributions, which assumes that $x\%$ of the SNPs explain $(100 - x)\%$ of the genetic variance, and the remaining $(100 - x)\%$ of SNPs explain the remaining $x\%$ of genetic variance (Grosfeld-Nir et al., 2007). Using this economic principle to assume the a priori distribution of the marker effect is not very convincing, leading to only a general predictive accuracy in Dong’s research.

For genomic selection, most of the focus has been on prediction accuracy and computational efficiency, but the computing limits are an increasingly important aspect that needs to be taken into account. The direct method of genomic selection can provide GEBV in a short computing time when the number of individuals in the population is small, but some studies have shown that when the dimensions of the kinship matrix exceed hundreds of thousands or even millions, the process to inverse the matrix inverse becomes very difficult due to the limitations in computer memory and computational time (Misztal, 2016). According to the Council on Dairy Cattle Breeding (https://queries.uscdcb.com/Genotype/cur_freq.html), more than 5,000,000 Holstein cows have been genotyped as of July 2021. With the accumulation of breeding data, there is an urgent need for a genomic selection model that can handle large-scale sample data.

In our study, we presented an ICE-based prediction model with four zero-mean normal distributions as the prior distribution and the variance of which have been obtained accurately based on the 374 standardized phenotypes in an F2 population. This model with four normal distributions and variances classified into four categories was referred to as FMixFN, where MixFN refers to the prior distribution of FMixFN was a mixture of four normal distributions, and the first “F” refers to “Fast.” As a test, the predictive ability and computation time obtained with GBLUP, SSgblup, Bayesian mixture regression (MIX), BayesR, BayesA, BayesB, and FMixFN were compared first based on six traits with different heritabilities and different genetic architectures by using cross-validation. Then FMixFN was evaluated by using data from Duroc and Asian rice experiments, respectively (Zhao et al., 2011; Ding et al., 2019). This study also evaluated the efficiency of FMixFN and its ability to handle large-scale sample data with 20 sets of sample data simulated by the QMsim software.

MATERIALS AND METHODS

All procedures including experimental animals established and tissue collection were performed in accordance with the guidelines approved by the Ministry of Agriculture of China. This study was approved by the ethics committee of Jiangxi Agricultural University.

Data

An F2 design resource population was developed between 2000 and 2006 (Guo et al., 2009) as follows: two White Duroc sires and 17 Erhualian dams were mated to produce F1 animals, from which 9 F1 boars and 59 F1 sows were intercrossed (avoiding full-sib mating) to produce 967 F2 males and 945 F2 females (in total $n = 1,912$) in six batches. All the F2 animals were kept under standard indoor conditions at the experimental farm of Jiangxi Agricultural University (China). Then the F2 piglets were weaned at 46 days, and males were castrated at 90 days. At 240 ± 6 days of age, 1,030 F2 animals including 549 gilts and 481 barrows were slaughtered at 70–120 kg live weight.

Genomic DNA was isolated from ear tissue with a standard phenol/chloroform extraction method. All DNA samples were diluted to a final concentration of 50 ng/μl in 96-well plates. In total, 933 F2 were genotyped with the Illumina PorcineSNP60 BeadChip on an iScan System (Illumina, United States) following the manufacturer's protocol (Ramos et al., 2009). Quality control procedures were implemented by PLINK (version 1.07) (Chang et al., 2015). Briefly, SNPs with unspecific positions on the genome build 10.2, a call rate lower than 90%, and a minor allele frequency (MAF) lower than 1% were eliminated, and animals with a missing typing rate higher than 10% were also removed. In total, 374 phenotypes were measured on the individuals of the F2 population, including carcass traits, reproductive traits, immune traits, meat traits, growth traits, and epigenetic traits (see Additional file 1: **Supplementary Table S1**). These 374 traits were then divided into three groups according to their heritability, i.e., 68 traits with high heritability ($h^2 > 0.4$), 148 traits with a moderate heritability ($0.2 < h^2 < 0.4$), and 158 with low heritability ($h^2 < 0.2$).

Estimation of Substitution Effects

We used the GEMMA software to calculate the substitution effects of 14,320,159 SNPs on the 374 traits included in the standard linear model (Zhou and Stephens, 2012). Sex was included as a fixed effect, and heritability was estimated by using the *-lmm* procedure implemented in GEMMA. Population stratification was adjusted by including a genomic relationship matrix. Briefly, the model was as follows:

$$y = Wa + X\beta + u + e; u \sim MVN_n(0, \sigma_u^2 K), e \sim MVN_n(0, \sigma_e^2 I_n)$$

where y is an n element vector of phenotypic values, all the traits were normalized before calculation so that the substitution effects were comparable among all the phenotypes, W is a design matrix of covariates, a is a vector of fixed effects, X is a vector of genotypes at each locus, β is the effect size of SNPs, and u is the vector of random effects following a multivariate normal distribution $MVN_n(0, \sigma_u^2 K)$, e is the vector of errors following $MVN_n(0, \sigma_e^2 I_n)$, σ_u^2 and σ_e^2 are polygenic variance and residual variance, respectively, which are

estimated based on the REML average information (AI) algorithm. K is a known kinship matrix estimated from genome sequence variants, and I_n being an $n \times n$ identity matrix.

Distribution of Additive Genetic Variance

Three genotypes “AA,” “Aa,” and “aa” were assumed each locus and were represented by 0, 1, and 2, respectively, with p and q the frequencies of alleles “A” and “a,” respectively. Assuming that the effect value of this locus is estimated as β (with no dominance), the additive genetic variance can be expressed as $2pq\beta^2$ (Park et al., 2011). In the group of traits with high heritability, all the loci for each phenotype were put together and ranked by additive genetic variance from large to small. And all the ordered loci were equally divided into four groups. For each group, the proportion of the sum of the additive genetic variances of all loci to the total additive genetic variance (or variance-ratio thereafter) was calculated, equal to a_1, b_1, c_1 , and d_1 , respectively (Subsequently called variance ratio). Similarly, the same method was used for the groups of traits with a moderate heritability and a low heritability, resulting in a_2, b_2, c_2, d_2 , and a_3, b_3, c_3, d_3 , respectively. Therefore, the four expected variances in each of the three groups can be expressed as:

Group of traits with high heritability:

$$\delta_1^2 = \frac{a_1 V_g}{\gamma M}; \delta_2^2 = \frac{b_1 V_g}{\gamma M}; \delta_3^2 = \frac{c_1 V_g}{\gamma M}; \delta_4^2 = \frac{d_1 V_g}{\gamma M}$$

Group of traits with a moderate heritability:

$$\delta_1^2 = \frac{a_2 V_g}{\gamma M}; \delta_2^2 = \frac{b_2 V_g}{\gamma M}; \delta_3^2 = \frac{c_2 V_g}{\gamma M}; \delta_4^2 = \frac{d_2 V_g}{\gamma M}$$

Group of traits with a low heritability:

$$\delta_1^2 = \frac{a_3 V_g}{\gamma M}; \delta_2^2 = \frac{b_3 V_g}{\gamma M}; \delta_3^2 = \frac{c_3 V_g}{\gamma M}; \delta_4^2 = \frac{d_3 V_g}{\gamma M}$$

Where V_g and M is the additive variance and the number of markers, respectively. γ is set to 0.25.

Analytical Derivation for FMixFN

The linear model for genomic prediction was as follows:

$$y = Xb + Bg + e$$

Where n individuals and m SNPs were assumed. Thus, y is a $n \times 1$ vector of phenotypes recorded; b is the vector of fixed effects; g is a $m \times 1$ vector of additive SNP effects; e is a vector of residual errors; X is the design matrix for fixed effects; and B is standardized design matrix for additive SNP effects (coded as 0 for genotype “AA,” 1 for “Aa” and 2 for “aa,” respectively).

In this study, the prior distribution with four zero-mean normal distributions was written as a function of prior distributions of SNP variance as determined above:

$$\pi(g_j) = \gamma\phi(g_j|0, \delta_1^2) + \gamma\phi(g_j|0, \delta_2^2) + \gamma\phi(g_j|0, \delta_3^2) + \gamma\phi(g_j|0, \delta_4^2); \gamma = 0.25 \quad (1)$$

$\pi(g_j)$ is the univariate normal distribution, the effects of SNPs are obtained by using the Iterative Conditional Expectation algorithm

(Meuwissen et al., 2009). In brief, assume that $E(g_j|y_{-j})$ is estimated, the current effects of all the other SNPs are used to calculate the y_{-j} as follows:

$$y_{-j} = y - Xb - \sum_{k \neq j} B_k g_k$$

where B_k is a vector from the k^{th} column of B , the expectation of SNP effect, $E(g_j|y_{-j})$, is then estimated by a Bayesian model in the next round:

$$\begin{aligned} E(g_j|y_{-j}) &= \int_{-\infty}^{+\infty} g_j f(g_j|y_{-j}) dg_j \\ &= \frac{\int_{-\infty}^{+\infty} g_j f(y_{-j}|B_j g_j, I\delta_e^2) \pi(g_j) dg_j}{f(y_{-j})} \\ &= \frac{\int_{-\infty}^{+\infty} g_j f(y_{-j}|B_j g_j, I\delta_e^2) \pi(g_j) dg_j}{\int_{-\infty}^{+\infty} f(y_{-j}|B_j g_j, I\delta_e^2) \pi(g_j) dg_j} \end{aligned} \quad (2)$$

$f(y_{-j})$ is a marginal distribution function of y_{-j} and can be calculated using the law of total cumulance: $\int_{-\infty}^{+\infty} f(y_{-j}|B_j g_j, I\delta_e^2) \pi(g_j) dg_j$. Calculating $f(y_{-j}|B_j g_j, I\delta_e^2)$ is computationally demanding because it is a multivariate normal density, which involves calculating the determinant and the inverse of the variance-covariance matrix for the data y_{-j} . Therefore, we simplified the derivation by using a univariate normal densities $f(Y|g_j, \delta^2)$ to replace $f(y_{-j}|B_j g_j, I\delta_e^2)$, where $Y = (B_j' B_j)^{-1} B_j' y_{-j}$ and $\delta^2 = (B_j' B_j)^{-1} \delta_e^2$; details of the derivation process are as follows:

$$\begin{aligned} f(y_{-j}|B_j g_j, I\delta_e^2) &\propto \exp\left[-\frac{(y_{-j} - B_j g_j)'(y_{-j} - B_j g_j)}{2\delta_e^2}\right] \\ &= \exp\left(-\frac{y_{-j}' y_{-j} - 2B_j y_{-j} g_j + B_j B_j' g_j^2}{2\delta_e^2}\right) \\ &= \exp\left[-\frac{g_j^2 - 2(B_j B_j)^{-1} B_j y_{-j} g_j + (B_j B_j)^{-1} y_{-j}' y_{-j}}{2(B_j B_j)^{-1} \delta_e^2}\right] \\ &= \exp\left\{-\frac{[g_j - (B_j B_j)^{-1} B_j y_{-j}]' [g_j - (B_j B_j)^{-1} B_j y_{-j}] + (B_j B_j)^{-1} y_{-j}' y_{-j}}{2(B_j B_j)^{-1} \delta_e^2}\right\} \\ &\propto \exp\left\{-\frac{[g_j - (B_j B_j)^{-1} B_j y_{-j}]^2}{2(B_j B_j)^{-1} \delta_e^2}\right\} \propto \exp\left[-\frac{(g_j - Y)^2}{2\delta^2}\right] \propto f(Y|g_j, \delta^2) \end{aligned}$$

Therefore, the equation $E(g_j|y_{-j})$ can be written as:

$$E(g_j|y_{-j}) = \frac{\int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \pi(g_j) dg_j}{\int_{-\infty}^{+\infty} f(Y|g_j, \delta^2) \pi(g_j) dg_j} \quad (3)$$

the numerator of **Eq. 3** can be broken down into four terms combined with **Eq. 1** as follows:

$$\begin{aligned} &\gamma \int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \phi(g_j|0, \delta_1^2) dg_j \\ &+ \gamma \int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \phi(g_j|0, \delta_2^2) dg_j \end{aligned}$$

$$\begin{aligned} &+ \gamma \int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \phi(g_j|0, \delta_3^2) dg_j \\ &+ \gamma \int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \phi(g_j|0, \delta_4^2) dg_j \end{aligned} \quad (4)$$

The first term of **Eq. 4** can be derived as follows:

$$\begin{aligned} &\gamma \int_{-\infty}^{+\infty} g_j f(Y|g_j, \delta^2) \phi(g_j|0, \delta_1^2) dg_j \\ &= \gamma \int_{-\infty}^{+\infty} g_j \frac{1}{\delta \sqrt{2\pi}} \exp\left[-\frac{(Y - g_j)^2}{2\delta^2}\right] \frac{1}{\delta_1 \sqrt{2\pi}} \exp\left[-\frac{g_j^2}{2\delta_1^2}\right] dg_j \\ &= \frac{\gamma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{g_j}{\delta \delta_1 \sqrt{2\pi}} \exp\left[-\frac{(Y - g_j)^2}{2\delta^2} - \frac{g_j^2}{2\delta_1^2}\right] dg_j \\ &= \frac{\gamma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{g_j}{\delta \delta_1 \sqrt{2\pi}} \exp\left[-\frac{\left(g_j - \frac{Y\delta_1^2}{\delta^2 + \delta_1^2} + \frac{Y^2\delta_1^2}{\delta^2 + \delta_1^2}\right)(\delta^2 + \delta_1^2)}{2\delta^2 \delta_1^2}\right] dg_j \\ &= \frac{\gamma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{g_j}{\delta \delta_1 \sqrt{2\pi}} \exp\left[-\frac{\left(g_j - \frac{Y\delta_1^2}{\delta^2 + \delta_1^2}\right)^2 (\delta^2 + \delta_1^2)}{2\delta^2 \delta_1^2} - \frac{Y^2}{2(\delta^2 + \delta_1^2)}\right] dg_j \\ &= \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_1^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_1^2}} \int_{-\infty}^{+\infty} \frac{g_j}{\delta \delta_1 \sqrt{2\pi}} \exp\left[-\frac{\left(g_j - \frac{Y\delta_1^2}{\delta^2 + \delta_1^2}\right)^2 (\delta^2 + \delta_1^2)}{2\delta^2 \delta_1^2}\right] dg_j \end{aligned}$$

Here, the last term of this formula equals $\frac{Y\delta_1^2}{\delta^2 + \delta_1^2}$ as it can be taken as calculating the expected value of g_j in the normal distribution with mean $\frac{Y\delta_1^2}{\delta^2 + \delta_1^2}$, and variance $\frac{\delta^2 \delta_1^2}{\delta^2 + \delta_1^2}$. Therefore, the first term of **Eq. 4** can be written as follows:

$$\frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_1^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_1^2}} \frac{Y\delta_1^2}{\delta^2 + \delta_1^2}$$

Here, the derivation process for the remaining terms of **Eq. 4** was the same as for this term, and therefore, the final form of the numerator of **Eq. 3** is:

$$\begin{aligned} &\frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_1^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_1^2}} \frac{Y\delta_1^2}{\delta^2 + \delta_1^2} \\ &+ \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_2^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_2^2}} \frac{Y\delta_2^2}{\delta^2 + \delta_2^2} \\ &+ \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_3^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_3^2}} \frac{Y\delta_3^2}{\delta^2 + \delta_3^2} \\ &+ \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_4^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_4^2}} \frac{Y\delta_4^2}{\delta^2 + \delta_4^2} \end{aligned} \quad (5)$$

there is no g_j in the integrand of the denominator in **Eq. 3** compared to that of the numerator. Therefore, it should calculate

the cumulative probability from $-\infty$ to $+\infty$, but not calculate the expected value, and this value is 1. Thus, the denominator in Eq. 3 can be written as:

$$\begin{aligned} & \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_1^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_1^2}} \\ & + \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_2^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_2^2}} \\ & + \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_3^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_3^2}} \\ & + \frac{\gamma}{\sqrt{2\pi}} \exp\left[-\frac{Y^2}{2(\delta^2 + \delta_4^2)}\right] \frac{1}{\sqrt{\delta^2 + \delta_4^2}} \end{aligned} \quad (6)$$

thus, the final form for Eq. 3 is derived:

$$E(g_j|y_{-j}) = \frac{\frac{\gamma^2}{\delta^2 + \delta_1^2} + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_2^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_2^2}} + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_3^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_3^2}} + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_4^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_4^2}}}{1 + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_2^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_2^2}} + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_3^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_3^2}} + \exp\left[\frac{\gamma^2}{2(\delta^2 + \delta_1^2)} - \frac{\gamma^2}{2(\delta^2 + \delta_4^2)}\right] \frac{\sqrt{\delta^2 + \delta_1^2}}{\sqrt{\delta^2 + \delta_4^2}}}$$

Here, the fixed effects are estimated at each iteration by the formula: $\hat{b} = (X'X)^{-1}X'(y - B\hat{g})$. Convergence of solutions at the t th iteration was judged based on the formula $\frac{(G^t - G^{t-1})'(G^t - G^{t-1})}{(G^t)'G^t} < 10^{-8}$, where $G = (\hat{b} \hat{g}')'$. It ends at the iteration when all the SNPs have been calculated once.

Analytical Models

In the following analysis, we used GBLUP (Meuwissen et al., 2001; VanRaden, 2008), SSgblup (Legarra et al., 2009; Christensen and Lund, 2010), FMixFN, MIX (Xavier et al., 2019), BayesR (Kemper et al., 2015), BayesA, and BayesB with respective model fittings to compare their performance, the variance components were pre-estimated using the mixed model. The details of these analyses were as follows:

GBLUP: GBLUP was used to estimate the effects of the markers by BLUP, assuming that each marker explains an equal proportion of the total genetic variance. The software GEMMA was used to implement the GBLUP calculation process (Zhou and Stephens, 2012).

SSgblup: Single-step genomic BLUP (SSgblup), which was developed by Aguilar et al. (2010) and Christensen and Lund (2010), opened the way to perform genomic prediction using phenotype, pedigree, and genomic information simultaneously on both genotyped and non-genotyped individuals via a combined relationship matrix (H). Implementation of SSgblup is completed by the R package “Hiblup” (<https://hiblup.github.io/>).

MIX: the MIXTURE model assumed that the marker effects came from a mixture of two distributions: one distribution with large variance (accommodating large marker effects) and one with small variance (accommodating small marker effects). The distribution to which the marker belongs is sampled from the Bernoulli distribution. The variances of the two distributions underlying the mixture are estimated using a noninformative chi-square distribution. Implementation of MIX is completed by the R package “VIGoR” (<https://cran.r-project.org/web/packages/VIGoR/index.html>).

BayesR: BayesR starts the hierarchical model and poses a mixture of four zero-mean normal distributions as a conditional prior for a specific SNP effect. We use BayesR software to implement the calculation process (<https://cnsgenomics.com/software.html>).

BayesA: BayesA assumes that the distribution of SNP effects follows a Student's t -distribution. Mathematically, it is assumed that each SNP effect comes from a normal distribution but σ^2 can be varied among the SNPs because the t -distribution is not easy to incorporate into a prediction of the marker. A scaled inverted chi distribution, $X^2(\nu, S)$ is usually used as prior for the variance components.

BayesB: The prior distribution of BayesB is a mixture distribution with some SNPs with zero effects and the rest with a t -distribution, and the prior hypothesis of the SNP with non-zero effect is the same as BayesA. The implementation of BayesA and BayesB is completed by the R package “BGLR” (Perez and de los Campos, 2014). All MCMC sampling was run for 50,000 cycles, and the first 20,000 cycles were discarded as burn-in for BayesR, BayesA, and BayesB.

The Verification of Predictive Ability and Computing Time

To test the performance of FMixFN in terms of predictive ability and calculation time, we did the following. Firstly, the variance ratio of the prior distribution of FMixFN was assumed to be random, then two traits were selected to verify the unbiasedness, and the compatibility of the variance ratio estimated in the F2 population. The specific assumptions of the variance ratio are shown in Additional file 1: **Supplementary Table S2**, the first one is to average all variances, i.e. to set the classification with the largest variance ratio at 50%, and the second was to centralize all variances, i.e., the classification with the largest variance ratio is assumed to be 90%. Secondly, we compared the predictive ability and calculation time of FMixFN and other mainstream genomic selection methods and selected two phenotypes with different genetic structures and different heritabilities from each group for cross-validation, one trait is controlled by numerous polygenic genes and the other one is controlled by several loci with large variances. **Supplementary Figure S1** (see Additional file 2: **Supplementary Figure S1**) shows that traits 1, 3, and 5 were controlled by SNPs with large variances, and traits 2, 4, and 6 were controlled by many SNPs, each with a very small effect. After quality control, the remaining number of individuals with the six traits were 839, 832, 834, 840, 784, and 838, respectively, with 33,901, 33,893, 33,891, 33,891, and 33,894 SNPs, respectively, the heritability of each trait was 0.600, 0.560, 0.380, 0.369, 0.145, and 0.107, respectively. Details on the number of individuals, number of SNPs, and heritability estimates are shown in Additional file 1: **Supplementary Table S3**. In addition, we also evaluated the stability of FMixFN by using data from the duroc experiment and Asian rice experiments, respectively, more specific information from this population is also shown in Additional file 1: **Supplementary Table S3**. Finally, to demonstrate that FMixFN can perform genomic prediction analysis based on large-scale sample data with no data overflow error, our study

simulated 20 sets of sample data using QMsim software (Sargolzaei and Schenkel, 2009), which contains 10,000, 20,000,, 190,000, and 200,000 individuals, respectively. Each set of data was obtained through eight generations of mating, combining genotype and phenotype data from generations 3–8, and determining the number of individuals per generation by parameter setting. Genomic information of each individual was set with 10 chromosomes, each chromosome is set 100 cM long and including 101 markers and 100 QTLs, respectively, with a marker mutation rate of 2.5 and QTL mutation rate of 3. Genomic prediction by a replicated training-testing method was used to evaluate the predictive results. Cross-validation of nine replicates was performed. All individuals were randomly and evenly divided into nine groups. In each replicate, one of the groups was selected as the testing data set while the remaining eight groups were used as the training data set, and the results of each cross-validation are shown in Additional file 1: **Supplementary Table S4**. Predictive ability is defined as the correlation between GEBV and the phenotypes adjusted for the covariates ($y - X\hat{u}$) (Meuwissen et al., 2001).

RESULTS

The Expected Variance Ratio

In this study, all traits of the F2 population were divided into three groups based on the heritability of the traits: high, moderate, and low. For the group of traits with high heritability, the calculated $a1$, $b1$, $c1$, and $d1$ were equal to 0.8752, 0.0958, 0.0256, and 0.0032, respectively. For the group of traits with a moderate heritability, the calculated $a2$, $b2$, $c2$, and $d2$ were equal to 0.8367, 0.1246, 0.0342, and 0.0043, respectively. And for the group of traits with low heritability, the calculated $a3$, $b3$, $c3$, and $d3$ were equal to 0.8225, 0.1413, 0.0324, and 0.0036, respectively. Those parameters were composed in the procedure of FMixFN, as FMixFN starts running, the program determines which group of variances is calculated based on the heritability of the experimental trait.

Verification of Unbiasedness

In this study, we selected phenotype 3 and phenotype 4 to verify the unbiasedness of the variance ratio of the prior distribution. When the variance ratio was assumed to be 0.5, 0.25, 0.125, and 0.125, the predictive ability of phenotype 3 and phenotype 4 are 0.4773 and 0.4911, respectively. When the variance ratio is assumed to be 0.9, 0.005, 0.045, 0.005, the predictive ability of each of these phenotypes are 0.4789 and 0.4911, respectively. In contrast, the predictive ability of these two phenotypes estimated by using the original parameters is 0.4787 and 0.4913, respectively.

Predictive Ability and Computing Time

The predictive ability of each of the six F2 traits under the six predictive methods is shown in **Figure 1**. For phenotypes 1, 3, and 5, the predictive ability with BayesR, BayesA, and BayesB was, respectively, 0.0377, 0.0308, and 0.0374 higher than that of FMixFN, and the predictive ability of FMixFN was slightly better than of GBLUP by 0.0045, 0.0013, and 0.0103, respectively. For those three phenotypes, there is almost no

difference between SSgblup and FMixFN in predictive ability, and FMixFN performs better than SSgblup for phenotype5. For phenotypes 2, 4, and 6, FMixFN performed best for phenotype 4, with a predictive ability 0.0129 higher than that of BayesR, BayesA, and BayesB. For phenotype 2, the predictive ability of the five software was similar and was highest with BayesA but only 0.0053 higher than that of FMixFN. For phenotype 6, FMixFN ranked second in the predictive ability, just 0.0027 lower than that of SSgblup. It was worth mentioning that the predictive ability of FMixFN was slightly better than that of other ICE-based Bayesian mixture regression (MIX) by 0.0221, 0.0116, 0.0801, and 0.0263 for phenotype 1, 4, 5, and 6, respectively. The specific information of the predictive ability was also shown in **Table 1**. **Table 2** reports the predictive ability performances of FMixFN and other methods using the Duroc and rice datasets. In the Duroc population, the prediction accuracies were 0.3655, 0.3300, 0.3998, 0.3476, and 0.3589 for FMixFN, GBLUP, BayesR, BayesA, and BayesB, respectively. From the mean value, we found that FMixFN performed slightly worse than BayesR, but outperformed GBLUP, BayesA, and BayesB. In general, the MCMC-based Bayes genome selection algorithm showed some advantages in the traits controlled by several major QTLs, which explained a large proportion of phenotypic variance in some SNPs, while FMixFN performs better than GBLUP. FMixFN is slightly better than some other mainstream methods when traits follow a polygenic model. In this study, we measured the calculation speed of five methods as the average time necessary for the first cross-validation of the six traits. As shown in **Figure 2A**, the average calculation time was 0.54, 0.37, 0.29, 30.2, 42, and 50.5 min for FMixFN, GBLUP, MIX, BayesR, BayesA, and BayesB, respectively.

FMixFN When Dealing With Large-Scale Sample Data

The computational time per iteration of FMixFN increases almost linearly as the number of individuals increases in the reference group, as shown in **Figure 2B**. Data reading time also increased linearly as the amount of sample data increased. Through simulation studies, we also found that FMixFN can calculate GEBV for 200,000 large samples without data overflow. Data overflow usually occurred in exponential functions, where data overflows or underflows could occur when the exponential part of the exponential function was very large or very small. The exponential part of **Eq. 5** and **Eq. 6** was $\exp\left[-\frac{Y^2}{2(\delta^2 + \delta_j^2)}\right]$, in which $Y = (B_j B_j)^{-1} B_j y_{-j}$, B_j was the j column of the accompanying matrix of SNPs. When B_j growth unlimited, the limit of $\exp\left[-\frac{Y^2}{2(\delta^2 + \delta_j^2)}\right]$ is negative infinity, and data underflows will occur. In this study, we found FMixFN could calculate GEBV for 200,000 samples without data overflow. So FMixFN has the potential for use on large-scale sample data.

In conclusion, when there are fewer individuals in the reference group, the computational speeds for ICE-based FMixFN are on the same order of magnitude with GBLUP, and they were much faster than the MCMC-based Bayesian methods. When the number of individuals in the reference

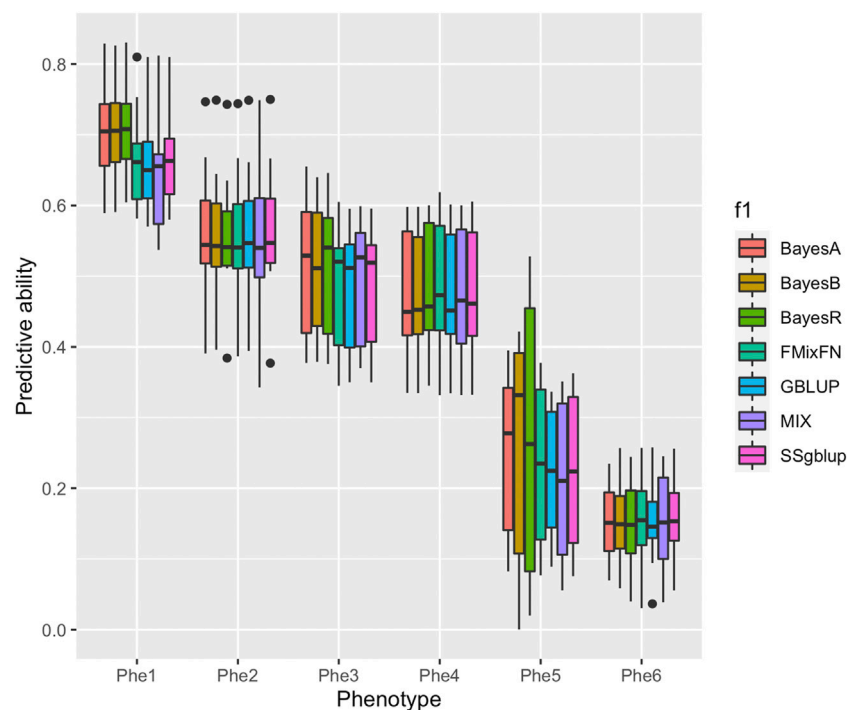


FIGURE 1 | Comparison of predictive ability of BayesA, BayesB, BayesR, FMixFN, GBLUP, MIX, and SSgblup in all six traits. The predictive ability performance of each method was measured by the correlation method, which is the average Pearson correlation between predicted values and phenotypic values.

dimension of the kinship matrix exceeds hundreds of thousands or even millions, the process to inverse the matrix becomes very difficult for the direct method of genomic selection. FMixFN has excellent computational efficiency and can handle large-scale sample data.

DISCUSSION

Our results show that the accuracy of genomic selection is affected by many factors, among which the a priori hypothesis on the size of the QTL effect values for traits is crucial. Usually, the most accurate method to predict genetic values or phenotypes based on SNP genotypes is to fit all SNPs simultaneously, treating the SNP effects as they are drawn from a prior distribution that matches the true distribution of SNP effects as closely as possible (Goddard and Hayes, 2009; Chatterjee et al., 2013). To date, the genetic architecture of many traits is still not entirely understood, which means that the prior hypothesis about the QTL effect distribution of all genomic selection models is empirical. In general, mixed normal distributions are more accurate than a single distribution, because the mixture of normal distributions can approximate a wide variety of distributions. It is important to note that this does not imply the SNP effects are drawn from a mixture of normal distributions, but it merely means that such a mixed distribution can approximate almost any distribution that might describe the distribution of effect sizes. BayesR provides an estimate of the number of causal variants that affect a trait and of the distribution of their effects by approximating the distribution

of effect sizes with a mixture of normal distributions. In our study, the prior distribution of the SNP effect was similar to BayesR, which came from a mixture of four normal distributions with a ratio of 1: 1: 1: 1. The difference with BayesR is that we used an Iterative Conditional Expectation (Zhao et al.) algorithm.

Narrow-sense heritability is defined as the proportion of additive variance to phenotype variance (Wray and Visscher, 2008), which means that a trait with a high heritability is more under the control of genes and is less affected by the environment. Therefore, we divided all 374 traits into three groups according to their heritability, and the variance of the mixed distribution is then calculated in each group. The phenotypes included in our study cover almost all the traits measured in pigs and thus are representative, resulting in a high unbiasedness and compatibility variance. This study randomly assumed two sets of variance ratios and used two representative phenotypes to verify the predictive ability but the results showed that the predictive ability obtained using the original variance may not be optimal. The distribution of marker effects for various traits was different, and no one genomic selection model or a priori hypothesis was optimal for all traits. The variance parameters (variance ratios) obtained in our study were expected to be unbiased as the F2 resource population contained a relatively sufficient number of individuals.

The results showed that the predictive ability of BayesR, BayesA, and BayesB was similar in phenotype1, 3, and 5, and was higher than that of the three other methods, which means the MCMC based Bayes genomic selection model has an advantage in predicting genomic breeding values when the trait is affected by large-effect QTL. This result confirms those reported by Chen et al. (2014). For the three phenotypes, GBLUP resulted in the

TABLE 1 | Prediction performance in all six traits under the seven predictive methods.

Traits/COR	Methods						
	GBLUP	SSgblup	FMixFN	MIX	BayesR	BayesA	BayesB
phe1	0.661	0.6658	0.6655	0.6434	0.7032	0.6962	0.6977
phe2	0.5637	0.5635	0.5591	0.5602	0.556	0.5564	0.561
phe3	0.4774	0.4803	0.4787	0.4833	0.5051	0.5096	0.509
phe4	0.4814	0.4861	0.4915	0.4799	0.4867	0.4786	0.48
phe5	0.2214	0.2232	0.2317	0.1516	0.2691	0.2485	0.2549
phe6	0.1524	0.1564	0.1537	0.1274	0.1512	0.1529	0.1541
Mean	0.4262	0.4292	0.4300	0.4076	0.4452	0.4403	0.4427

COR: The Pearson correlation coefficient between predicted values and phenotypic value.

TABLE 2 | Comparison of predictive ability performances of six methods by using Duroc dataset and rice dataset.

Traits/COR	Methods					
	GBLUP	FMixFN	MIX	BayesR	BayesA	BayesB
Duroc	0.33	0.3655	0.3579	0.3998	0.3476	0.3589
FT	0.4347	0.4506	0.4381	0.4415	0.4295	0.4342
CH	0.6000	0.5696	0.5786	0.5830	0.5809	0.5737
Mean	0.4549	0.4619	0.4582	0.4747	0.4526	0.4556

COR: The Pearson correlation coefficient between predicted values and phenotypic value.

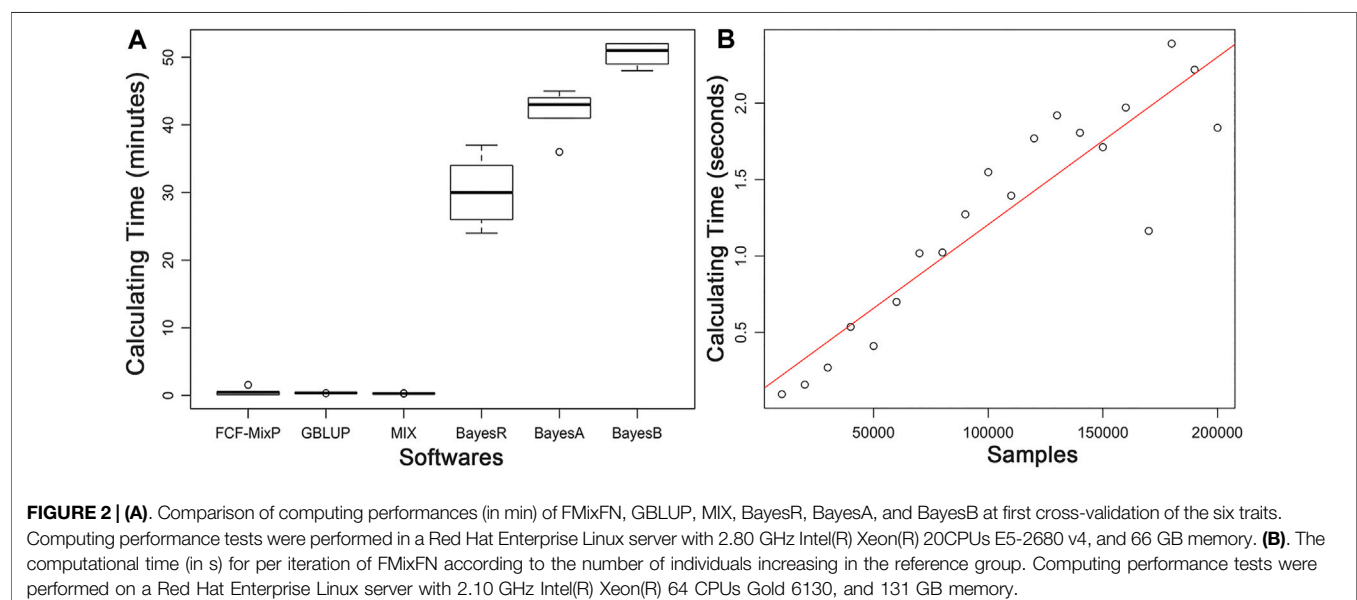
least prediction results because its prior assumption did not match reality as it assumed that traits are controlled by many SNPs, each with a small effect. FMixFN performed slightly better than GBLUP for these three traits, but worse than the Bayes-based methods. Although the prior distribution of FMixFN was a mixture of normal distributions, the posterior variances of SNP effects were not updated, which is a potential drawback for these ICE-based methods. The predictive ability obtained with SSgblup was similar to that with FMixFN because of the addition of pedigree information. However, all the methods yielded almost

the same result for phenotypes 2, 4, and 6, a reasonable explanation may be that when the traits are controlled by many polymorphisms of very small effect, the prior hypothesis of the Bayes-based method is closed to that of GBLUP.

In addition to resulting in stable predictive ability, two other advantages of FMixFN are computational efficiency and its ability to deal with large-scaled sample data. The level of the computational efficiency of the direct method of genomic selection was the same as that of FMixFN when the number of individuals in the reference population was small, but if this number increases, the direct method will not be efficient because the process to invert the matrix will become very difficult due to the limitations in computer memory and computational time. Our study demonstrates the stability of FMixFN and its potential for use on large-scale sample data.

CONCLUSION

We have developed a Bayes-based genomic selection model called FMixFN, which combines stable predictive ability and computational efficiency. Besides, when the number of individuals in the reference population is large, FMixFN is one



of the best choices for genomic selection. FMixFN is a stable, big data-oriented genomic selection model, which could meet the needs of large breeding companies or combined breeding schedules.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by the ethics committee of Jiangxi Agricultural University.

AUTHOR CONTRIBUTIONS

ZZ conceived and designed the experiments. WX and XL analyzed the data. SX, MZ, TY, and LH contributed to

materials and analysis tools. WX and ML wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31760656).

ACKNOWLEDGMENTS

We are grateful to all members who participated in this study from the State Key Laboratory for Pig Genetic Improvement and Production Technology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.721600/full#supplementary-material>

REFERENCES

- Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot Topic: a Unified Approach to Utilize Phenotypic, Full Pedigree, and Genomic Information for Genetic Evaluation of Holstein Final Score. *J. Dairy Sci.* 93 (2), 743–752. doi:10.3168/jds.2009-2730
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8
- Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the Performance of Risk Prediction Based on Polygenic Analyses of Genome-wide Association Studies. *Nat. Genet.* 45 (4), 400–405. doi:10.1038/ng.2579
- Chen, L., Li, C., Sargolzaei, M., and Schenkel, F. (2014). Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction. *PLoS One* 9 (7), e101544. doi:10.1371/journal.pone.0101544
- Cheng, H., Qu, L., Garrick, D. J., and Fernando, R. L. (2015). A Fast and Efficient Gibbs Sampler for BayesB in Whole-Genome Analyses. *Genet. Sel. Evol.* 47, 80. doi:10.1186/s12711-015-0157-x
- Christensen, O. F., and Lund, M. S. (2010). Genomic Prediction when Some Animals Are Not Genotyped. *Genet. Sel. Evol.* 42, 2. doi:10.1186/1297-9686-42-2
- Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-wide Approach. *PLoS One* 3 (10), e3395. doi:10.1371/journal.pone.0003395
- Ding, R., Yang, M., Quan, J., Li, S., Zhuang, Z., Zhou, S., et al. (2019). Single-Locus and Multi-Locus Genome-wide Association Studies for Intramuscular Fat in Duroc Pigs. *Front. Genet.* 10, 619. doi:10.3389/fgene.2019.00619
- Dong, L., Fang, M., and Wang, Z. (2017). Prediction of Genomic Breeding Values Using New Computing Strategies for the Implementation of MixP. *Sci. Rep.* 7 (1), 17200. doi:10.1038/s41598-017-17366-2
- Duchemin, S. I., Colombani, C., Legarra, A., Baloché, G., Larroque, H., Astruc, J.-M., et al. (2012). Genomic Selection in the French Lacaune Dairy Sheep Breed. *J. Dairy Sci.* 95 (5), 2723–2733. doi:10.3168/jds.2011-4980
- Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., et al. (2012). Improving Accuracy of Genomic Predictions within and between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels. *J. Dairy Sci.* 95 (7), 4114–4129. doi:10.3168/jds.2011-5019
- Goddard, M. E., and Hayes, B. J. (2009). Mapping Genes for Complex Traits in Domestic Animals and Their Use in Breeding Programmes. *Nat. Rev. Genet.* 10 (6), 381–391. doi:10.1038/nrg2575
- Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., and Hayes, B. J. (20161835). Genetics of Complex Traits: Prediction of Phenotype, Identification of Causal Polymorphisms and Genetic Architecture. *Proc. R. Soc. B.* 283, 20160569. doi:10.1098/rspb.2016.0569
- Goddard, M. (2009). Genomic Selection: Prediction of Accuracy and Maximisation of Long Term Response. *Genetica* 136 (2), 245–257. doi:10.1007/s10709-008-9308-0
- Grosfeld-Nir, A., Ronen, B., and Kozlovsky, N. (2007). The Pareto Managerial Principle: when Does it Apply? *Int. J. Prod. Res.* 45 (10), 2317–2325. doi:10.1080/00207540600818203
- Guo, Y., Mao, H., Ren, J., Yan, X., Duan, Y., Yang, G., et al. (2009). A Linkage Map of the Porcine Genome from a Large-Scale White Duroc × Erhualian Resource Population and Evaluation of Factors Affecting Recombination Rates. *Anim. Genet.* 40 (1), 47–52. doi:10.1111/j.1365-2052.2008.01802.x
- Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian Alphabet for Genomic Selection. *BMC Bioinformatics* 12, 186. doi:10.1186/1471-2105-12-186
- Ibáñez-Escriche, N., Forni, S., Noguera, J. L., and Varona, L. (2014). Genomic Information in Pig Breeding: Science Meets Industry Needs. *Livestock Sci.* 166, 94–100. doi:10.1016/j.livsci.2014.05.020
- Kemper, K. E., Reich, C. M., Bowman, P. J., Vander Jagt, C. J., Chamberlain, A. J., Mason, B. A., et al. (2015). Improved Precision of QTL Mapping Using a Nonlinear Bayesian Method in a Multi-Breed Population Leads to Greater Accuracy of Across-Breed Genomic Predictions. *Genet. Sel. Evol.* 47, 29. doi:10.1186/s12711-014-0074-4
- Legarra, A., Aguilar, I., and Misztal, I. (2009). A Relationship Matrix Including Full Pedigree and Genomic Information. *J. Dairy Sci.* 92 (9), 4656–4663. doi:10.3168/jds.2009-2061
- Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009). The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* 183 (3), 1119–1126. doi:10.1534/genetics.109.107391
- Mclachlan, G. J., and Basford, K. E. (1988). Mixture Models: Inference and Applications to Clustering. *J. R. Stat. Soc. Ser. A Stat. Soc.* 152 (1), 126–127. doi:10.2307/2982840
- Meuwissen, T. H. (2009). Accuracy of Breeding Values of 'unrelated' Individuals Predicted by Dense SNP Genotyping. *Genet. Sel. Evol.* 41, 35. doi:10.1186/1297-9686-41-35

- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157 (4), 1819–1829. doi:10.1093/genetics/157.4.1819
- Meuwissen, T. H., Solberg, T. R., Shepherd, R., and Woolliams, J. A. (2009). A Fast Algorithm for BayesB Type of Prediction of Genome-wide Estimates of Genetic Value. *Genet. Sel. Evol.* 41, 2. doi:10.1186/1297-9686-41-2
- Miszta, I. (2016). Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. *Genetics* 202 (2), 401–409. doi:10.1534/genetics.115.182089
- Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *Plos Genet.* 11 (4), e1004969. doi:10.1371/journal.pgen.1004969
- Mrode, R., Ojango, J. M. K., Okeyo, A. M., and Mwacharo, J. M. (2018). Genomic Selection and Use of Molecular Tools in Breeding Programs for Indigenous and Crossbred Cattle in Developing Countries: Current Status and Future Prospects. *Front. Genet.* 9, 694. doi:10.3389/fgene.2018.00694
- Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., et al. (2011). Distribution of Allele Frequencies and Effect Sizes and Their Interrelationships for Common Genetic Susceptibility Variants. *Proc. Natl. Acad. Sci.* 108 (44), 18026–18031. doi:10.1073/pnas.1114759108
- Pérez, P., and de los Campos, G. (2014). Genome-wide Regression and Prediction with the BGLR Statistical Package. *Genetics* 198 (2), 483–495. doi:10.1534/genetics.114.164442
- Pollak, E. J., Bennett, G. L., Snelling, W. M., Thallman, R. M., and Kuehn, L. A. (2012). Genomics and the Global Beef Cattle Industry. *Anim. Prod. Sci.* 52 (3), 92–99. doi:10.1071/an11120
- Preisinger, R. (2012). Genome-wide Selection in Poultry. *Anim. Prod. Sci.* 52, 121–125. doi:10.1071/an11071
- Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A., Amaral, A. J., Archibald, A. L., Beaver, J. E., et al. (2009). Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLoS One* 4 (8), e6524. doi:10.1371/journal.pone.0006524
- Samorè, A. B., and Fontanesi, L. (2016). Genomic Selection in Pigs: State of the Art and Perspectives. *Ital. J. Anim. Sci.* 15 (2), 211–232. doi:10.1080/1828051x.2016.1172034
- Sargolzaei, M., and Schenkel, F. S. (2009). QMSim: a Large-Scale Genome Simulator for Livestock. *Bioinformatics* 25 (5), 680–681. doi:10.1093/bioinformatics/btp045
- Silverman, B. W. (1996). Smoothed Functional Principal Components Analysis by Choice of Norm. *Ann. Stat.* 24 (1), 1–24. doi:10.1214/aos/1033066196
- VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980
- VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited Review: Reliability of Genomic Predictions for North American Holstein Bulls. *J. Dairy Sci.* 92 (1), 16–24. doi:10.3168/jds.2008-1514
- Verbyla, K. L., Bowman, P. J., Hayes, B. J., and Goddard, M. E. (2010). “Sensitivity of Genomic Selection to Using Different Prior Distributions,” BMC Proc. 4 Suppl. 1 in Proceedings of the 13th European workshop on QTL map), S5. doi:10.1186/1753-6561-4-S1-S5
- Wray, N. R., and Visscher, P. M. (2008). Estimating Trait Heritability. *Nat. Edu.* 1 (1), 29.
- Xavier, A., Muir, W. M., and Rainey, K. M. (2019). bWGR: Bayesian Whole-Genome Regression. *Bioinformatics* 24, btz794. doi:10.1093/bioinformatics/btz794
- Xu, S. (2010). An Expectation-Maximization Algorithm for the Lasso Estimation of Quantitative Trait Locus Effects. *Heredity* 105 (5), 483–494. doi:10.1038/hdy.2009.180
- Xu, S. (2003). Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics* 163 (2), 789–801. doi:10.1093/genetics/163.2.789
- Yi, N., and Banerjee, S. (2009). Hierarchical Generalized Linear Models for Multiple Quantitative Trait Locus Mapping. *Genetics* 181 (3), 1101–1113. doi:10.1534/genetics.108.099556
- Yu, X., and Meuwissen, T. H. (2011). Using the Pareto Principle in Genome-wide Breeding Value Estimation. *Genet. Sel. Evol.* 43, 35. doi:10.1186/1297-9686-43-35
- Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide Association Mapping Reveals a Rich Genetic Architecture of Complex Traits in *Oryza Sativa*. *Nat. Commun.* 2, 467. doi:10.1038/ncomms1467
- Zhou, X., and Stephens, M. (2012). Genome-wide Efficient Mixed-Model Analysis for Association Studies. *Nat. Genet.* 44 (7), 821–824. doi:10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Xu, Liu, Liao, Xiao, Zheng, Yao, Chen, Huang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Circular RNA Expression and Regulation Profiling in Testicular Tissues of Immature and Mature Wandong Cattle (*Bos taurus*)

Ibrar Muhammad Khan^{1†}, Hongyu Liu^{1*†}, Jingyi Zhuang¹, Nazir Muhammad Khan², Dandan Zhang¹, Jingmeng Chen¹, Tengting Xu¹, Lourdes Felicidad Córdova Avalos¹, Xinqi Zhou¹ and Yunhai Zhang^{1*}

¹Anhui Provincial Laboratory of Local Livestock and Poultry Genetical Resource Conservation and Breeding, College of Animal Science and Technology, Anhui Agricultural University, Hefei, China, ²Department of Zoology, University of Science and Technology, Bannu, Pakistan

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Olanrewaju B. Morenikeji,
University of Pittsburgh at Bradford,
United States
Ruihua Dang,
Northwest A&F University, China

*Correspondence:

Hongyu Liu
liuhongyu@ahau.edu.cn
Yunhai Zhang
yunhaizhang@ahau.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 25 March 2021

Accepted: 13 October 2021

Published: 22 November 2021

Citation:

Khan IM, Liu H, Zhuang J, Khan NM,
Zhang D, Chen J, Xu T, Avalos LFC,
Zhou X and Zhang Y (2021) Circular
RNA Expression and Regulation
Profiling in Testicular Tissues of
Immature and Mature Wandong Cattle
(*Bos taurus*).
Front. Genet. 12:685541.
doi: 10.3389/fgene.2021.685541

Wandong cattle are an autochthonous Chinese breed used extensively for beef production. The breed tolerates extreme weather conditions and raw feed and is resistant to tick-borne diseases. However, the genetic basis of testis development and sperm production as well as breeding management is not well established in local cattle. Therefore, improving the reproductive efficiency of bulls via genetic selection is crucial as a single bull can breed thousands of cows through artificial insemination (AI). Testis development and spermatogenesis are regulated by hundreds of genes and transcriptomes. However, circular RNAs (circRNAs) are the key players in many biological developmental processes that have not been methodically described and compared between immature and mature stages in *Bovine* testes. In this study, we performed total RNA-seq and comprehensively analyzed the circRNA expression profiling of the testis samples of six bulls at 3 years and 3 months of developmental age. In total, 17,013 circRNAs were identified, of which 681 circRNAs (p -adjust < 0.05) were differentially expressed (DE). Among these DE circRNAs, 579 were upregulated and 103 were downregulated in calf and bull testes. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses revealed that the identified target genes were classified into three broad functional categories, including biological process, cellular component, and molecular function, and were enriched in the lysine degradation, cell cycle, and cell adhesion molecule pathways. The binding interactions between DE circRNAs and microRNAs (miRNAs) were subsequently constructed using bioinformatics approaches. The source genes *ATM*, *CCNA1*, *GSK3B*, *KMT2C*, *KMT2E*, *NSD2*, *SUCLG2*, *QKI*, *HOMER1*, and *SNAP91* were found

Abbreviations: GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DE, differentially expressed; RT-qPCR, real-time quantitative polymerase chain reaction; FPKM, fragments per kilobase of transcript sequence per millions base pairs sequenced; CNCI, coding-non-coding index; CPC, coding potential calculator; nc-RNAs, noncoding RNAs; cds, coding sequence; BP, biological process; CC, cellular components; MF, molecular function; dNTP, deoxyribonucleotide triphosphate; dATP, deoxyadenosine triphosphate; dGTP, deoxyguanosine triphosphate; dUTP, deoxyuridine triphosphate; dCTP, deoxycytidine triphosphate.

to be actively associated with bull sexual maturity and spermatogenesis. In addition, a real-time quantitative polymerase chain reaction (RT-qPCR) analysis showed a strong correlation with the sequencing data. Moreover, the developed model of *Bovine* testes in the current study provides a suitable framework for understanding the mechanism of circRNAs in the development of testes and spermatogenesis.

Keywords: wandong cattle, testicular growth, spermatogenesis, circRNAs, total RNA sequencing

INTRODUCTION

The local Wandong cattle are viewed as one of the primary breeds in Anhui, China, and moreover, they are considered as one of the most economically important breeds in the beef industry. However, their development and utilization are still in the initial stages (Zhou et al., 2001). The reproductive efficiency of breeding bulls is an important consideration in livestock production systems, and the genetic mainstream is obtained through the selective application of germ cells from the desirable sire (Waqas et al., 2019). Hence, molecular strategies are required to improve the reproduction traits and to explore the genetic potential for economically important traits in local cattle breeds.

Mammalian spermatogenesis can be divided into three distinct stages: i) self-renewal and mitosis in spermatogonia to form spermatocytes, ii) meiosis in spermatocytes to generate elongated haploid spermatids, and iii) postmeiotic modifications in haploid spermatids to form spermatozoa (Hecht, 1998; Griswold, 2016; Ibtisham et al., 2017). The crucial stages of spermatogenesis are regulated at the transcriptional, post-transcriptional, and epigenetic levels by a well-coordinated genomics network (Eddy and O'Brien, 1997; Yu et al., 2003). Male gonads consist of complicated transcriptomic elements, with more than 15,000 differentially expressed genes (DEGs) involved in spermatogonial maturation (Ramsköld et al., 2009; Soumillon et al., 2013). Recent studies have demonstrated that noncoding RNAs (nc-RNAs) play an important role at the post-transcriptional level during spermatogenesis (Mukherjee et al., 2014; Robles et al., 2017; Gao et al., 2019).

Circular RNAs (circRNAs) are nc-RNAs with a circular closed loop and play a vital regulatory role in many biological processes. The circRNAs are resistant to RNase degradation because of the lack of 5'- and 3'-ends (Sanger et al., 1976; Lasda and Parker, 2014). In the modern era of bioinformatics and sequencing technology, a significantly high number of circRNAs have been revealed in the mouse brain and testis. It has been shown that the testis consists of the second highest number of tissue-specific circRNA-level genes (You et al., 2015). A total of 15,996 circRNAs have been identified in humans, of which 10,792 (67%) have been observed in the testis (Guo et al., 2014; Dong et al., 2016). Remarkably, a primary report has shown that 1,017 circRNA-linked genes are mostly related to spermatogenesis, sperm motility, and fertilization (Dong et al., 2016). However, little is known about the global profile and characteristics of circRNAs in mammalian testis development and spermatogenesis.

In the present study, we provided a comprehensive catalog of circRNAs in the testes of bulls at two different developmental stages and detailed new insights into the functional activities of circRNAs in testicular development and spermatogenesis. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomics (KEGG) analyses were performed to ascertain the molecular mechanisms of reproduction that were modulated during growth and development in bulls.

MATERIALS AND METHODS

Animals and Sample Collection

Three physically healthy individuals ($n = 3$) of two different age groups, andrologically mature bulls (3 ± 0.014 years) and immature calves (3 ± 0.24 months), were selected for the current study (Tomlinson et al., 2017). The native cattle are also known as Wandong in the Anhui province of China. All testicular samples were obtained immediately after the slaughtering process under the veterinary surgical protocol in Fengyang County, Anhui. All pairs of testes were processed by incising the scrotum medially and uncovering the right and left testicles within the tunica vaginalis (Lunstra and Echternkamp, 1982). After removing the fascia tissues, tunica vaginalis, and tunica albuginea from the testis, three slices of testicular samples were cross-cut from the middle aspect of the testis through fine-scale dissection. One cross-cut section was stored in 4% formaldehyde solution for histological assessment, and the others were immediately immersed in liquid nitrogen and stored until total RNA extraction (Wu et al., 2020).

Histological Assessment

Testicular tissue samples that had been preserved in 4% formaldehyde (Wuhan Servicebio Technology Co., Ltd.) for 72 h were used for histological sections (Fischer et al., 2008). The tissues were sectioned into 6- μ m-thick sections and stained with Masson's trichrome stain. The histomorphology of testicular tissues was assessed using different magnification powers.

RNA Isolation, Quantification, and Qualification

The testicular samples were sent to a commercial sequencing facility (Beijing Novogene Corporation, China) for Seq-analysis. Total RNA was isolated from the testis samples using the Trizol reagent (Takara, Beijing, China) under sterile conditions and treated with the DNase I enzyme to remove endogenous DNA

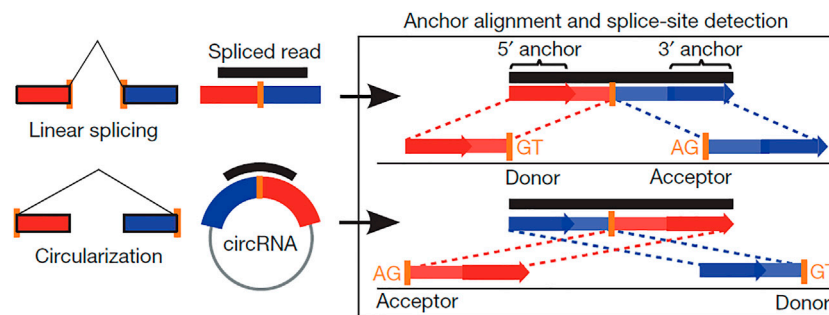


FIGURE 1 | Scheme showing the difference between linear and back-splicing and also illustrating the RNA reads' accommodation to each other that makes the stable circular RNAs.

contamination according to the manufacturer's protocol. RNA quality and contamination were assessed by 1% agarose gel electrophoresis. The purity was checked using a NanoPhotometer[®] spectrophotometer (IMPLEN, CA, United States). The total RNA concentration was assessed using a Qubit[®] RNA Assay Kit in a Qubit[®] 2.0 fluorometer (Life Technologies, CA, United States). RNA integrity was measured using an RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, United States).

cDNA Library Construction for circRNA Sequencing Analysis

A total of 3 µg of RNA from each testis sample was used to construct the cDNA libraries using the NEB-Next[®] Ultra-TM RNA Library Prep Kit for Illumina[®] (NEB, United States). Novogene constructed a chain-specific library by removing ribosomal RNA (Parkhomchuk et al., 2009). First, the ribosomal RNA was removed from the total RNA, and then, the RNA was divided into many short slices of 250–300 bp. The first cDNA strand manufactured from the fragmented RNA was used as the template strand, and random oligonucleotides were used as primers. Subsequently, the RNA strand was degraded with RNase H, and then, the second cDNA strand was manufactured along with dNTPs (dATP, dGTP, dUTP, and dCTP) using the DNA polymerase I system. After purification, the double-stranded cDNA was fixed, followed by the addition of a poly A-tail and sequencing adaptors. A cDNA of approximately 200 bp was separated using the AMPure XP beads. Finally, the USER enzyme was used to degrade the second (uracil-containing) strand of the cDNA, and PCR amplification was performed to obtain the library.

Alignment of RNA-Sequencing Reads and circRNA Identification

To ensure the quality and reliability of the data analysis, it is necessary to filter the original raw data. The low-quality reads (exceeding 50% low-quality bases, e.g., where Q-phred score ≤ 20), reads exceeding 10% possessing (N) nucleotides (where N,

the proportion of reads that cannot be determined, is greater than 0.002), and reads with seq-adaptors were removed to obtain the clean reads, which were aligned to the ribosomal RNA database using the Bowtie2 tool (Langmead and Salzberg, 2012). The ribosomal RNA-free reads of each calf and bull sample were charted to the reference genome (*Bos taurus*-UMD 3.1.1) by TopHat2 (v. 2.0.3.12) as described previously (Kim et al., 2013). The charted reads of each sample were assembled using String-Tie v1.3.1 (Pertea et al., 2016). Subsequently, 20-mer sequences were removed from both ends of the unmapped reads and aligned according to the reference genomes to identify the exclusive anchor positions within the splicing position. The anchor reads were aligned in the reverse head-to-tail direction that exhibited back splicing (Hansen et al., 2013) and led to key identification of circRNAs (Figure 1). The identified circRNAs were blast for the specific functions against circBase, a database for circRNAs (<http://www.circbase.org/>) (Glažar et al., 2014), and assigned to different features, such as genomic classification, length distribution, and chromosomal distribution. The circRNAs that did not find their annotations were considered novel.

Putative Identification and Coding Potential of circRNAs

Putative circRNAs were identified by filtration of anonymous transcripts. To reduce the rates of false positives, the assembled transcripts were filtrated to gain putative circRNAs by the following steps: i) the transcripts having more than one exon were selected, ii) transcripts having more than 200 bp in length and free from exon region overlapping were desired, and iii) transcripts with high expression levels where the FPKM value was more than 2 log₂ (fold change) were selected. All known transcripts of the database were analyzed using the Cuffcompare software. The coding potential of circRNAs was determined using the software programs Coding–Non-Coding Index (CNCI) (Sun et al., 2013), Pfam-scan (Bateman et al., 2004; Finn et al., 2014), and coding potential calculator (CPC) (Kong et al., 2007). All transcripts identified with coding potential were refined, and those without coding potential were recognized as putative circRNAs.

Expression Pattern of Differentially Expressed circRNAs and miRNA Endogenous Sponge

The circRNA gene expression level was assessed using the fragments per kilobase of transcript sequence per million base pairs sequenced (FPKM) value. The significance of DE circRNAs at the gene or transcript level was analyzed, and functional genes relevant to the developmental groups were identified. Cuffdiff (v2.1.1) software was used to project the TPM levels of circRNAs (Trapnell et al., 2010). An edgeR kit (<http://www.rproject.org/>) was used to classify differentially expressed circRNAs through samples. CircRNAs with a fold change greater than or equal to 2 and a p -value < 0.05 in comparison between the samples are differentially expressed. We identified the endogenous sponge interaction between the differentially expressed circRNA and miRNA by means of three software programs: miRanda (v. 3.3a), MIREAP, and TargetScan (v. 7.0).

Bioinformatics Analysis of GO and KEGG

The GO term analysis and classification provides significant source genes that are involved in different biological functions, and DE circRNA genes were employed by the GO-seq R package (Young et al., 2010). All source genes in the pathways and the background genes were charted to the GO terms in the database (<http://www.geneontology.org/>). The GO terms with the corrected p -value ($p < 0.05$) were recognized as relatively enriched by DE genes according to the definition of the hyper-geometric test. The significance analysis of the term enrichment was corrected by FDR, and the corrected p -value (Q-value) was obtained (Ashburner et al., 2000). The KEGG pathway-based study further elaborated the explanation of source genes and their biological functions (<http://www.genome.jp>). KOBAS software was used to test the enriched DE source genes in the KEGG pathways, as suggested by (Mao et al., 2005). The enrichment analysis significantly identified the signaling and metabolic pathways, to which circRNA source genes and total background genes were charted.

Validation of the circRNAs Gene via RT-qPCR

Whole RNA was extracted from immature and mature groups using the Trizol reagent (Life Technologies, 182805, United States), and its concentration was measured using a Nanodrop spectrophotometer. The good-quality RNAs were then converted into cDNAs using a QuantiTect Reverse Transcription Kit (Qiagen, 205311, Germany) in accordance with the manufacturer's protocol. The software programs Primer 3 web version 4.0.0 and basic local alignment search tool (BLAST; <https://blast.ncbi.nlm.nih>) were used to design gene-specific primers. The obtained PCR products were detected by 3% agarose gel and Sanger sequencing (Sangon Biotech, Shanghai, China). The divergent primers were used in this study and are presented in **Supplementary File 1**. The expression of circRNAs was detected by using a StepOnePlus Real-Time PCR System (Applied Biosystems, United States)

using the FastStart SYBR Green Master Mix (Roche, Germany). Each PCR comprised 7.5 μ l 2 \times SYBR Green PCR Master Mix, 1.5 μ l cDNA, 0.5 μ l each of reverse and forward primers, 4.7 μ l nuclease-free water, and 10.3 μ l ROX dye. The following q-PCR thermal cycles were carried out: i) pre-denaturation at 95°C for 10 min, ii) followed by 45 denaturation amplification cycles at 95°C for 15 s, iii) annealing at 60°C for 10 s, and iv) an extension cycle at 72°C for 20 s. The quantification cycle (Cq) of each target gene was normalized to that of the reference *GAPDH* gene. The q-PCR experiment was performed in triplicate to minimize the risk of error in the experiments. The Cq values were obtained and transferred to a Microsoft Excel sheet for further relative quantification analysis using the 2 $^{-\Delta\Delta C_t}$ method (Sherman and Lempicki, 2009).

Statistical Analysis

Data on circRNA transcripts of interest from immature and mature testicular tissues were analyzed using the student t -test (SPSS 17.0) and presented as mean \pm SEM. Before conducting the t -test, the data distribution and variances between the two groups were evaluated and found to be normal and homogeneous, respectively. Mean values were believed to be significantly different with $*p < 0.05$ and $**p < 0.01$.

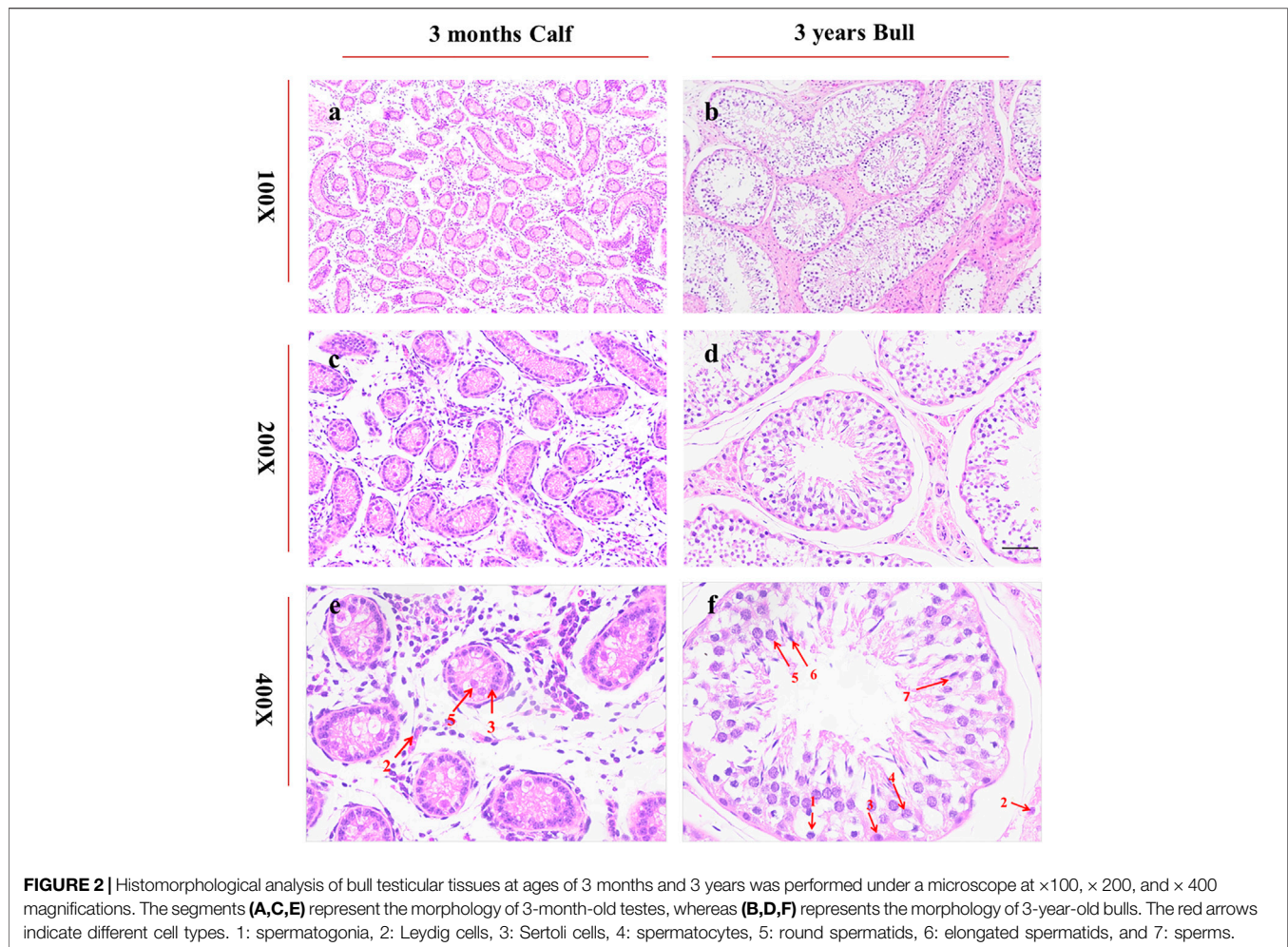
RESULTS

Histomorphology of Testes

A histological study of bull (*B. taurus*) testes showed a noteworthy difference between immature and mature stages. Under 100X microscopic examination, the diameter of seminiferous tubules was much smaller in immature than in mature testes. At the same time, similar conditions were observed in the interstitial connective tissue of immature and mature testes, as highlighted in **Figures 2A,B**. Under 400X microscopic examination, we obtained further details and observed more developmental stages of biological spermatozoa in mature than in immature testes. Sertoli cells were significantly increased in number and were found very near the basement membrane of seminiferous tubules in mature testes, as shown in **Figures 2E,F**.

Transcriptome Sequencing of circRNAs in Immature and Mature Bull Testes

A total of six experimental animals, three immature (3 ± 0.014 months) and three mature (3 ± 0.24 years), were selected from a local cattle station, and their testis samples were collected. To investigate the circRNA profiles of the immature and mature testis tissues, we prepared the RNA-seq libraries of all tissue samples. The raw reads of sequencing data were acquired using an Illumina HiSeq 4000 platform (Illumina, Inc, San Diego, CA, United States). Raw reads were filtered and classified into different segments, including clean reads (70,333,218, 98.95%), N-containing reads (38,769, 0.05%), low-quality reads (126,760, 0.18%), and adapter-related reads (584,096, 0.82%) (**Figure 3**). The raw reads for each sample ranged between 118,785,372 and



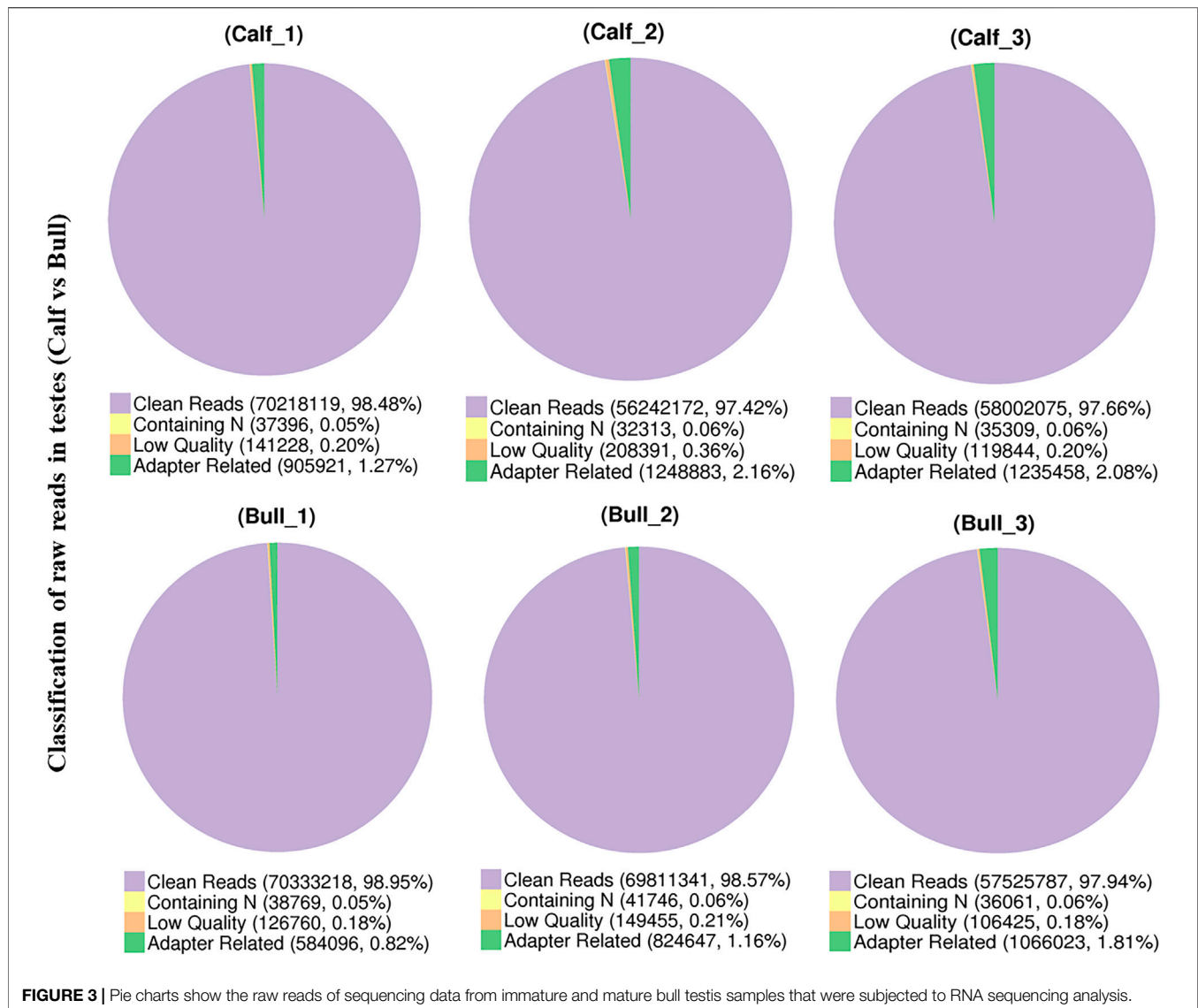
142,165,686, while the clean reads per sample ranged between 116,004,150 and 140,666,436. Thus, the raw and clean reads together yielded 90 GB of data, and the GC content ranged from 47.09 to 53.57% (**Supplementary File 2**). All Q20 values of the subjected reads in the six samples exceeded 97%, as shown in **Supplementary File 3**. More than 95.5% of the clean reads were coordinated to *B. taurus* UMD3.1.

Correlation and Differential Analysis Between the Testis Samples

The testis samples were collected from three calves (Calf1, Calf2, and Calf3) and three bulls (Bull1, Bull2, and Bull3). The Pearson's correlation heatmap was constructed according to the expression profiles for samples' quality control. All samples in the calf and bull groups showed the linear correlation (**Figure 4A**). The differences of transcripts of two groups demonstrated in the sample are shown in the PCA 3D map (**Figure 4B**). We also discovered that bull and calf groups have noticeable differences, whereas the testis samples of calves and bulls clustered separately and showed the difference.

Identification and Characterization of circRNAs

A total of 17,013 putative circRNAs were recognized with at least one read spanning and head-to-tail splicing in immature and mature testes. Based on their genomic location and features, circRNAs were divided into three subclasses, denoted as circ-exon (79.0%), circ-intergenic (14.0%), and circ-intron (7.0%) (**Figure 5A**). The total length plot quantile of circRNAs (approximately 75%) was not more than 1,000 nt, and the average length was 400 nt, as shown in **Figure 5B**. According to their host gene site, all the circRNAs were widely distributed on all sets of chromosomes, while none of the circRNA was mapped to the mitochondrial genome. The total available set of chromosomes generated 100 circRNAs in both developmental stages of the *Bovine* testis (**Figure 5C**). The coding potential of circRNAs was analyzed using the software programs CPC2, CNCI, and PFAM, and 6,011 coding circRNAs were identified in all samples (**Figure 5D**). Detailed information on the top 40 up- and downcoding circRNAs is listed in **Table 1**.



Differentially Expressed Gene and circRNAs in Testicular Tissues

After the quantification process, the expression pattern of circRNAs was identified using Cuffdiff and Ballgown tools. A total of 681 DE circRNAs at the gene level were detected in immature and mature stages. The expression levels of circRNAs in the testis samples were calculated, taking into account the parameter of significance, whereas the log₂ fold change was considered higher than or equal to 2 and *p*-adjusted < 0.05. According to the significance criteria, we detected 578 upregulated and 102 downregulated circRNAs between bull and calf testis samples. These up and down highlights of DE genes are displayed in volcanic plots (Figures 6A,B), and the list of total and DE circRNAs is shown in Supplementary Files 4–6. We also analyzed the DE circRNA by the hierarchical clustering method, which is another way to display DE genes and to cluster all the genes with similar expression patterns that may have common functions in metabolic and signaling pathways. The gene clusters on the left side were formed because of similar

expression patterns (fold change >2, *p* < 0.05); the columns presented calves and bulls, and the expression from blue to red was gradually upregulated (Figure 6C).

GO and Enriched KEGG Pathways Analysis

CircRNAs are closed-loop structures that are most commonly generated by back-splicing events between the exons of coding genes. In a limited way, circRNAs regulate the host gene and thus perform the desired function at the genetic level. Hence, the analysis of GO terms and circRNA source gene functions might provide new insights. Many GO classification terms were significantly enriched in the source genes of circRNAs. A list of 26,815 source genes significantly associated with DE circRNAs was submitted to the database for functional annotation and GO classification analysis, resulting in the identification of 50 significant GO terms, as shown in Figure 7. The functional annotation study revealed that several source genes actively participate in spermatogenesis, sperm morphology,

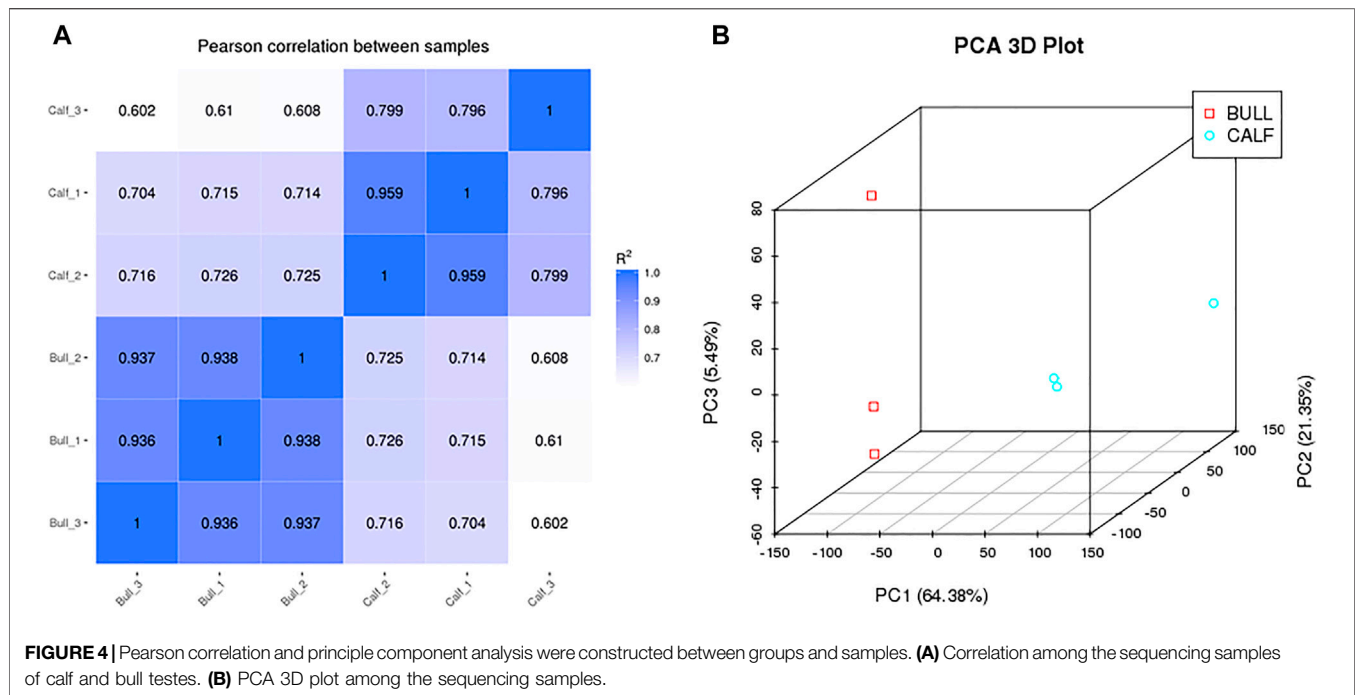


FIGURE 4 | Pearson correlation and principle component analysis were constructed between groups and samples. **(A)** Correlation among the sequencing samples of calf and bull testes. **(B)** PCA 3D plot among the sequencing samples.

developmental stages, and reproduction. The functional classification of source genes showed that their products were assigned to the three GO categories of biological processes, cellular components, and molecular functions (**Supplementary File 7**). To further understand the role of circRNA host genes in the developmental stages of the *Bovine*, a pathway analysis was applied using the KEGG pathway database and a total of 147 signaling pathways linked to circRNA gene products were identified (**Supplementary File 8**). The top 20 statistically enriched pathways ($p < 0.05$) were subjected to further analysis and were found to be significantly associated with signaling pathways such as lysine degradation, cell cycle, propanoate metabolism, adherens junction, and cell adhesion molecules (**Figure 8**). The host genes such as *KMT2E*, *EHMT1*, and *NSD2* were mediated by the lysine degradation signaling pathways but actively participated in the reproductive traits of the *Bovine*. Similarly, *ATM*, *GSK3B*, and *CCNA1* are host genes of the cell cycle signaling pathway and influence the biological descriptions relevant to spermatogenesis and testis development. The most significantly enriched pathways and DEGs are listed in **Table 2**.

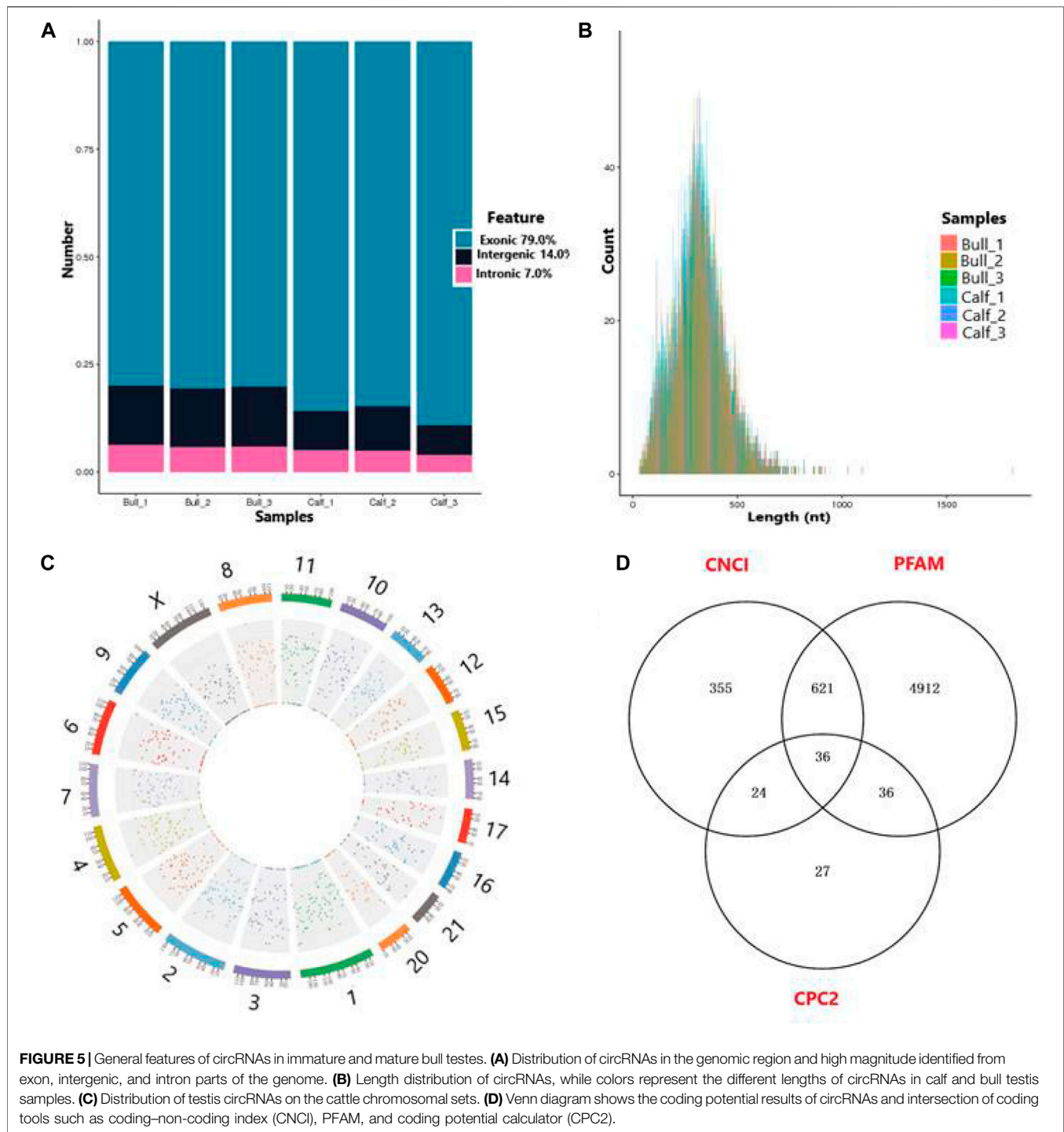
Prediction of Differentially Expressed circRNA and miRNA Endogenous Sponges

Some studies in recent years have suggested that miRNAs play a role in all aspects of spermatogenesis and testis formation; thus, miRNAs could be used as biomarkers for reproductive traits (Yadav and Kotaja, 2014). Reports also suggested that circRNAs act as miRNA sponges and play an important role in regulating gene expression in posttranscriptional processes (Kulcheski et al., 2016). We analyzed further to determine whether circRNAs

found in immature and mature testes act as endogenous miRNA sponges. We noticed that a total of 4,300 circRNAs have potential binding miRNA sites, but other circRNAs were predicted to have no possible binding targets for miRNAs. Each circRNA may attach to one or more targeted miRNAs. As a result, the percentage of circRNAs consisting of different numbers of miRNA targets was further tested. The majority of circRNAs have at least two miRNA-binding sites; here, the proportion of circRNAs containing 6–10 miRNA targets is the largest, as illustrated in **Figure 9A**. The binding interactions between selected DE circRNAs in testicular tissues and their predicted miRNA targets are shown in **Figure 9B**. Recent studies have proposed that circRNAs act as miRNA sponges and play a critical role in regulating gene expression in pathways through complex (circRNAs–miRNAs genes) chains (Hansen et al., 2013; Memczak et al., 2013). Approximately 758 miRNAs have been predicted, and many of them, such as bta-miR-204, bta-miR-532, and bta-miR-34 groups, have been correlated with spermatogenesis (Zhang et al., 2015).

Validation of DE circRNA-Seq Results by RT-qPCR

To justify the sequencing analysis of DE circRNAs through RT-qPCR, we selected the truly significant and exonic DE circRNAs for validation. We randomly selected nine circRNAs of different nucleotide richness values and lengths, and divergent primers were prepared to verify the expression levels of circRNAs in the testis. Ten host genes were proposed for circRNA detection, of which seven, including ATM serine/threonine kinase (*ATM*), cyclin A1 (*CCNA1*), glycogen synthase kinase 3 beta (*GSK3B*), lysine methyltransferase 2C (*KMT2C*), lysine methyltransferase 2E



(*KMT2E*), nuclear receptor binding SET domain protein 2 (*NSD2*), and succinate-CoA ligase GDP-forming subunit beta (*SUCLG2*), were upregulated. The remaining three genes, including QKI, KH domain-containing RNA binding (*QKI*), homer scaffold protein 1 (*HOMER1*), and synaptosome-associated protein 91 (*SNAP91*), were downregulated. The RT-qPCR analysis validated the expression of all nine circRNAs, which were found to regulate the functions of the aforementioned genes, as shown in **Figures**

10A,B. The gel electrophoresis findings showed that the PCR products of circRNA were single bands, which designated the presence of specific circRNAs, **Figure 10C**. Additionally, the back-splice junction of circRNAs was highlighted by further Sanger sequencing of the PCR products, **Figure 10D**. These genes work in specific reproductive-related pathways, such as lysine degradation, cell adhesion molecules (CAMs), focal adhesion, cell cycles, and adherens junctions, thus supporting

TABLE 1 | Top 40 regulated coding circular RNAs in immature and mature testicular tissues of Wandong cattle.

Novel circRNA Id	Chromosome	Total cds length	Classification	Source gene	Fold change	<i>P</i> _{adj} value
Top 20 novel upregulated coding circRNAs						
novel_circ_0024958	Chr 4	608	Exonic	<i>KMT2E</i>	3.2126	0.008797
novel_circ_0001818	Chr 11	325	Exonic	<i>EHMT1</i>	7.1764	0.014192
novel_circ_0027596	Chr 6	308	Exonic	<i>NSD2</i>	7.4895	0.009261
novel_circ_0006800	Chr 15	596	Exonic	<i>ATM</i>	2.0689	0.028979
novel_circ_0005451	Chr 13	402	Exonic	<i>RBLI</i>	3.0064	0.014784
novel_circ_0012012935	Chr 1	397	Exonic	<i>GSK3B</i>	3.206	0.026266
novel_circ_0029602	Chr 7	477	Exonic	<i>CDC25C</i>	4.0429	0.001681
novel_circ_0026085	Chr 5	506	Exonic	<i>SMC1B</i>	4.375	0.013226
novel_circ_0024710	Chr 4	402	Exonic	<i>DBF4</i>	3.816	9.32E-07
novel_circ_0003626	Chr 12	451	Exonic	<i>CCNA1</i>	8.2913	0.001968
novel_circ_0031633	Chr 9	436	Exonic	<i>AFDN</i>	2.4391	0.04536
novel_circ_0004558	Chr 13	630	Exonic	<i>PARD3</i>	2.8041	0.007949
novel_circ_0006892	Chr 15	594	Exonic	<i>CADMI</i>	7.4286	0.009651
novel_circ_0017085	Chr 24	546	Exonic	<i>LAMA3</i>	7.0164	0.07625
novel_circ_0009598	Chr 17	359	Exonic	<i>PXN</i>	7.1673	0.014472
novel_circ_0015698	Chr 22	431	Exonic	<i>SUCLG2</i>	4.4091	0.024862
novel_circ_0018225	Chr 25	270	Exonic	<i>ABAT</i>	6.456	0.038294
novel_circ_0029184	Chr 7	307	Exonic	<i>TJP3</i>	6.9245	0.020237
novel_circ_0031633	Chr 9	436	Exonic	<i>AFDN</i>	2.4391	0.04536
novel_circ_0024833	Chr 4	592	Exonic	<i>HGF</i>	7.3916	0.010582
Top 20 novel downregulated coding circRNAs						
novel_circ_0025557	Chr 4	655	Exonic	<i>AASS</i>	-4.2615	0.04536
novel_circ_0023110	Chr 3	272	Exonic	<i>VAV3</i>	-7.4813	0.010743
novel_circ_0024329	Chr 5	455	Exonic	<i>KMT2C</i>	-3.0993	0.031917
novel_circ_0033406	Chr X	399	Exonic	<i>TMLHE</i>	-3.0993	0.016812
novel_circ_0012012847	Chr 1	906	Exonic	<i>NECTIN3</i>	-2.9046	0.016725
novel_circ_0012012395	Chr 1	561	Exonic	<i>NCAM2</i>	-4.3218	0.009566
novel_circ_0023110	Chr 3	272	Exonic	<i>VAV3</i>	-7.4813	0.010743
novel_circ_0000398	Chr 10	607	Exonic	<i>NEOI</i>	-3.4837	0.031452
novel_circ_0007021	Chr 15	298	Exonic	<i>RRAS2</i>	-2.5529	0.001203
novel_circ_0031262	Chr 8	618	Exonic	<i>GKAPI</i>	-9.4111	0.000179
novel_circ_0022366	Chr 2	305	Exonic	<i>UNC80</i>	-5.385	0.004455
novel_circ_0032345	Chr 9	368	Exonic	<i>ZNF292</i>	-2.1687	0.00742
novel_circ_0024463	Chr 4	546	Exonic	<i>DYNC11</i>	-4.7909	0.007542
novel_circ_001012674	Chr 1	319	Exonic	<i>ST3GAL6</i>	-3.4838	0.007765
novel_circ_0012012395	Chr 1	561	Exonic	<i>NCAM2</i>	-4.3218	0.009566
novel_circ_0028513	Chr 6	164	Exonic	<i>EPHA5</i>	-7.5188	0.00975
novel_circ_0031677	Chr 9	696	Exonic	<i>RIMS1</i>	-3.1598	0.010582
novel_circ_0020402	Chr 29	187	Exonic	<i>MYRF</i>	-7.506	0.010879
novel_circ_0014496	Chr 21	264	Exonic	<i>SH3GL3</i>	-3.4807	0.002134
novel_circ_0026164	Chr 5	468	Exonic	<i>TMTC2</i>	-3.8141	0.012555

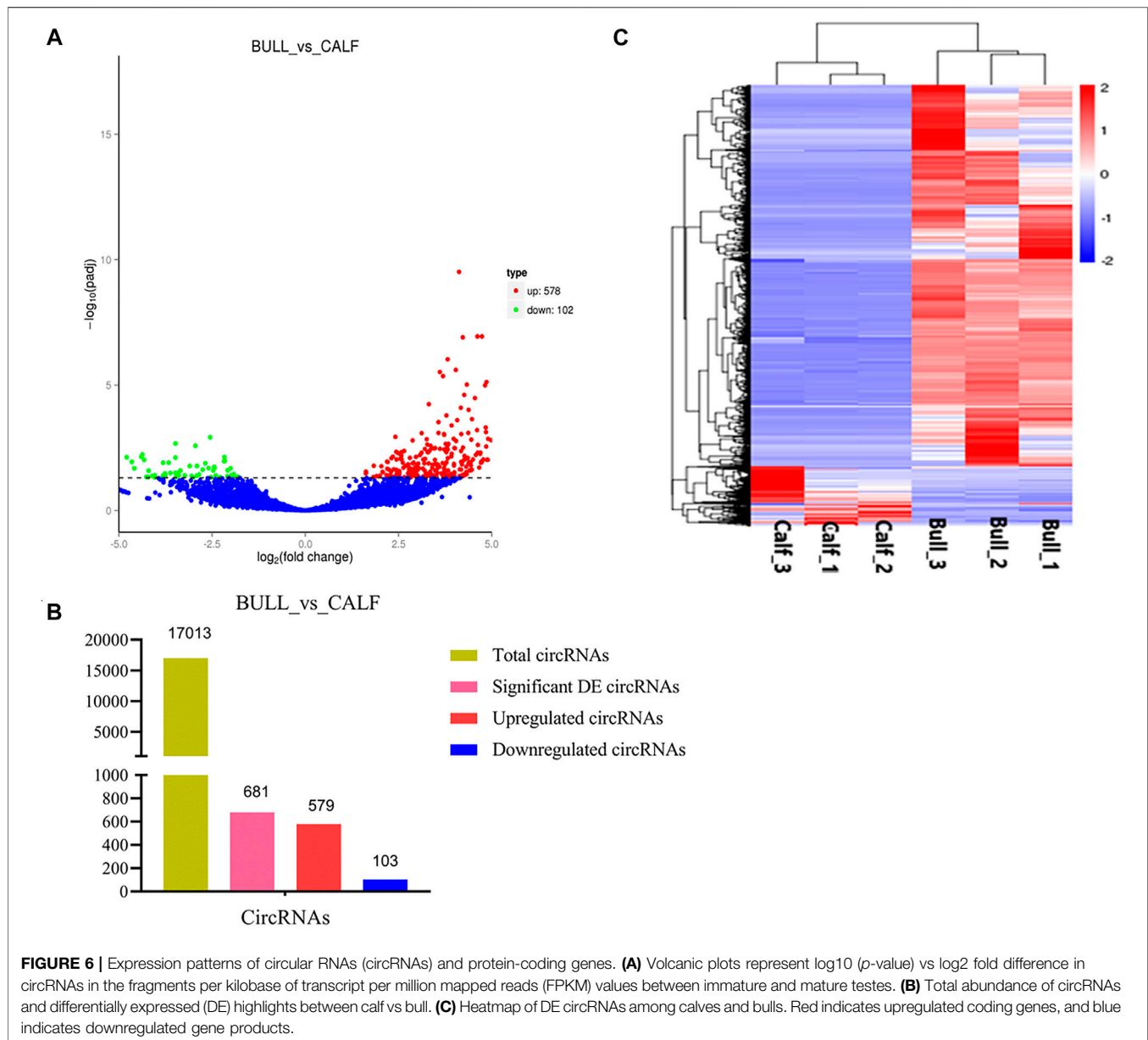
reproductive functions in animals. Therefore, the obtained results confirmed the accuracy of the ribo-depleted RNA-seq results.

DISCUSSION

Mammalian testicular growth and spermatogenesis are complex biological transformation processes that are regulated by a strong combination of coding and noncoding genes and transcriptomes (Card et al., 2013; Djureinovic et al., 2014). Testicular growth includes the growth of testicular tissues during the embryonic and postnatal phases (Svingen and Koopman, 2013). The bull testes first gradually expand until approximately 25 weeks of

growth and then rapidly expand until adolescence at 37–50 weeks of age (Rawlings et al., 2008). To analyze the molecular biology of testis development and reproductive transformation in *Bovine* species, we selected $n = 3$ animals of two different age groups: 3-month-old (immature) calves and 3-year-old (adult) bulls of a local Wandong cattle breed. Gao et al. (2018) tested the similar experimental design in the Qinchuan cattle testis and selected $n = 1$ animals of two different age groups: neonatal (1 week) and mature bulls (4 years).

Previously, nc-RNAs were regarded as useless stocks or junk RNAs. However, various types of ncRNAs, such as miRNAs, PIWI-interacting RNAs (pi-RNAs), long noncoding RNAs (lnc-RNAs), and circRNAs (Lee et al., 1993; Girard et al., 2006;



Memczak et al., 2013; Ulitsky and Bartel, 2013), have been identified and reported to play important roles in cellular and tissue development. In recent years, research has focused on the functional features of nc-RNAs, and much work has been reported specifically in animal reproduction (Svingen and Koopman, 2013). In addition, lnc-RNAs play a regulatory role in the testis and are expressed more strongly in mammalian testes than in other organs (Soumillon et al., 2013). While circRNAs are essential members of the nc-RNA family, they are also known to be involved in animal reproduction and disease control (Wang et al., 2016). Limited studies have characterized circRNAs in reproductive organs, such as placenta, embryos, and testes, and in cell types including oocytes, granulosa cells, immature sperm cells, seminal plasma cells, and mature sperm cells (Dong et al., 2016; Quan and Li, 2018; Chioccarelli et al., 2019). circRNAs

were identified between the good and poor sperm subpopulations in terms of kinetic parameters and morphological characteristics. These comparative data highlighted 148 DE circRNAs between the two semen quality classes (Dong et al., 2016). Previous studies have reported that circRNAs are the primary regulators of multiple biological processes, but they have not been methodically explored in the *Bovine* testis during reproductive development. In this study, we identified and marked circRNAs in *Bovine* testes and discovered possible genes that determine the development of testes.

A total of 17,013 and 681 coding circRNAs were identified in the *Bovine* testes at two different immature and mature stages. You et al. (2015) reported that the second highest number of circRNAs was found in the testes compared to other body organs.

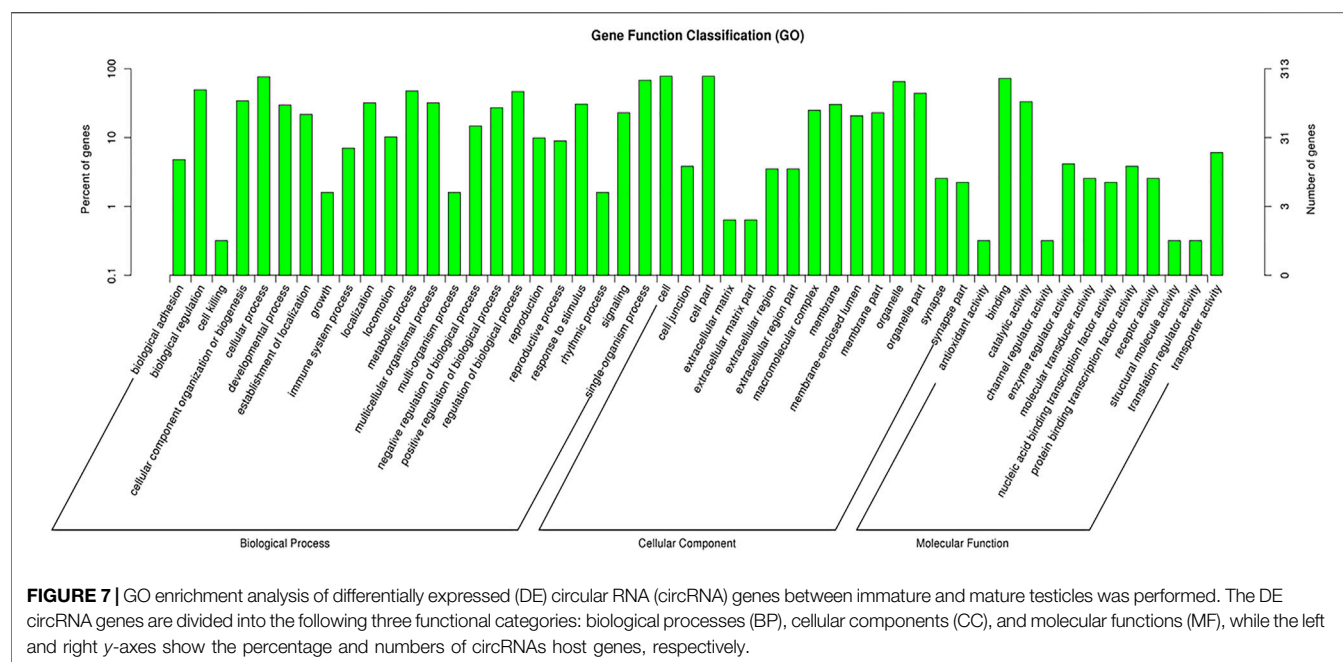


FIGURE 7 | GO enrichment analysis of differentially expressed (DE) circular RNA (circRNA) genes between immature and mature testicles was performed. The DE circRNA genes are divided into the following three functional categories: biological processes (BP), cellular components (CC), and molecular functions (MF), while the left and right y-axes show the percentage and numbers of circRNAs host genes, respectively.

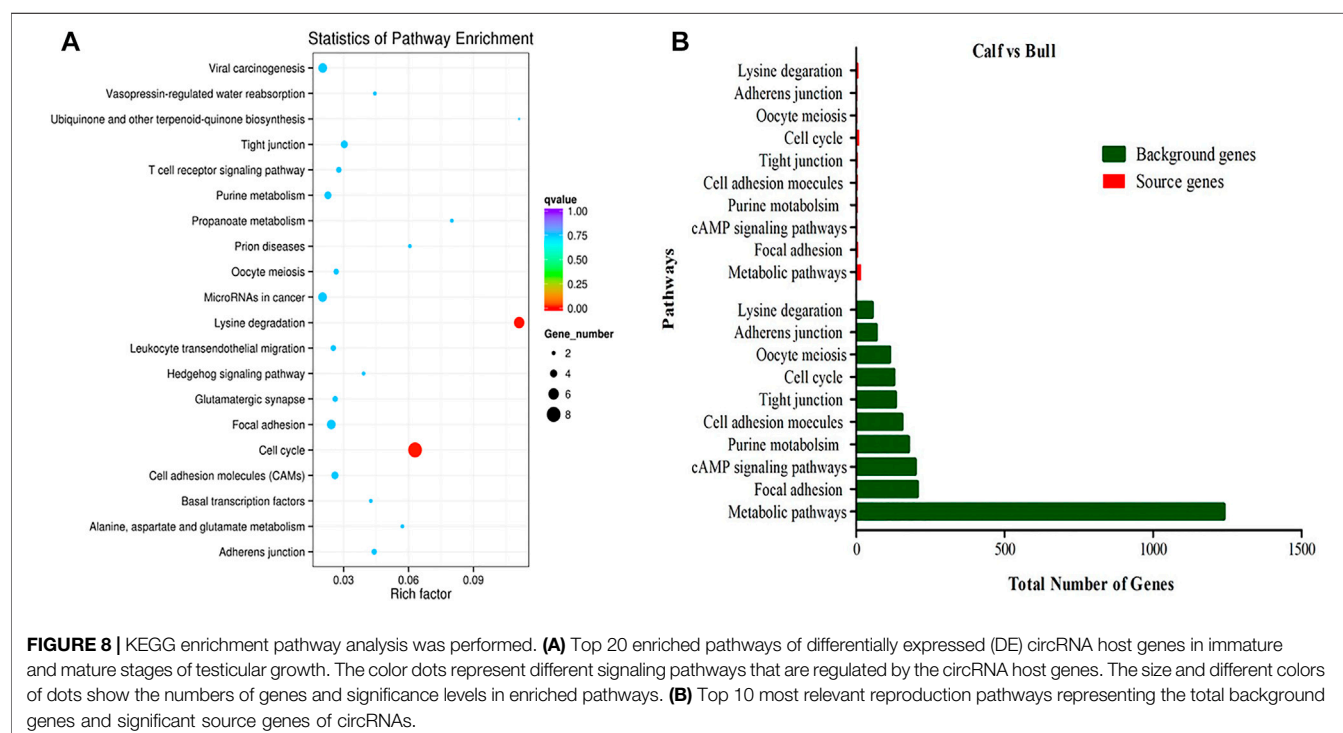


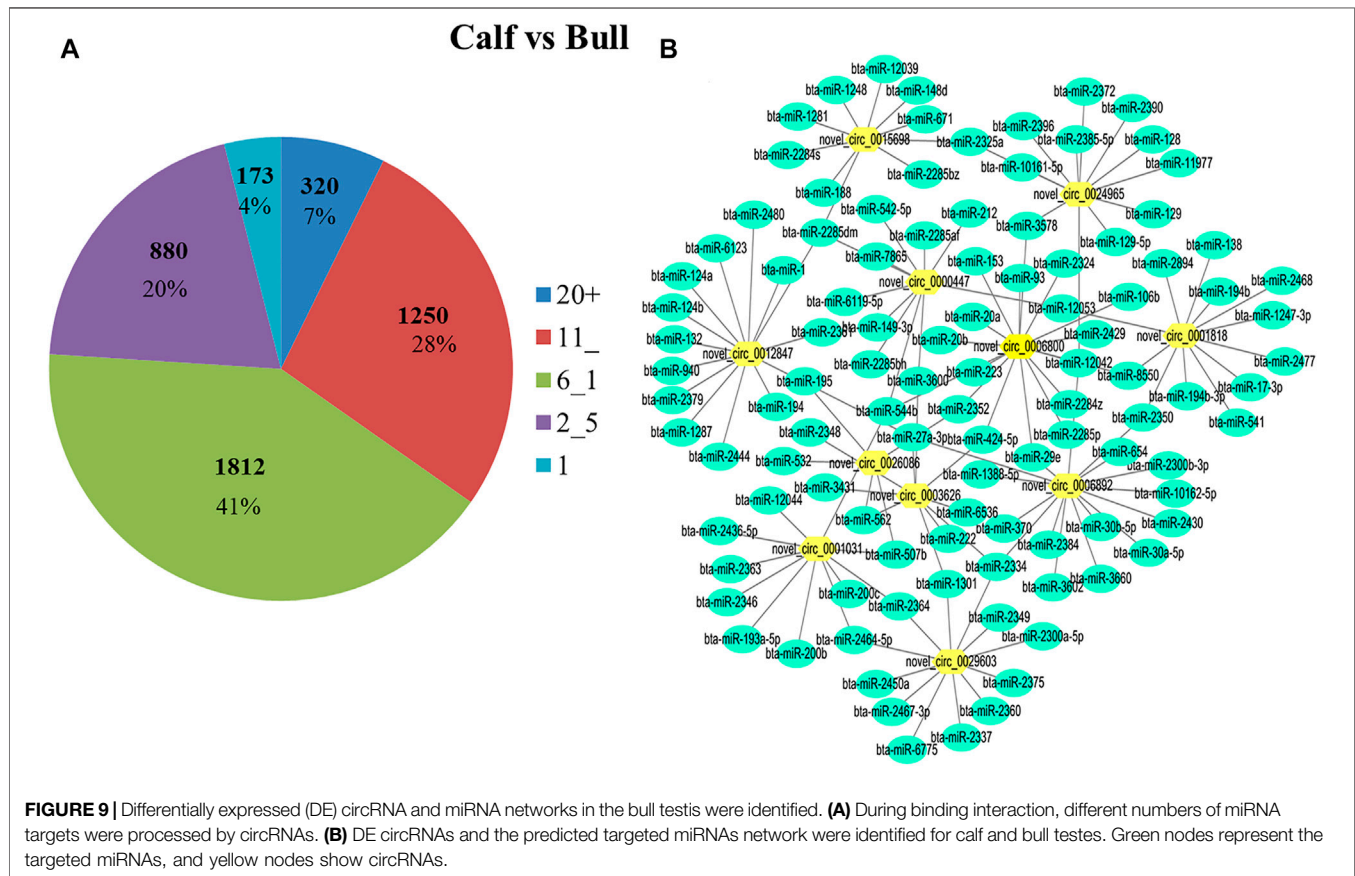
FIGURE 8 | KEGG enrichment pathway analysis was performed. **(A)** Top 20 enriched pathways of differentially expressed (DE) circRNA host genes in immature and mature stages of testicular growth. The color dots represent different signaling pathways that are regulated by the circRNA host genes. The size and different colors of dots show the numbers of genes and significance levels in enriched pathways. **(B)** Top 10 most relevant reproduction pathways representing the total background genes and significant source genes of circRNAs.

The magnitude of circRNAs in testicular samples was significantly greater than that of circRNAs found in other organs, for example, chicken theca cells contained 14,502 circRNAs (Shen et al., 2020), pig pituitaries had 5,148 circRNAs (Ulitsky and Bartel, 2013; Chen et al., 2020), and buffalo fat adipose tissues had 5,141 circRNAs (Huang et al., 2019). We identified 681 coding circRNAs, of which 579 circRNAs were upregulated and 103 were downregulated in

immature and mature bull testes. The study composed of Qinchuan cattle identified a total of 21,753 candidate circRNAs, where 2023 circRNAs were downregulated and 2,225 circRNAs were upregulated (Gao et al., 2018). A total of 2,326 differentially expressed circRNAs (DECs) were found in testicular development, of which 1,003 circRNAs were upregulated in adult boar testes and 1,323 circRNAs were downregulated; an analysis of transcriptomic changes in boar

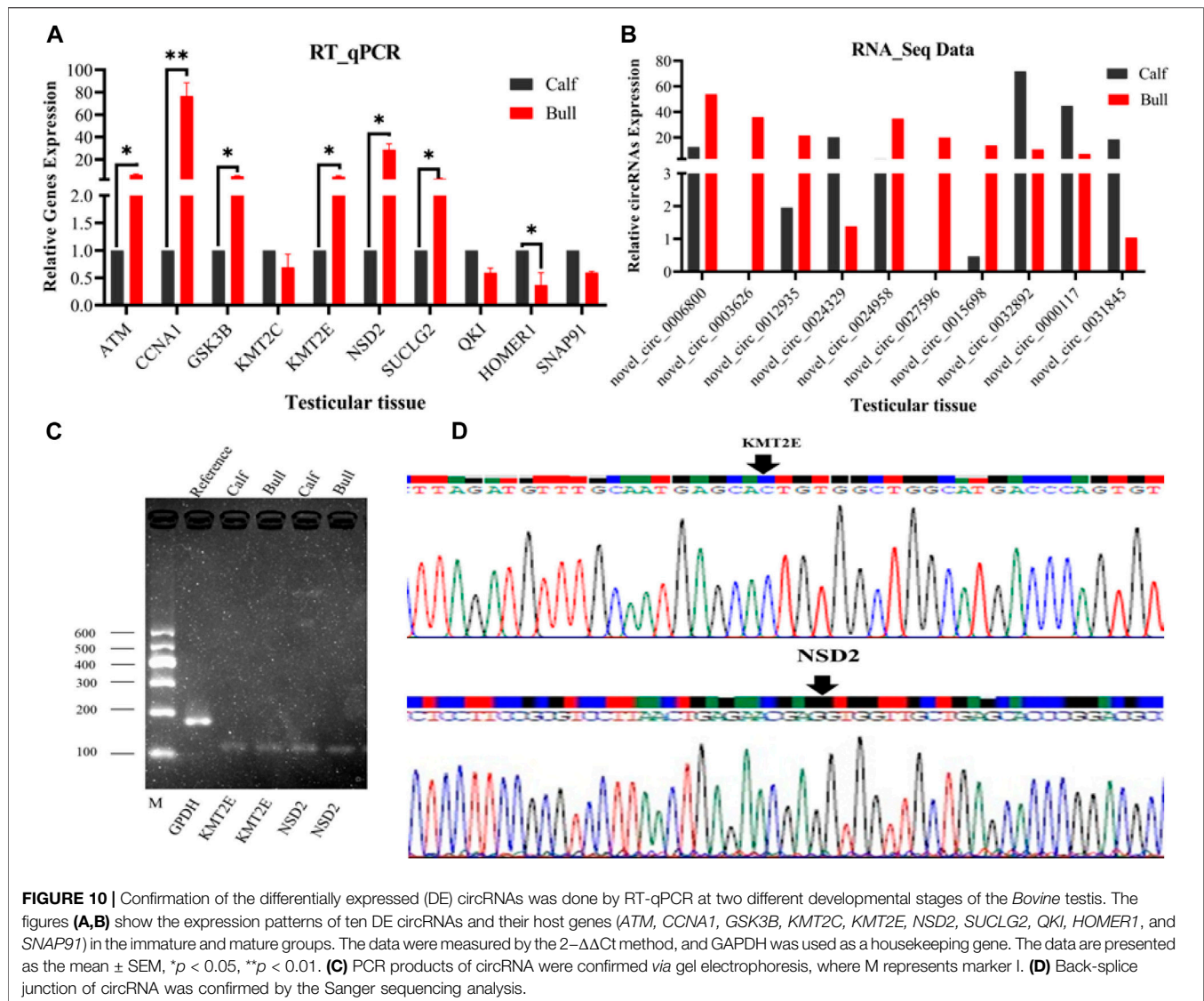
TABLE 2 | Most enriched biological pathway terms and regulated DE genes in immature and mature testes.

Pathway terms	Rich factor	q-value	Gene number	Gene name
Lysine degradation	0.111111111	0.006486719	6	<i>AASS,KMT2C,KMT2E,EHMT1,TMLHE, NSD2</i>
Cell cycle	0.062992126	0.007955094	8	<i>ATM,RBL1,GSK3B,CDC25C, SMC1B, DBF4,CCNA1</i>
Propanoate metabolism	0.08	0.753456013	2	<i>SUCLG2,ABAT</i>
Adherens junction	0.044117647	0.753456013	3	<i>NECTIN3,AFDN,PARD3</i>
Prion diseases	0.060606061	0.753456013	2	<i>NCAM2,C5</i>
Tight junction	0.03030303	0.753456013	4	<i>TJP3,RRAS2,AFDN,PARD3</i>
Alanine, aspartate, and glutamate metabolism	0.057142857	0.753456013	2	<i>GLS,ABAT</i>
Focal adhesion	0.024271845	0.753456013	5	<i>VAV3,LAMA3 LAMA3, GSK3B, PXN</i>
Vasopressin-regulated water reabsorption	0.044444444	0.753456013	2	<i>AQP4,DYNC1I1</i>
Cell adhesion molecules (CAMs)	0.025974026	0.753456013	4	<i>NECTIN3,NCAM2,NEO1,CADM1</i>
Basal transcription factors	0.042553191	0.753456013	2	<i>GTF2A1,TAF2</i>
Hedgehog signaling pathway	0.039215686	0.753456013	2	<i>GSK3B,CSNK1G3</i>
T-cell receptor signaling pathway	0.027777778	0.753456013	3	<i>GSK3B, DLG1, VAV3</i>
Oocyte meiosis	0.026548673	0.753456013	3	<i>SMC1B, CDC25C, FBX O 43</i>
Purine metabolism	0.022727273	0.753456013	4	<i>PDE10A, FHIT, PRIM2,AK8</i>
Glutamatergic synapse	0.026086957	0.753456013	3	<i>GLS, HOMER1, DLGAP1</i>
Viral carcinogenesis	0.020325203	0.753456013	5	<i>DLG1, CCNA1, GTF2A1, PXN, RBL1</i>
MicroRNAs in cancer	0.020242915	0.753456013	5	<i>ZFPM2, ATM, GLS, CDC25C, EZH2</i>
Leukocyte transendothelial migration	0.025210084	0.753456013	3	<i>VAV3, AFDN, PXN</i>



testicular tissues revealed an agreement with our results (Zhang et al., 2021). A study of transcriptomic changes in cattle-yak testes showed a tendency toward upregulation rather than downregulation and identified 679 upregulated and 2,281 downregulated genes (Cai et al., 2017). An RNA-seq analysis showed that a total of 10,095

genes were substantially differentially expressed in porcine testes, of which 5,199 were upregulated and 4,896 were downregulated in immature and mature porcine testes (Song et al., 2015). Approximately 242 genes were associated with porcine spermatogenesis. The key reason for testis transcriptome profiling



is to discover the transcriptional factors that play crucial roles in testis development and spermatogenesis and thus clarify the genetic structure of the testis.

We performed GO annotation and KEGG pathway enrichment analysis to better describe the circRNAs involved in the biological processes of testis development and spermatogenesis. These processes may indicate the roles of circRNAs in the testis or reflect the cellular differences between immature and mature bull testes. In GO annotation, 60 host genes were assigned to spermatogenesis processes, 12 genes to testis development, and 23 host genes were found solely relevant to reproduction phenomena, such as behavior and primary sexual characteristics. Gao et al. (2018) executed the GO annotation, and 44 host genes were assigned to reproduction and the reproductive process, and 25 host genes were associated with spermatogenesis, including *PIWIL1* and the spermatogenesis-associated protein 6 (*SPATA6*). We found different candidate genes by analyzing these GO terms, and some of these genes, such as *ATM* serine/

threonine kinase (*ATM*), glycogen synthase kinase 3 beta (*GSK3B*), and cyclin A1 (*CCNA1*), were found to be significantly enriched in the cell cycle pathway and associated with spermatozoa differentiation. The *ATM* protein is associated with low-fertile buffalo bull spermatozoa (M. K et al., 2019); in mouse spermatocytes, the meiotic recombination is initiated by SPO11-induced double-strand breaks (*DSBs*), and *ATM* controls this process (Huang et al., 2019; Paiano et al., 2020). *GSK3B* showed significant expression changes across these pubertal stages in the hypothalamus of piglets and is also involved in crucial biological processes that regulate economically relevant traits associated with beef cattle fertility (Fonseca et al., 2018). The GO and KEGG pathway enrichment analyses showed that the DE circRNA genes were associated with pathways, such as the cell cycle, lysine degradation, and tight junction, whereas the top 20 most significant enriched pathways were evaluated in this study. A total of 47 significant signaling pathways in the Qinchuan cattle testis were found, and key host genes including *PIWIL1*, *DPY19L2*,

SLC26A8, *IFT81*, *SMC1B*, *IQCG*, *TTL5*, and *ACVR2A* belong to the tight junction, the adherens junction, the TGF β signaling pathway, progesterone-mediated oocyte maturation, and oocyte meiosis (Gao et al., 2018).

The probable reasons of variation between the findings of the present study and the results of Gao et al. regarding candidate genes and signaling pathways are due to the breed variations. In the current study, we collected samples from Wandong, while Gao et al. used Qinchuan breed testicular tissues for RNA-seq analysis. A study found that many candidate genes were differentially expressed (DE) and also differential alternative spliced in three different cattle breeds, including the Holstein, Jersey, and Cholistani. It suggests that breed-specific gene expression occurs at the transcriptional level and posttranscriptional modifications such as splicing, 5' capping, and poly A-tail addition. The differentially expressed genes are enriched in KEGG pathways including translation, electron transport, and immune responses (Huang et al., 2012). In the Bovine breeds, in Merino vs Poll Dorset, *SLC35A5* and *ITM2C* were identified as key DEGs; both have been reported for their association with conception and embryonic development (Hodge et al., 2021). We found different candidate genes and signaling pathways from the findings of Gao et al. due to variation in age, where they analyzed the testicular tissue in days 7 and in 4 years of age. However, we analyzed the same tissue in the age of 3 months and 3 years. As gene expression is greatly influenced by age, especially genes related to developmental phases, a comparative study was conducted on candidate genes having functional roles in regulation of muscle development at various growth stages in goat. In this study, i) 2 vs 9 months groups were compared and intramuscular fat-linked genes, e.g., *MYH13*, *IFIT1*, *METTL21C*, *SERPINE1*, *EGR1*, and *ESRRG*, were found; ii) 2 vs 24 months groups were matched and significant DEGs *RPS25*, *MYH13*, *COMP*, *LOC102186300*, and *NR4A2* were found; and iii) 9 vs 24 months goat muscles were transcriptionally matched and *PARM1*, *LOC102177715*, *MRF1-like*, *ARID5B*, and *PFKFB3* were found (Lin et al., 2017). Therefore, the candidate genes and signaling pathways are different in these two different studies of similar nature.

In mammalian testes, the cell cycle, lysine degradation, and tight junction and adherens junction pathways and host genes function together to control testis growth, spermatogenesis, and primary sexual activity (Young et al., 2015). The gene expression profiling of testes showed that some lysin degradation pathway genes, including lysine methyltransferase 2E (*KMT2E*) and nuclear receptor-binding SET domain protein 2 (*NSD2*), are active during spermatogenesis. Zhang et al. (2015) reported that the *KMT2E* gene plays key roles in various biological processes, including cell cycle progression, adult hematopoiesis, and spermatogenesis. A vital question raised in developmental biology is that how pluripotent stem cells can give rise to the cells derived from the three germ layers. The *NSD2* gene has a dual role in pluripotency exit and germ layer specification of embryonic stem cells (Tian et al., 2019).

Exonic circRNAs have been shown to act preferentially in post-transcriptional regulation in mice and humans (Bose and Ain, 2018). Similarly, we found that the majority of DE circRNAs identified in this study have putative target miRNA-binding sites, suggesting that most circRNAs are

likely to function as miRNA sponges. circRNA-miRNA network analysis further revealed that circRNAs can interact with miRNAs in multiple directions at different rates in the testes, which is consistent with observations in previous studies (Liang et al., 2017). Therefore, when translated, some circRNAs can inhibit or relieve miRNA repression. In this transcriptomic network, we selected 11 DE circRNAs that were significantly matched with 160 miRNAs, wherein the bta-miR-34, bta-miR-532, and bta-miR-204 families were involved in cattle spermatogenesis (Tscherner et al., 2014; Zhang et al., 2015). The related circRNAs may function as miRNA sponges to control the growth of testes and spermatogenesis. The circRNAs act as miRNA sponges, and research on Qinchuan cattle showed that 758 miRNAs matched with significantly differentially expressed circRNAs which play an important role in regulating gene expression in bull testes (Gao et al., 2018). The RT-qPCR validation results indicated that the dynamic expression of the circRNA had a strong interaction with the RNA-seq data. Previous studies have shown that circRNAs exhibit significant developmental stage-precision activity in human testes, suggesting their potential contribution to spermatogenesis (Dong et al., 2016). In rodents, circRNA patterns have been analyzed in three distinct stages, mitotic, meiotic, and postmeiotic germ cells (Lin et al., 2016), and correlated with testis development (Zhou et al., 2018).

CONCLUSION

In conclusion, our study presented an extensive analysis of circRNA expression profiles in immature and mature Wandong cattle bull testes. The landscapes of circRNA expression in testes were detected and used for interpreting their regulatory structure in testis development and spermatogenesis in cattle bulls. To date, there is a lack of the genetical data to establish baseline factors related to reproductive performance of local cattle bulls in China. Nevertheless, the insights from this study could lead to a breakthrough in the reproductive efficiency of local cattle resources.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183347>, <https://www.ncbi.nlm.nih.gov/geo/info/linking.html>.

ETHICS STATEMENT

The animal study was reviewed and approved by the ethical committee of the faculty of veterinary, permit number SYXK 2016-007. Every effort was made to reduce the number of animals used and minimize the sufferings of the animals.

AUTHOR CONTRIBUTIONS

IK and HL contributed to conceptualization. JZ contributed to the methodology. IK contributed to software. DZ, TX, and XZ contributed to validation. JC contributed to formal analysis. HL contributed to investigation. YZ contributed to resources. HL contributed to data curation. IK contributed to the original draft. YZ contributed to final draft and editing. NK and LA contributed in visualization. HL contributed in supervision. YZ contributed to project administration. YZ contributed in funding acquisition.

FUNDING

This research was funded by the National Research and Development Program of China (No. 2018YFD0501700), the

National Natural Science Foundation of China (No. 31101696), and Anhui Province Modern Agriculture Industry (Cattle, Goat, and Sheep) Technology System (No. AHCYJSTX-07).

ACKNOWLEDGMENTS

We highly acknowledge Fengyang County Daming Agriculture and Animal Husbandry Technology Development Co., Ltd., who provided the animals and physical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.685541/full#supplementary-material>

REFERENCES

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam Protein Families Database. *Nucleic Acids Res.* 32, D138–D141. doi:10.1093/nar/gkh121
- Bose, R., and Ain, R. (2018). Regulation of Transcription by Circular RNAs. *Circular RNAs*. 1087, 81–94. doi:10.1007/978-981-13-1426-1_7
- Cai, X., Yu, S., Mipam, T., Yang, F., Zhao, W., Liu, W., et al. (2017). Comparative Analysis of Testis Transcriptomes Associated with Male Infertility in Cattle. *Theriogenology* 88, 28–42. doi:10.1016/j.theriogenology.2016.09.047
- Card, C. J., Anderson, E. J., Zamberlan, S., Krieger, K. E., Kaproth, M., and Sartini, B. L. (2013). Cryopreserved Bovine Spermatozoal Transcript Profile as Revealed by High-Throughput Ribonucleic Acid Sequencing. *Biol. Reprod.* 88, 49. doi:10.1095/biolreprod.112.103788
- Chen, Z., Pan, X., Kong, Y., Jiang, Y., Zhong, Y., Zhang, H., et al. (2020). Pituitary-Derived Circular RNAs Expression and Regulatory Network Prediction during the Onset of Puberty in Landrace × Yorkshire Crossbred Pigs. *Front. Genet.* 11, 135. doi:10.3389/fgene.2020.00135
- Chioccarelli, T., Manfrevola, F., Ferraro, B., Sellitto, C., Cobellis, G., Migliaccio, M., et al. (2019). Expression Patterns of Circular RNAs in High Quality and Poor Quality Human Spermatozoa. *Front. Endocrinol.* 10, 435. doi:10.3389/fendo.2019.00435
- Djureinovic, D., Fagerberg, L., Hallström, B., Danielsson, A., Lindskog, C., Uhlén, M., et al. (2014). The Human Testis-specific Proteome Defined by Transcriptomics and Antibody-Based Profiling. *Mol. Hum. Reprod.* 20, 476–488. doi:10.1093/molehr/gau018
- Dong, W. W., Li, H. M., Qing, X. R., Huang, D. H., and Li, H. G. (2016). Identification and Characterization of Human Testis Derived Circular RNAs and Their Existence in Seminal Plasma. *Sci. Rep.* 6, 39080–39111. doi:10.1038/srep39080
- Eddy, E., and O'Brien, D. (1997). 5 Gene Expression during Mammalian Meiosis. *Curr. Top. Dev. Biol.* 37, 141–200. doi:10.1016/s0070-2153(08)60174-x
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the Protein Families Database. *Nucl. Acids Res.* 42, D222–D230. doi:10.1093/nar/gkt1223
- Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. (2008). Hematoxylin and Eosin Staining of Tissue and Cell Sections. *Cold Spring harbor Protoc.* 2008, pdb.prot4986. doi:10.1101/pdb.prot4986
- Fonseca, P. A. d. S., Id-Lahoucine, S., Reverter, A., Medrano, J. F., Fortes, M. S., Casellas, J., et al. (2018). Combining Multi-OMICS Information to Identify Key-Regulator Genes for Pleiotropic Effect on Fertility and Production Traits in Beef Cattle. *PLoS One* 13, e0205295. doi:10.1371/journal.pone.0205295
- Gao, Y., Li, S., Lai, Z., Zhou, Z., Wu, F., Huang, Y., et al. (2019). Analysis of Long Non-coding RNA and mRNA Expression Profiling in Immature and Mature Bovine (*Bos taurus*) Testes. *Front. Genet.* 10, 646. doi:10.3389/fgene.2019.00646
- Gao, Y., Wu, M., Fan, Y., Li, S., Lai, Z., Huang, Y., et al. (2018). Identification and Characterization of Circular RNAs in Qinchuan Cattle Testis. *Royal society open science*. 5, 180413. doi:10.1098/rsos.180413
- Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A Germline-specific Class of Small RNAs Binds Mammalian Piwi Proteins. *Nature* 442, 199–202. doi:10.1038/nature04917
- Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a Database for Circular RNAs. *Rna* 20, 1666–1670. doi:10.1261/rna.043687.113
- Griswold, M. D. (2016). Spermatogenesis: The Commitment to Meiosis. *Physiol. Rev.* 96, 1–17. doi:10.1152/physrev.00013.2015
- Guo, J. U., Agarwal, V., Guo, H., and Bartel, D. P. (2014). Expanded Identification and Characterization of Mammalian Circular RNAs. *Genome Biol.* 15, 409. doi:10.1186/s13059-014-0409-z
- Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA Circles Function as Efficient microRNA Sponges. *Nature* 495, 384–388. doi:10.1038/nature11993
- Hecht, N. B. (1998). Molecular Mechanisms of Male Germ Cell Differentiation. *Bioessays* 20, 555–561. doi:10.1002/(sici)1521-1878(199807)20:7<555::aid-bies6>3.0.co;2-j
- Hodge, M. J., de Las Heras-Saldana, S., Rindfleisch, S. J., Stephen, C. P., and Pant, S. D. (2021). Characterization of Breed Specific Differences in Spermatozoal Transcriptomes of Sheep in Australia. *Genes* 12 (2), 203. doi:10.3390/genes12020203
- Huang, J., Zhao, J., Zheng, Q., Wang, S., Wei, X., Li, F., et al. (2019). Characterization of Circular RNAs in Chinese Buffalo (*Bubalus Bubalis*) Adipose Tissue: a Focus on Circular RNAs Involved in Fat Deposition. *Animals* 9, 403. doi:10.3390/ani9070403
- Huang, W., Nadeem, A., Zhang, B., Babar, M., Soller, M., and Khatib, H. (2012). Characterization and Comparison of the Leukocyte Transcriptomes of Three Cattle Breeds. *PLoS ONE* 7 (1), e30244. doi:10.1371/journal.pone.0030244
- Ibtisham, F., Wu, J., Xiao, M., An, L., Banker, Z., Nawab, A., et al. (2017). Progress and Future prospect of *In Vitro* Spermatogenesis. *Oncotarget* 8, 66709–66727. doi:10.18632/oncotarget.19640
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions. *Genome Biol.* 14, R36–R13. doi:10.1186/gb-2013-14-4-r36
- Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., et al. (2007). CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine. *Nucleic Acids Res.* 35, W345–W349. doi:10.1093/nar/gkm391

- Kulcheski, F. R., Christoff, A. P., and Margis, R. (2016). Circular RNAs Are miRNA Sponges and Can Be Used as a New Class of Biomarker. *J. Biotechnol.* 238, 42–51. doi:10.1016/j.jbiotec.2016.09.011
- Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923
- Lasda, E., and Parker, R. (2014). Circular RNAs: Diversity of Form and Function. *Rna* 20, 1829–1842. doi:10.1261/rna.047126.114
- Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14. *cell* 75, 843–854. doi:10.1016/0092-8674(93)90529-y
- Liang, G., Yang, Y., Niu, G., Tang, Z., and Li, K. (2017). Genome-wide Profiling of *Sus scrofa* Circular RNAs across Nine Organs and Three Developmental Stages. *DNA Res.* 24, 523–535. doi:10.1093/dnares/dsx022
- Lin, X., Han, M., Cheng, L., Chen, J., Zhang, Z., Shen, T., et al. (2016). Expression Dynamics, Relationships, and Transcriptional Regulations of Diverse Transcripts in Mouse Spermatogenic Cells. *RNA Biol.* 13, 1011–1024. doi:10.1080/15476286.2016.1218588
- Lin, Y., Zhu, J., Wang, Y., Li, Q., and Lin, S. (2017). Identification of Differentially Expressed Genes through RNA Sequencing in Goats (*Capra hircus*) at Different Postnatal Stages. *PLoS one* 12 (8), e0182602. doi:10.1371/journal.pone.0182602
- Lunstra, D.-D., and Echtenkamp, S. E. (1982). Puberty in Beef Bulls: Acrosome Morphology and Semen Quality in Bulls of Different Breeds. *J. Anim. Sci.* 55, 638–648. doi:10.2527/jas1982.553638x
- Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated Genome Annotation and Pathway Identification Using the KEGG Orthology (KO) as a Controlled Vocabulary. *Bioinformatics* 21, 3787–3793. doi:10.1093/bioinformatics/bti430
- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency. *Nature* 495, 333–338. doi:10.1038/nature11928
- M. K. M. A., Kumaresan, A., Yadav, S., Mohanty, T. K., and Datta, T. K. (2019). Comparative Proteomic Analysis of High- and Low-fertile buffalo Bull Spermatozoa for Identification of Fertility-Associated Proteins. *Reprod. Domest. Anim.* 54, 786–794. doi:10.1111/rda.13426
- Mukherjee, A., Koli, S., and Reddy, K. V. R. (2014). Regulatory Non-coding Transcripts in Spermatogenesis: Shedding Light on 'dark Matter'. *Andrology* 2, 360–369. doi:10.1111/j.2047-2927.2014.00183.x
- Paiano, J., Wu, W., Yamada, S., Sciascia, N., Callen, E., Paola Cotrim, A., et al. (2020). ATM and PRDM9 Regulate SPO11-Bound Recombination Intermediates during Meiosis. *Nat. Commun.* 11, 857–915. doi:10.1038/s41467-020-14654-w
- Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitch, S., et al. (2009). Transcriptome Analysis by Strand-specific Sequencing of Complementary DNA. *Nucleic Acids Res.* 37, e123. doi:10.1093/nar/gkp596
- Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi:10.1038/nprot.2016.095
- Quan, G., and Li, J. (2018). Circular RNAs: Biogenesis, Expression and Their Potential Roles in Reproduction. *J. Ovarian Res.* 11, 9–12. doi:10.1186/s13048-018-0381-4
- Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *Plos Comput. Biol.* 5, e1000598. doi:10.1371/journal.pcbi.1000598
- Rawlings, N., Evans, A., Chandolia, R., and Bagu, E. (2008). Sexual Maturation in the Bull. *Reprod. Domest. Anim.* 43, 295–301. doi:10.1111/j.1439-0531.2008.01177.x
- Robles, V., Herráez, P., Labbé, C., Cabrita, E., Pšenička, M., Valcarce, D. G., et al. (2017). Molecular Basis of Spermatogenesis and Sperm Quality. *Gen. Comp. Endocrinol.* 245, 5–9. doi:10.1016/j.ygcen.2016.04.026
- Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids Are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-like Structures. *Proc. Natl. Acad. Sci.* 73, 3852–3856. doi:10.1073/pnas.73.11.3852
- Shen, M., Wu, P., Li, T., Wu, P., Chen, F., Chen, L., et al. (2020). Transcriptome Analysis of circRNA and mRNA in Theca Cells during Follicular Development in Chickens. *Genes* 11, 489. doi:10.3390/genes11050489
- Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44. doi:10.1038/nprot.2008.211
- Song, H., Zhu, L., Li, Y., Ma, C., Guan, K., Xia, X., et al. (2015). Exploiting RNA-Sequencing Data from the Porcine Testes to Identify the Key Genes Involved in Spermatogenesis in Large White Pigs. *Gene* 573, 303–309. doi:10.1016/j.gene.2015.07.057
- Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., et al. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cel Rep.* 3, 2179–2190. doi:10.1016/j.celrep.2013.05.031
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing Sequence Intrinsic Composition to Classify Protein-Coding and Long Non-coding Transcripts. *Nucleic Acids Res.* 41, e166. doi:10.1093/nar/gkt646
- Svingen, T., and Koopman, P. (2013). Building the Mammalian Testis: Origins, Differentiation, and Assembly of the Component Cell Populations. *Genes Develop.* 27, 2409–2426. doi:10.1101/gad.228080.113
- Tian, T. V., Di Stefano, B., Stik, G., Vila-Casadesús, M., Sardina, J. L., Vidal, E., et al. (2019). Whsc1 Links Pluripotency Exit with Mesoderm Specification. *Nat. Cel Biol* 21, 824–834. doi:10.1038/s41556-019-0342-1
- Tomlinson, M., Jennings, A., Macrae, A., and Truysers, I. (2017). The Value of Trans-scrotal Ultrasonography at Bull Breeding Soundness Evaluation (BBSE): The Relationship between Testicular Parenchymal Pixel Intensity and Semen Quality. *Theriogenology* 89, 169–177. doi:10.1016/j.theriogenology.2016.10.020
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nat. Biotechnol.* 28, 511–515. doi:10.1038/nbt.1621
- Tscherner, A., Gilchrist, G., Smith, N., Blondin, P., Gillis, D., and Lamarre, J. (2014). MicroRNA-34 Family Expression in Bovine Gametes and Preimplantation Embryos. *Reprod. Biol. Endocrinol.* 12, 85–89. doi:10.1186/1477-7827-12-85
- Ulitisky, I., and Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* 154, 26–46. doi:10.1016/j.cell.2013.06.020
- Wang, K., Long, B., Liu, F., Wang, J.-X., Liu, C.-Y., Zhao, B., et al. (2016). A Circular RNA Protects the Heart from Pathological Hypertrophy and Heart Failure by Targeting miR-223. *Eur. Heart J.* 37, 2602–2611. doi:10.1093/eurheartj/ehv713
- Waqas, M. S., Ciccirelli, M., Oatley, M. J., Kaucher, A. V., Tibary, A., and Oatley, J. M. (2019). Enhanced Sperm Production in Bulls Following Transient Induction of Hypothyroidism during Prepubertal Development. *J. Anim. Sci.* 97, 1468–1477. doi:10.1093/jas/sky480
- Wu, S., Mipam, T., Xu, C., Zhao, W., Shah, M. A., Yi, C., et al. (2020). Testis Transcriptome Profiling Identified Genes Involved in Spermatogenic Arrest of Cattle. *PLoS one* 15, e0229503. doi:10.1371/journal.pone.0229503
- Yadav, R. P., and Kotaja, N. (2014). Small RNAs in Spermatogenesis. *Mol. Cell. Endocrinol.* 382, 498–508. doi:10.1016/j.mce.2013.04.015
- You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., et al. (2015). Neural Circular RNAs Are Derived from Synaptic Genes and Regulated by Development and Plasticity. *Nat. Neurosci.* 18, 603–610. doi:10.1038/nn.3975
- Young, J. C., Wakitani, S., and Loveland, K. L. (2015). TGF- β Superfamily Signaling in Testis Formation and Early Male Germline Development. *Semin. Cel. Dev. Biol.* 45, 94–103. doi:10.1016/j.semcdb.2015.10.029
- Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias. *Genome Biol.* 11, R14–R12. doi:10.1186/gb-2010-11-2-r14
- Yu, Z., Guo, R., Ge, Y., Ma, J., Guan, J., Li, S., et al. (2003). Gene Expression Profiles in Different Stages of Mouse Spermatogenic Cells during Spermatogenesis. *Biol. Reprod.* 69, 37–47. doi:10.1095/biolreprod.102.012609
- Zhang, F., Zhang, X., Ning, W., Zhang, X., Ru, Z., Wang, S., et al. (2021). Expression Analysis of Circular RNAs in Young and Sexually Mature Boar Testes. *Animals* 11, 1430. doi:10.3390/ani11051430
- Zhang, S., Zhang, Y., Yang, C., Zhang, W., Ju, Z., Wang, X., et al. (2015). TNF1 Functional SNPs in Bta-miR-532 and Bta-miR-204 Target Sites Are Associated with Semen Quality Traits in Chinese Holstein Bulls. *Biol. Reprod.* 92 (139), 139–110. doi:10.1095/biolreprod.114.126672
- Zhou, G. H., Liu, L., Xiu, X. L., Jian, H. M., Wang, L. Z., Sun, B. Z., et al. (2001). Productivity and Carcass Characteristics of Pure and Crossbred Chinese Yellow Cattle. *Meat Sci.* 58, 359–362. doi:10.1016/s0309-1740(00)00160-1

Zhou, T., Xie, X., Li, M., Shi, J., Zhou, J. J., Knox, K. S., et al. (2018). Rat BodyMap Transcriptomes Reveal Unique Circular RNA Features across Tissue Types and Developmental Stages. *Rna* 24, 1443–1456. doi:10.1261/rna.067132.118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of

the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Copyright © 2021 Khan, Liu, Zhuang, Khan, Zhang, Chen, Xu, Avalos, Zhou and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Transcriptome-Wide Analyses Identify Dominant as the Predominantly Non-Conservative Alternative Splicing Inheritance Patterns in F1 Chickens

Xin Qi, Hongchang Gu* and Lujiang Qu*

Department of Animal Genetics and Breeding, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Lifan Zhang,
Nanjing Agricultural University, China
Paolo Zambonelli,
University of Bologna, Italy

*Correspondence:

Lujiang Qu
qulij@163.com
Hongchang Gu
guhgc@cau.edu.cn

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 11 September 2021

Accepted: 05 November 2021

Published: 03 December 2021

Citation:

Qi X, Gu H and Qu L (2021)
Transcriptome-Wide Analyses Identify
Dominant as the Predominantly Non-
Conservative Alternative Splicing
Inheritance Patterns in F1 Chickens.
Front. Genet. 12:774240.
doi: 10.3389/fgene.2021.774240

Transcriptome analysis has been used to investigate many economically traits in chickens; however, alternative splicing still lacks a systematic method of study that is able to promote proteome diversity, and fine-tune expression dynamics. Hybridization has been widely utilized in chicken breeding due to the resulting heterosis, but the dynamic changes in alternative splicing during this process are significant yet unclear. In this study, we performed a reciprocal crossing experiment involving the White Leghorn and Cornish Game chicken breeds which exhibit major differences in body size and reproductive traits, and conducted RNA sequencing of the brain, muscle, and liver tissues to identify the inheritance patterns. A total of 40 515 and 42 612 events were respectively detected in the brain and muscle tissues, with 39 843 observed in the liver; 2807, 4242, and 4538 events significantly different between two breeds were identified in the brain, muscle, and liver tissues, respectively. The hierarchical cluster of tissues from different tissues from all crosses, based on the alternative splicing profiles, suggests high tissue and strain specificity. Furthermore, a comparison between parental strains and hybrid crosses indicated that over one third of alternative splicing genes showed conserved patterns in all three tissues, while the second prevalent pattern was non-additive, which included both dominant and transgressive patterns; this meant that the dominant pattern plays a more important role than suppression. Our study provides an overview of the inheritance patterns of alternative splicing in layer and broiler chickens, to better understand post-transcriptional regulation during hybridization.

Keywords: chicken, hybridization, RNA-seq, dominant, alternative splicing, inheritance pattern

INTRODUCTION

Splicing of pre-mRNA is a crucial post-transcriptional process that increases proteome diversity in eukaryotes. Alternative splicing (AS) generates multiple isoforms from a single gene using different combinations of exons. AS is a widespread and complex component of gene regulation in humans and domestic animals, and increasing evidence suggests that aberrant AS functionality can be the cause or consequence of many diseases, and may also associating with economically important traits in domestic animals (Pan et al., 2008; Merkin et al., 2012; Gao et al., 2018; Dlamini et al., 2021). Therefore, splice-altering therapies using animal models have been extensively studied for many diseases such as neurodegeneration and muscular dystrophies, and AS events have also emerged as

new biomarkers in some circumstances (Montes et al., 2019; Zhao, 2019). Changes in AS are regulated by the interactions between cis- and trans-acting elements, and studies in *Camellia* and *Drosophila* suggest that parental genetic divergence may affect the regulation patterns in hybrids due to these interactions (Coolon et al., 2014; Zhang et al., 2019). Therefore, it is important to study the AS regulatory mechanisms in birds.

Based on different combinations of the constitutive and alternative exons, AS is divided into seven types: exon skipping (SE), intron retention (RI), alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), mutually exclusive exons (MXE), alternative promoters (AFEs and ALEs), and alternative polyadenylation (tandem 3'UTRs). SE is the most prevalent AS event in approximately 40% of higher eukaryotes, and commonly generates functional isoforms, while RI is the dominant type in plants (Barbazuk et al., 2008; Weatheritt et al., 2016; Cardoso et al., 2018; Chen S.-Y. et al., 2019). With the rapid development of sequencing technology, a broad range of bioinformatics approaches can identify and classify AS events using isoform-based and count-based methods, of which several tools perform robustly and exhibit excellent overall performance (Mehmood et al., 2019). However, there is a lack of systematic analysis of AS in chickens, and it is necessary to study the components and divergence patterns of splicing events, providing an alternate view of transcriptome plasticity.

Hybridization is ubiquitous in nature—involving more than 25 and 10% of plants and animals, respectively—and widely utilized in breeding programs (Whitney et al., 2010). Some hybrids show enhanced environmental adaption and growth rate, whereas others are infertile or exhibit negative economic traits (Abasht and Lamont, 2007; Chen, 2010; Chen et al., 2013; Zhang et al., 2015; Clasen et al., 2017). Hybridizing two different strains can remodel the parental gene patterns, with the genes in hybrids diverging from the mid-parental value, leading to “transcriptome shock” (Hegarty et al., 2006; Han et al., 2014; Cui et al., 2015). These genes mainly contribute to some transgressive phenotypes called over- and under-dominant genetic patterns (Zhuang and Tripp, 2017). Additive and dominant patterns also represent phenotypical variations, while conserved patterns show parental similarity. To take advantage of genome-wide methods for expression analysis, the classic hypotheses of inheritance, dominance, over-dominance, and epistasis should have more contributions at the molecular level to explore the mechanism of heredity (Shull, 1908; Bateson, 1910; Jones, 1917; Mcmanus et al., 2010). However, the identified predominant genetic patterns regulating phenotype divergence are not always consistent among studies because of different genetic backgrounds, species, and tissues employed in those studies. Studies on *Camellia* and coffee have indicated that the non-additive expression prevailed over other patterns (Marie-Christine et al., 2015; Zhang et al., 2019). On the other hand, several studies have suggested that additivity is the predominant genetic pattern in maize, rice, and cotton (Swanson-Wagner et al., 2006; Li et al., 2008; Rapp et al., 2009), while divergent outcomes suggest that additive or high parent-dominance is the major

pattern in chickens (Mai et al., 2019; Zhuo et al., 2019). Most previous studies have identified different inheritance patterns based on gene expression, and there is rarely a transcriptomic study that investigates AS event patterns, considering the association between AS and gene regulation at the post-transcriptional level.

In this study, Cornish Game (CG) and White Leghorn (WL), representing broilers and layers, respectively, were used as the parental strains to produce purebred and reciprocal crossed progenies. Taking advantage of RNA-sequencing (RNA-seq), tissue samples from the brain, liver, and breast muscle were collected and sequenced. Splicing events from each sample were identified, classified, and quantified, and events significantly different between purebreds were detected for further study. Finally, five main types of inheritance patterns—conserved, additive, parental-enhancing/suppressing, dominant, and transgressive—were categorized, and compared between purebred strains and hybrid crosses. Changes in alternatively spliced genes indicated that the diverse AS inheritance patterns have different influences on heredity, and variation during hybridization. AS is an effective and novel approach for investigating genetic patterns, and understanding the molecular mechanisms of heterosis.

MATERIALS AND METHODS

Sample Collection and RNA Sequencing

We used CG, a broiler breed with superior growth and muscle development, and WL, a layer breed with high egg production, acquired from the National Engineering Laboratory for Animal Breeding of the China Agricultural University. The four breeding patterns used in this study resulted in pure-bred progeny, CG, and WL, representing the first generation (F1) with parents of the same type, and reciprocal cross progeny WL ♂ × CG ♀ (LC), and CG ♂ × WL ♀ (CL) representing F1 hybrids (**Supplementary Figure S1**). Each group had six offspring (three female and three male), except CL where only two females were obtained. We collected the brain, breast muscle, and liver from 23 one-day-old chicks, and extracted total RNA from the tissue samples using Trizol reagent (Invitrogen, Carlsbad, CA, United States). The DNA and RNA quality was assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., United States) and agarose gel electrophoresis. After synthesizing cDNA, PCR amplifications, and library construction, total RNA was sequenced, using paired-end 100-bp reads with a 300-bp insert, on an Illumina HiSeq 2500 platform (Illumina Inc., San Diego, CA, United States) using standard Illumina RNA-seq protocols.

RNA-Sequencing Data Alignment and Alternative Splicing Analysis

The RNA-seq data were aligned to the chicken reference genome (*Gallus gallus*-6.0) using STAR v2.7.5 (Dobin et al., 2012). Duplicate reads were removed to eliminate potential bias, using SAMtools (Li, 2009; Dozmorov et al., 2015). Putative AS

events were detected and annotated from aligned RNA-seq data using rMATS v4.1.0 (Wang et al., 2017). Five major types of AS events were identified: A3SS, A5SS, RI, MXE, and SE. To quantify and compare event variation, the percent spliced-in (PSI) value of each AS was calculated for each sample using reads on target and junction counts, where PSI was equal to, the number of reads specific to exon inclusion isoform divided by the sum of reads specific to exon inclusion and exclusion isoforms. Moreover, a hierarchical model for paired replicates and false discovery rate (FDR) correction ($FDR < 0.05$) was used to determine the statistical differences between the parental strains (Shen et al., 2012). Only the events occurring on autosomes were considered in this study because of incomplete dosage compensation in the chicken sex chromosome. Additionally, if the sum of inclusion and skipping read counts is less than 10 (average in replicate samples), AS was considered low quality and then filtered. Liver samples from the hybrid females were removed because the detection process for putative AS events provided abnormal results.

Classification of Alternative Splicing Inheritance Patterns

To measure differences between parents and hybrids in order to identify inheritance patterns of AS, samples were recombined as male cross (MC: CG ♂, WL ♂, CL ♂), female cross (FC: CG ♀, WL ♀, CL ♀), male reciprocal cross (MR: CG ♂, WL ♂, LC ♂), and female reciprocal cross (FR: CG ♀, WL ♀, LC ♀). First, shared AS was determined by merging expressed events in hybrids with significantly different ($FDR < 0.05$) events in pure-bred, and calculating the average PSI value for biological replicates. A 1.25-fold threshold was set as the criterion for classification as conserved or non-conserved splicing (Gu et al., 2020). Non-conserved AS was classified into eight types: events for which quantification in the hybrid was less than in CG and greater than in WL (or vice versa) was defined as additive (2 types); splicing for which quantification in the hybrid was remarkably higher or lower than in one of the parents and similar to that in another was categorized as dominant (4 types); and splicing for which quantification in the hybrid was significantly greater or lesser than in both parents was classified as transgressive inheritance (two types). Non-conserved genes which were identified in the same mode in more than two groups in a tissue, would be considered as exhibiting that inheritance mode in the corresponding tissue. All eight AS types are listed in **Supplementary Figure S1**. Exploring the biological function of these genes further, Gene Ontology (GO) classification and enrichment analysis were performed using PANTHER v14 (Thomas et al., 2003), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was carried out using the KOBAS v3.0 (Bu et al., 2021).

Statistical Test of Inheritance Patterns

R (v4.0.2) was used for most of the statistical analyses in this study. Principal component analysis was performed using Prcomp for statistically different AS between the parental lines. Besides, the Venn diagram was visualized by the

“VennDiagram” package, and heat-map was prepared using “pheatmap,” with hierarchical clustering among samples performed by “hclust.” After classifying inheritance patterns, the Kruskal–Wallis test was carried out to identify differences among the three tissues, while the Mann–Whitney test was used to compare divergence between CG/WL-like dominant, up/down-regulation dominant, and over/under-dominant cases in each tissue.

RESULTS

Alternative Splicing Divergence Between Cornish Game and White Leghorn

We collected RNA-seq data from brain, liver, and breast muscle tissues of two inbred chicken strains, CG, and WL, representing parental lines, as well as their reciprocal crossed progenies. There was 246.3 Gb of RNA-seq data and 3.6 million mapped reads for each sample. After filtering low-quality reads using the NGS QC Toolkit v2.3 (Patel et al., 2012) and mapping them onto the reference genome, we obtained, on average, 22.8, 21.3, and 17.8 million mapped reads per individual for the brain, muscle, and liver tissues, respectively. AS events were quantified as PSI and classified into five types—A3SS, A5SS, MXE, RI, and SE—while statistically significant differences between WL and CG were identified.

The number of putative splicing events was 45668, 47983, and 46313 in the brain, muscle, and liver tissues, respectively, most of which were expressed (86%), and related to more than 7000 genes (**Figure 1**). SE formed a large proportion of splicing in the brain and muscle tissues (approximately 36%), A3SS was slightly higher than SE in the liver tissue, and RI only accounted for 3% of the tissues. 28, 38, and 44% of genes only underwent one splicing (simple event), and over 56% of events among the tissues were complex events (**Figure 2A**). Most of the events were primarily distributed on chromosome-1 and numbered over 5200, while those located on chromosome-30 were less than 50 in number. In addition, over 6000 event locations were positioned on chromosome-4 in the liver and over 2700 in chromosome-7 in muscle tissues (**Figure 2B**). On the other hand, several genes participated in more than two types of events; 100 genes covered five types, with approximately 60% related to multiple types of events, among the three tissues (**Figure 3**). To summarize, locations of alternatively spliced events widely exist in tissues, and have a complex correlation with genes.

From among 40000 expressed AS, we used a hierarchical model to detect 2807, 4242, and 4538 ($FDR < 0.05$) events as significantly different between the two strains, in the brain, muscle, and liver tissues, respectively. These spliced events related to 1823, 2020, and 1568 genes, approximately half of which could be tissue-specific. In addition, the number of up-regulated splicing events was 3.4–5.5% (1387, 2038, and 2188), while down-regulated events accounted for 3.5–5.9% (1420, 2204, and 2350), between layers and broilers. Based on PSI values of divergence events in each sample, principal component analysis was performed (**Figure 4**); the tissue was significantly influenced by AS, and the strain probably played an important role in the

Tissue	AS Type	Alternative 3' Spliced Sites (A3SS)	Alternative 5' Spliced Sites (A5SS)	Mutually exclusive Exons (MXE)	Retained Introns (RI)	Skipped Exons (SE)	total
Brain	Putative events	14601	9885	2315	2319	16548	45668
	Expressed AS	13533	9006	2082	1197	14697	40515
	Related Gene	5572	4500	934	757	5608	8691
Muscle	Putative events	15163	10225	2794	2514	17287	47983
	Expressed AS	14073	9362	2480	1385	15312	42612
	Related Gene	5313	4353	883	788	5525	8280
Liver	Putative events	13310	9497	7303	2451	13752	46313
	Expressed AS	12278	8633	6191	1307	11434	39843
	Related Gene	4409	3726	636	676	4168	7157

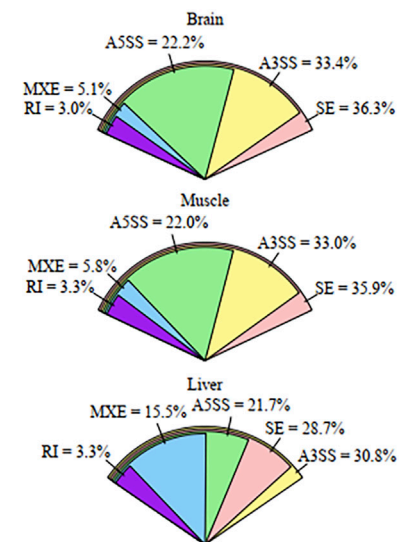


FIGURE 1 | Overview of alternative splicing in Cornish and White Leghorn.

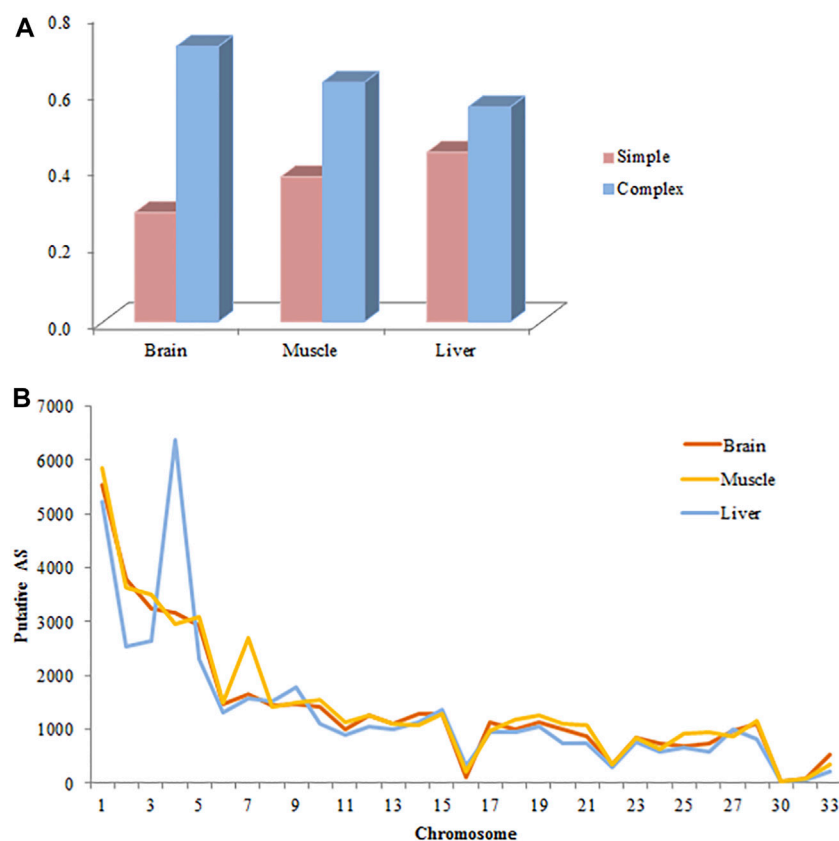


FIGURE 2 | Basic information of alternative splicing events in each tissue. **(A)** Complexity of AS events per gene. **(B)** The distribution of AS events in the chicken genome. The distribution of 45668, 47983, and 46313 putative AS events identified from brain, muscle and liver on the chicken autosomes are shown. Most of the events distributed primarily on chromosomes 1 and 2.

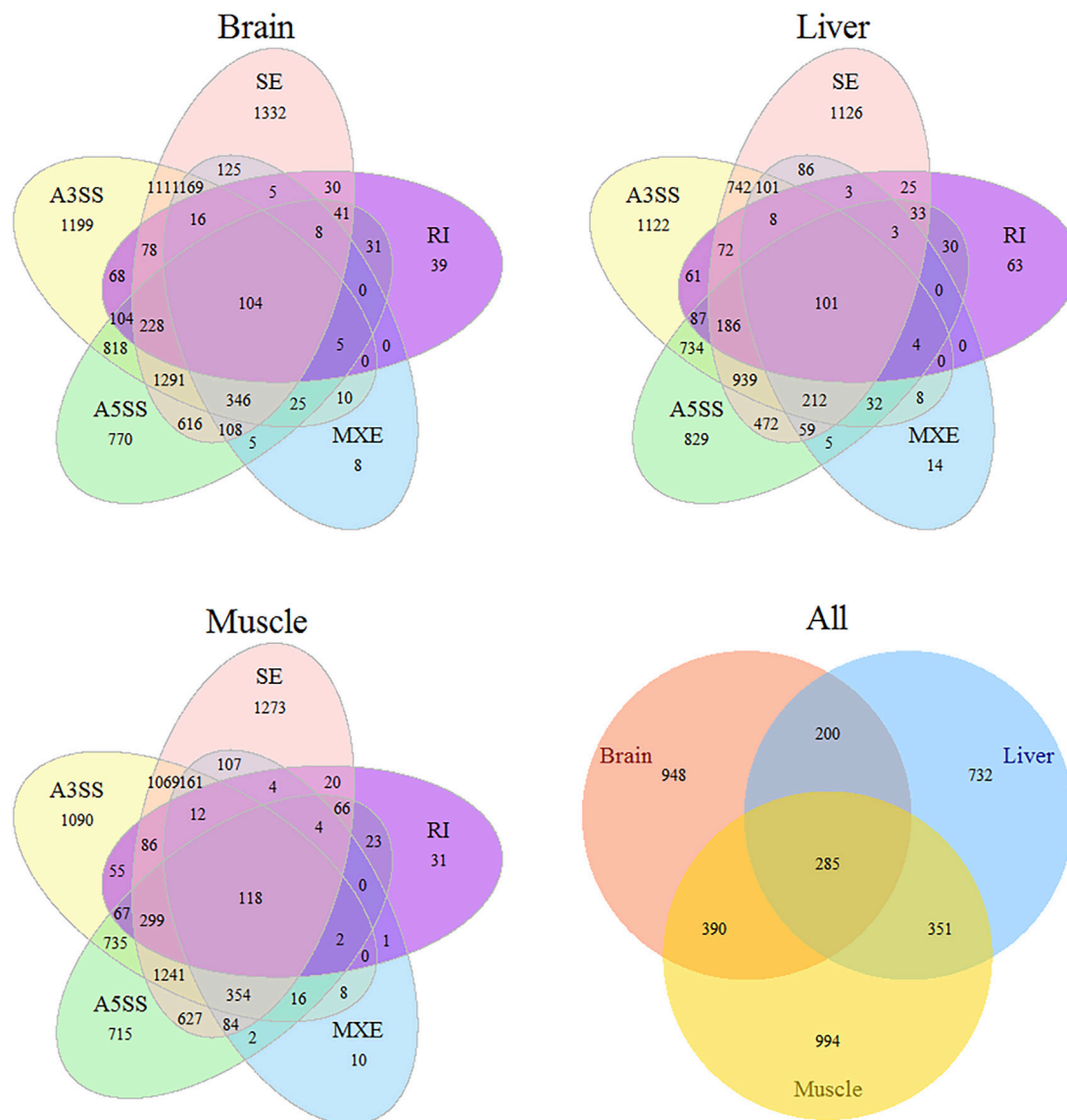


FIGURE 3 | Venn diagrams illustrating statistical results of different types AS in tissue and divergence AS among tissues.

liver and brain. The parent-of-origin effect might have influenced muscle formation because hybrids were observed to be slightly closer to the maternal group, while it showed insignificant influence on sex determination of the offspring. Further, hierarchical clustering using Spearman correlation—where the samples were classified into three tissue-based clusters and purebred were always categorized in the same group—showed that tissue- and strain-based clustering tended to be stronger than sex-based clustering (Figure 5).

Classification of Alternative Splicing Inheritance Patterns

After filtering and removing sex chromosome splicing, AS between parental strains was merged with expressed splicing

in hybrids; 754 splicing events in the liver were used for further analysis as against 1030 and 1950 AS events in the brain and muscle because only liver samples from male hybrids were available. Based on the difference in expression between parental and hybrid splicing, events were categorized into nine types in all four groups (MC, MR, FC, and FR) (Figure 6).

Based on statistical results (Table 1), we detected a significant difference in the sum of conserved, additive, WL-like dominant, and enhancing/suppressing dominant patterns among the three tissues (Kruskal–Wallis test, p -value $\approx 0.02 < 0.05$), whereas there was no divergence in the number of CG-like dominant (p -value = 0.51) and transgressive patterns ($p = 0.11$). Conserved splicing was predominant, with above 34% in all three tissue types, while additive splicing accounted for a small proportion of non-

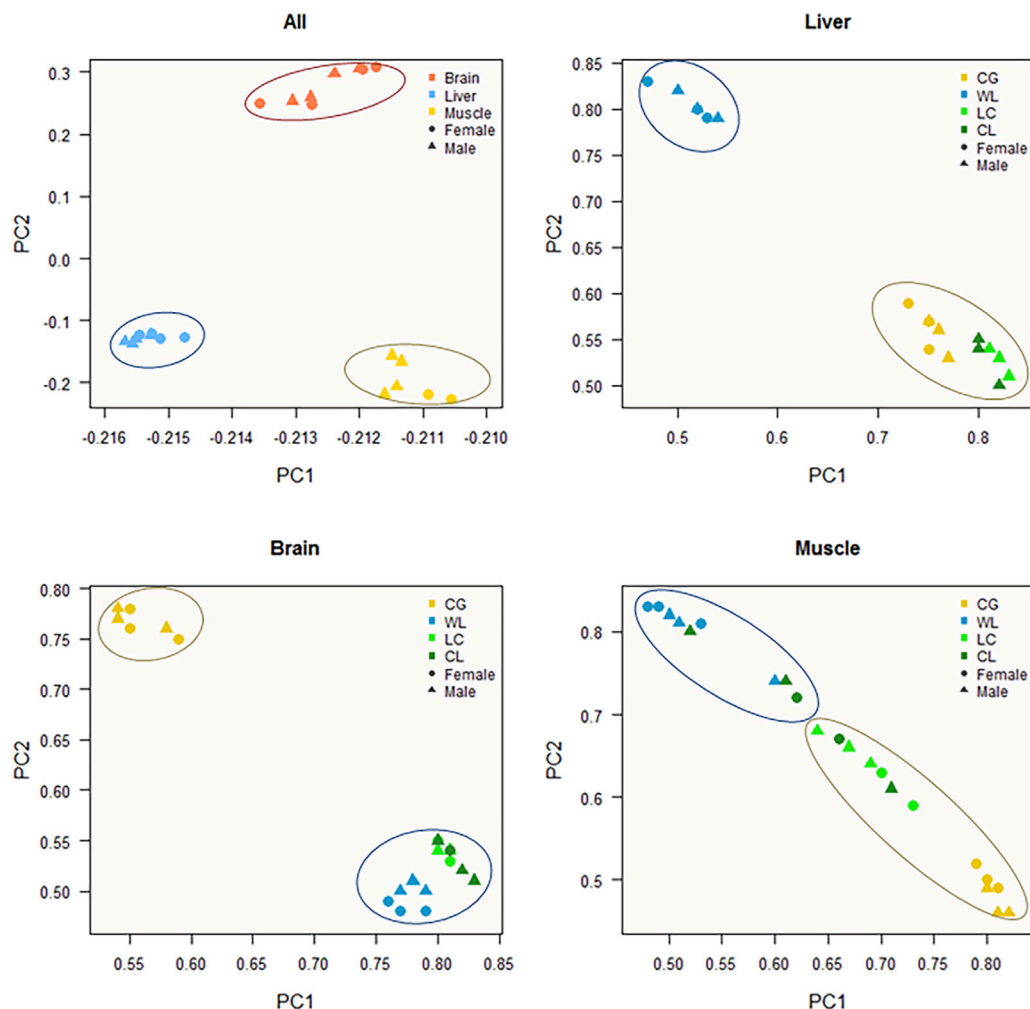


FIGURE 4 | Principal Component Analysis of alternative splicing. PCA analysis is performed using PSI value for significantly different ($FDR < 0.05$) AS events between purebreds, which events located on the sex chromosomes has been excluded.

conserved patterns, with an average of 8.7, 9.3, and 8.0% in the brain, muscle, and liver tissues, respectively. A greater proportion of events between the hybrids and their parents exhibited a non-additively expressed pattern, in which the sum of parental dominance was approximately 30%, and transgressivity accounted for approximately 22.7, 11.6, and 22.5% in the brain, muscle, and liver tissues, respectively. Therefore, up-regulated dominance was higher than down-regulated, in muscle and brain tissues (Mann–Whitney test, p -value = 0.03 < 0.05). Interestingly, the relative proportion of CG-like dominance was larger in brain tissues of the female groups and liver tissues of the male groups, while WL-like dominance was greater in the brain and muscle tissues of the male groups. No differences were detected between over-dominance and under-dominance in each tissue type. To check whether our conclusion was sensitive to specific statistical methods, we used Fisher's exact test to identify events with significant divergence between hybrids and parents (**Supplementary Table S1**), with the resulting p -values controlled for an FDR ($FDR < 0.05$). Overall, we

drew the same conclusion that most non-conserved events were dominant, with transgressive modes during hybridization, while a few events displayed additive patterns among the tissues.

Functional Analysis of Non-Conserved AS Genes

We selected non-conserved genes that showed the same pattern in a particular organ tissue in more than two groups, to ensure accuracy prior to functional annotation. There were 148, 211, 203, and 307 AS genes with additive, CG dominant, WL dominant, and transgressive patterns, respectively, in the brain tissue, while the corresponding numbers for the muscle tissue were 179, 254, 302, and 188, and those for the liver tissue were 25, 84, 46, and 125, respectively. GO functional enrichment analysis was used for of four categories ($FDR < 0.05$) in the brain and muscle tissues, while they were filtered for $p < 0.001$ for the liver tissue due to fewer selected gene samples (**Figure 7**). The results showed that expressed AS genes in the brain and muscle tissues are mostly

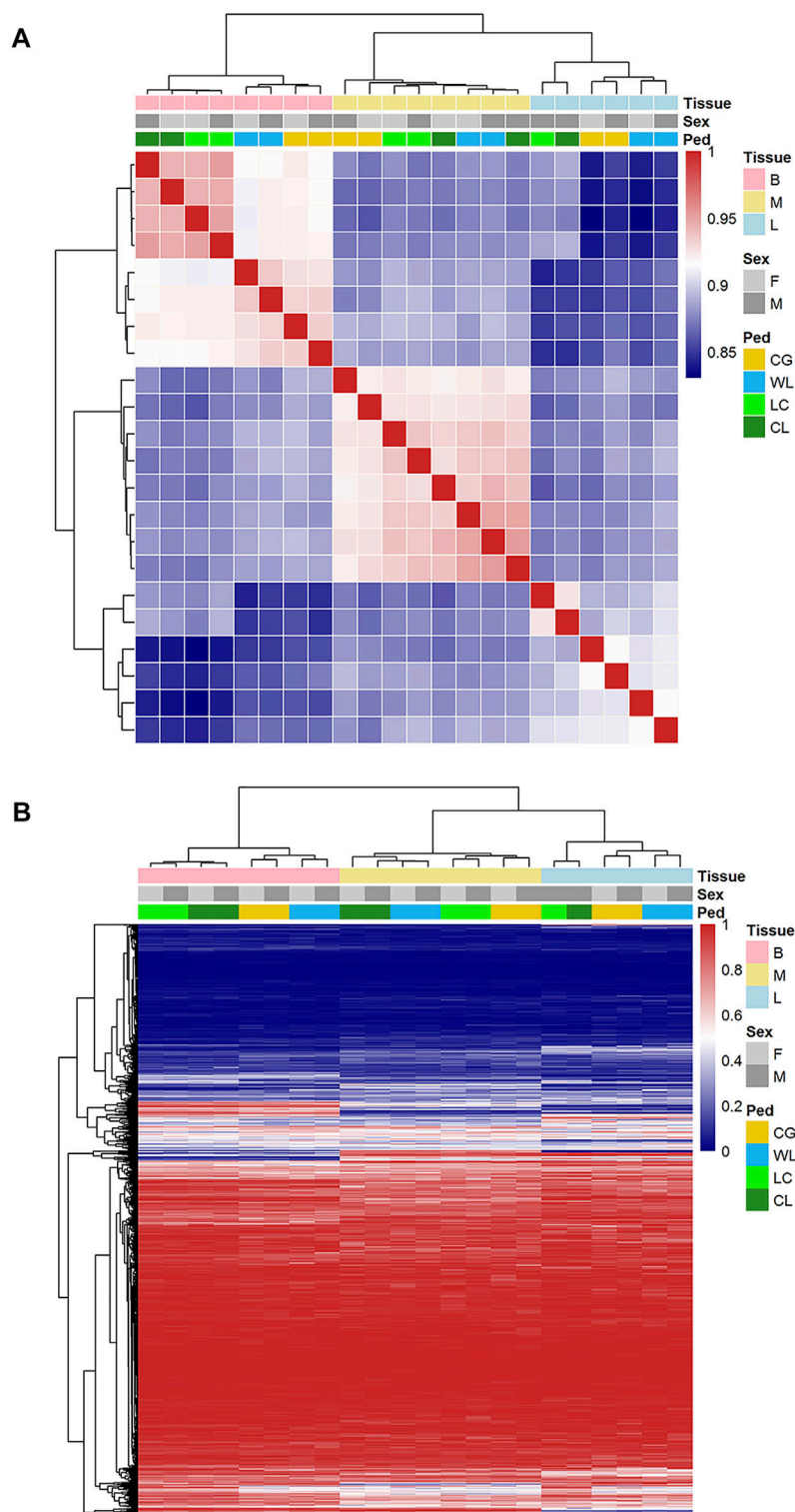


FIGURE 5 | Correlations and hierarchical clustering of alternative splicing. **(A)** Spearman correlation based on alternative splicing (PSI) expressed in brain, liver, and muscle. **(B)** Hierarchical cluster for AS events in each group.

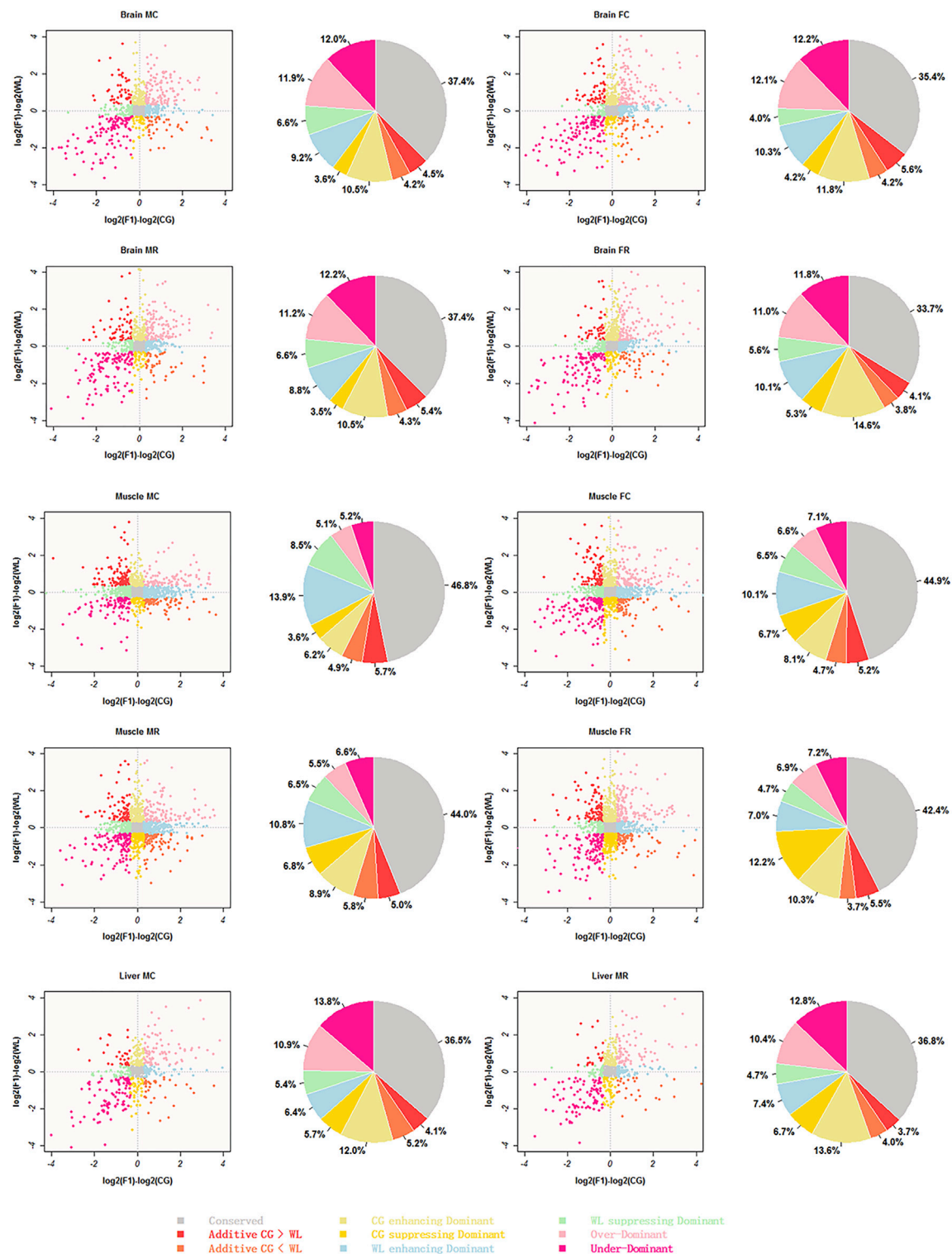


FIGURE 6 | Scatter-plot and pie representing different inheritance patterns. inheritance patterns identified from reciprocal crosses and parental lines brain, muscle, and liver tissues were classified into nine clusters listed at the bottom of the images, respectively. Scatterplots represent relationships between F1 hybrids and their two parents, and pie charts show the relative proportions of these patterns for each cluster using different colors.

TABLE 1 | Summary statistics for splicing patterns in reciprocal crosses among tissues.

Tissue	Group	Conser vative	Additivity		CG-like dominant		WL-like dominant		Transgressivity	
			CG>WL	CG<WL	Up	Down	Up	Down	Over-dominant	Under-dominant
brain	FC	350	55	42	117	42	102	40	120	121
	MC	371	45	42	104	36	91	65	118	119
	FR	333	41	38	144	52	100	55	109	117
	MR	371	54	43	104	35	87	65	111	121
muscle	FC	808	94	84	145	121	182	117	119	128
	MC	840	103	87	111	65	250	153	91	93
	FR	769	100	68	186	222	127	86	126	130
	MR	794	91	104	160	123	195	118	100	120
liver	MC	261	29	37	86	41	46	39	78	99
	MR	259	26	28	96	47	52	33	73	90

related to cellular components such as cytoplasm and intracellular anatomical structure, which are mainly associated with biological processes including cellular amino acid catabolic process as well as small molecule metabolic and catabolic processes in the liver. Specifically, AS genes of classified non-conserved mode from the muscle tissue were involved in the development of ribosomes, and muscle structure such as actin filament, myofibril, contractile fiber, sarcomere, and sarcolemma.

KEGG pathway enrichment analysis was performed to identify whether AS genes are involved in signal transduction pathways and biochemical metabolic pathways. We focused on the 10 most significant pathways in each tissue (**Figure 8, Supplementary Table S2**). Six pathways which are involved in many important biological processes such as metabolic pathways, energy metabolism, genetic information processing in RNA translation and transcription, cellular community, and cell motility, overlap among tissues. In the brain, CG-like dominant and under-dominant genes that contained an average of 17.6 exons were involved in several pathways related to energy metabolism and RNA translation, such as mRNA surveillance, RNA transport, and ribosome and cellular processes including regulation of the actin cytoskeleton and adherens junction. Most of the CG-like dominant genes in the muscle were enriched by ribosome, spliceosome, and oxidative phosphorylation, while additive genes were associated with carbohydrate metabolism including glycolysis and pyruvate metabolism. In the liver, we found that metabolic pathways, glycolipid metabolism pathways, and several amino acid metabolism-related pathways that included tryptophan, histidine, cysteine, and methionine were significantly enriched in WL-like dominant and under-dominant genes.

DISCUSSION

A ubiquitous and complex component of gene regulation in humans, animals, and plants is AS (Pan et al., 2008; Barbosa-Morais et al., 2012; Marquez et al., 2012). Using comparative transcriptome analysis, a previous study had found that 23% of chicken genes undergo AS, compared to 68% in humans and 57% in mice (Elsa and Shoba, 2009). Li et al. observed that AS genes

make up approximately 36.85% of a genome, and over 40% of them are specifically expressed during muscle development (Li et al., 2018). There is a growing recognition of the contributions of AS to myogenesis, and the refinement of muscle function, neuronal development, and the function of mature neurons (Bland et al., 2010; Vuong et al., 2016; Nikonova et al., 2020). On one hand, we detected over 45000 putative splicing events, where 56% of genes undergo complex splicing; SE is the most common event, while RI is the rarest, which is consistent with the findings of Rogers et al. (Rogers et al., 2021). The proportion of different splicing types is dynamic, depending on the location of tissues, species, size, and development periods. The muscle and brain showed a similar proportion of five types of events, whereas the liver exhibited slightly different results. Previous studies have shown that A3SS is most common in Luning chickens and sheep, and RI is the major event in muscle tissues of the Gushi chicken (Zhang et al., 2013; Zhang et al., 2017; Li et al., 2018). Additionally, RI is the most common type of event in bovine embryos compared to calves and adults, with the frequency of all splicing events sharply decreasing after birth (Sun et al., 2015). Splicing events are unevenly distributed on chromosomes, with the frequency of events consistent with chromosome length (chromosomes-1 and chromosome-2 are the longest); a similar result was reported in a sheep study (Zhang et al., 2013). For example, in the breast muscle, more events are positioned on chromosome-7, and *MSTN* and *NEB* are specific to muscle development and growth (Donner et al., 2004; Chen P. R. et al., 2019). Most analyses are centered on the effect of spatial and temporal transcriptomes; measurement of tissue-specific differences between two modern chicken lines by us indicated that 7–11% of divergent genes (FDR < 0.05) undergo AS, with more than half the gene detected in one tissue. Compared to the muscle and liver, the brain tissue undergoes more complex splicing, and exhibits fewer divergence events than other tissues, suggesting that it is relatively conserved between broilers and layers. AS was relatively conserved in the brain tissue, with 7% fewer divergent events than in the muscle and liver tissues, indicating high tissue and strain specificity. With over 95% of splicing events being tissue-specific in both proteomics and RNA-seq analyses, the brain has strongly conserved splicing signatures, while the nervous system has also been shown to

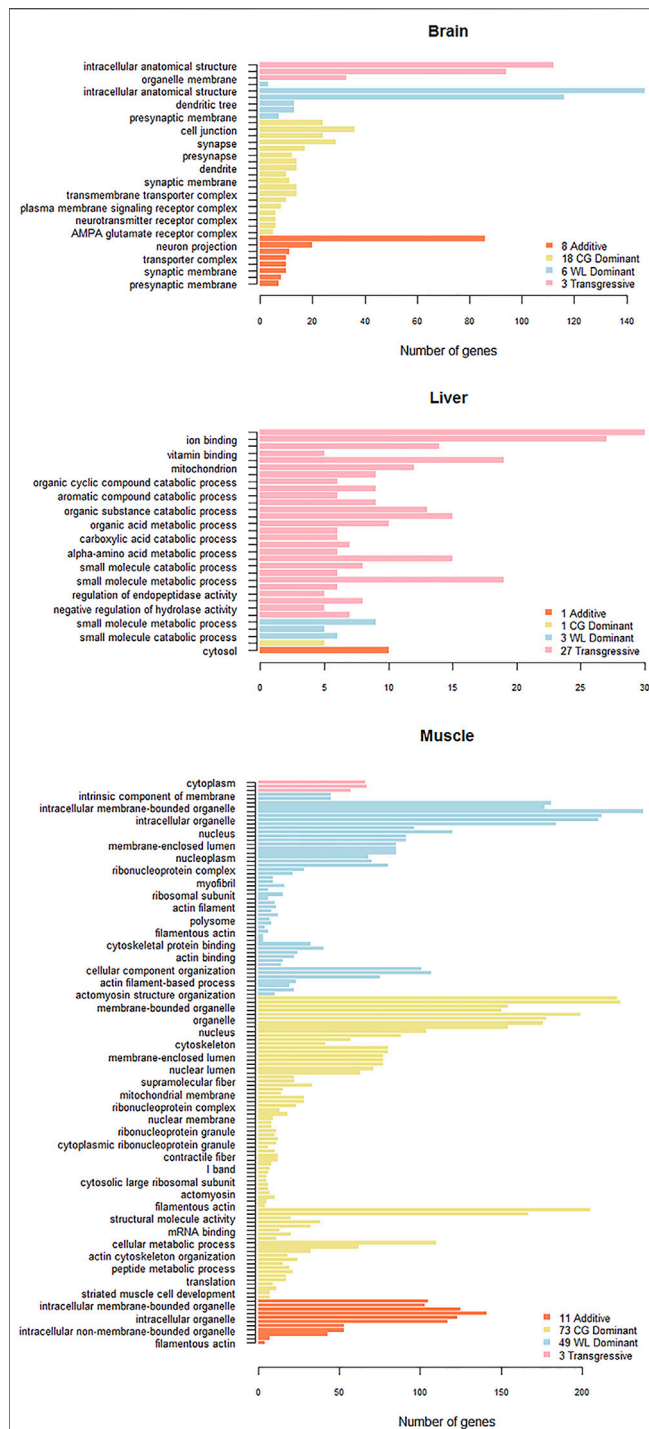


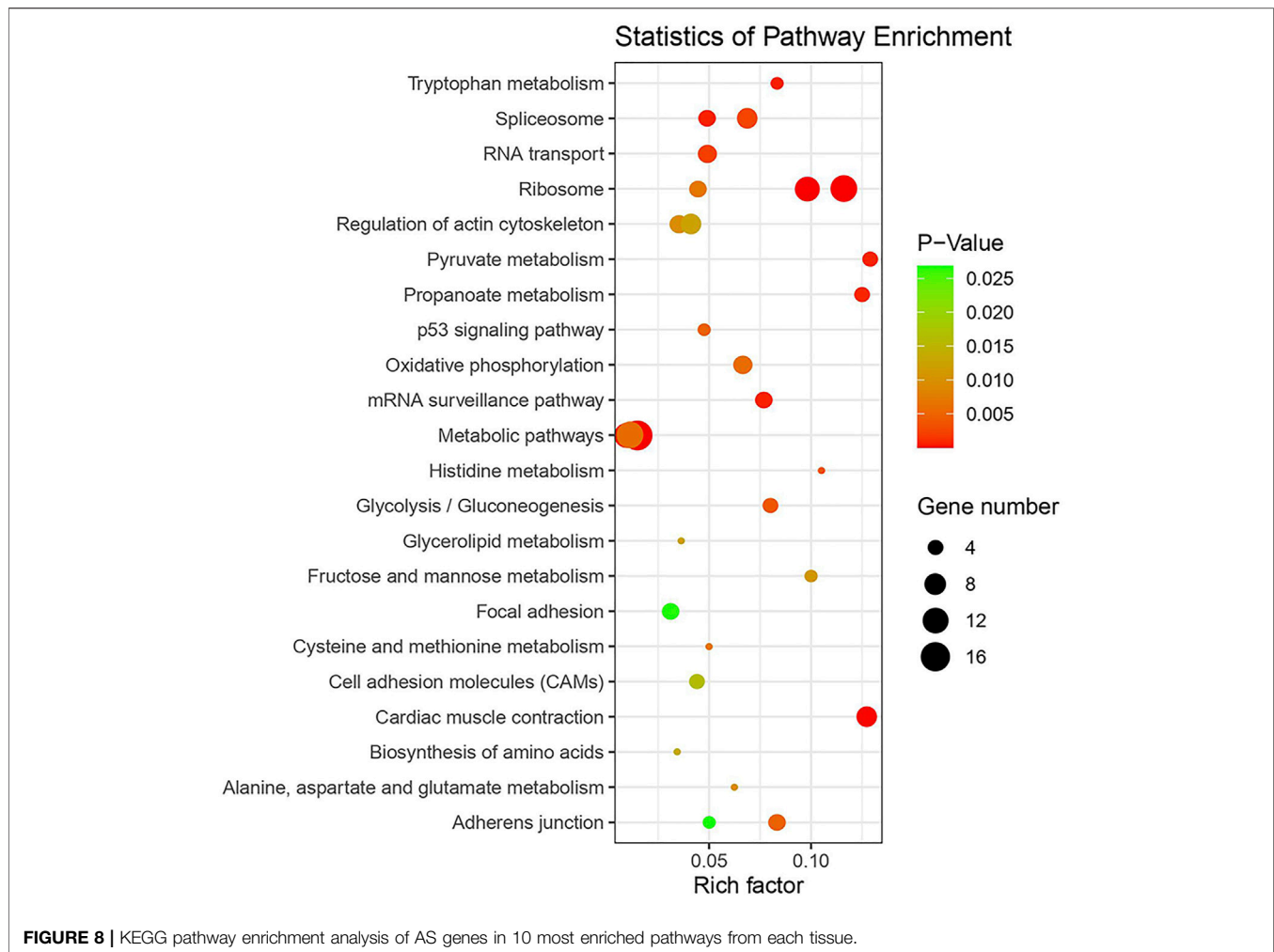
FIGURE 7 | Gene Ontology (GO) analysis of un-conserved pattern AS genes in three tissues. GO functional enrichment analysis is performed on additive, dominant and transgressive AS genes in brain, liver, and muscle.

record many conserved tissue-specific splicing events (Gu et al., 2020; Rodriguez et al., 2020). Merkin et al. also indicated that splicing in the brain is well conserved, through measured transcriptome variations among multiple vertebrate species,

and a similar result was also obtained by gene expression analysis (Merkin et al., 2012; Gu et al., 2020).

Approximately half of the splicing events that are significantly different between CG and WL are classified into nine types due to the divergence between hybrids and parents. Considering that the time of divergence of the two breeds is relatively recent, more than one-third of the events exhibited a conserved pattern in the three tissue types, the other patterns being the additive, dominant, and transgressive patterns. Especially in the muscle tissue, four groups consistently indicate events that are conserved during hybridization (41%); the muscle structure and basic developmental mechanisms are highly conserved (Nikonova et al., 2020). The greater prevalence of this category is also observed in coffee and *Drosophila* hybrids (Marie-Christine et al., 2015; Lopez-Maestre et al., 2017). We recognized that a small proportion of events were additive, while the majority of events were dominant and transgressive; similar categorization of gene expression have been reported by other studies (Gu et al., 2020). The predominance of non-additively expressed splicing in hybrids also supports the idea that most hybridizations can cause “transcriptome shock,” a widely studied phenomenon in hybrid plants, and associated with the divergence between parental species (Hegarty et al., 2006). In addition, the pattern of enhancing dominance was more prevalent than suppressing dominance in most groups, indicating that up-regulated dominant events potentially play important roles in heterosis. Previous studies on body weight in *Drosophila* and gene expression modes in the liver of chickens have shown similar results (Mai et al., 2019). However, due to different cluster methods used, other studies have reported that additivity is the main gene expression pattern in embryonic chickens (Wu et al., 2016; Zhuo et al., 2019). They classified “expression dominance” as being additive, where the genes in hybrids are not significantly different from those of one of the parents and from mid-parental values, and genes in this parent and mid-parental values are significantly different from those of the other parent (Rapp et al., 2009). This approach is more suitable for polyploid-like cotton and *cobitis* (Rapp et al., 2009; Oldrich et al., 2019). The method being unsuitable in our case, we used the Fisher test compared with the threshold criteria (**Supplementary Table S1**), in which more than 60% of events overlapped in each pattern. Stringent criteria may lead to slightly different results, although the two methods still draw the same conclusion.

Functional enrichment analyses in muscle tissue showed pyruvate metabolism, an intersection of key pathways of energy metabolism including *LDHB*, *LDHA*, *PDHA2*, and *ACSS2*. *LDHB* controlled lactate metabolism, and the mRNA and protein expression levels were significantly higher in wooden breast chicken (Zhao et al., 2019). *ACSS2* encodes a cytosolic enzyme that catalyzes the activation of acetate for lipid synthesis and energy generation. In addition, oxidative phosphorylation forms adenosine triphosphate (ATP) as a result of the transfer of electrons from NADH or FADH₂ to O₂ by a series of electron carriers related to six genes namely *ATP6V1A*, *UQCRCQ*, *NDUFA8*, *NDUFV3*, *NDUFB3*, and *UQCRC2*. This pathway is also correlated with body weight in *Drosophila*, and is associated with heterosis in plants (McDaniel and Grimwood, 1971; Katara



et al., 2020). By detecting ATP content and ATPase activity in reciprocal cross chickens, Mai et al. validated that oxidative phosphorylation is the major genetic and molecular factor in growth (Mai et al., 2021). The liver is the major site of amino acid metabolism and fat deposition in the body, and 37 non-conserved AS genes are significantly enriched in metabolic pathways, amino acid metabolism, and glycerolipid metabolism.

The relative frequency of cis-regulatory elements, and trans-acting BP divergence has great influence on the inheritance of gene expression patterns in hybrids (Lemos et al., 2008). Allelic-specific gene expression tests revealed that cis-regulatory divergence is a predominant contributor in mouse and *Camellia*, whereas trans-regulatory divergence is an important driving force in *Drosophila*, and greater parental genetic divergence decreases inheritance of patterns in coffee (Bell et al., 2013; Gao et al., 2015; Marie-Christine et al., 2015; Zhang et al., 2019). Compared with mouse and *Drosophila* inbred lines, the divergence time between commercial chicken strains is relatively short, resulting in a small number of single nucleotide polymorphisms that could be studied. In addition, short read-based data are limited to identifying allele-specific AS events, while full-length transcript sequencing might improve the

accuracy and robustness of AS analysis. Studies have already explored a bulk of new transcript isoforms in chicken (Thomas et al., 2014; Sun et al., 2021); numerous studies have reported that coordinated action between AS and nonsense-mediated RNA decay controls the ratio of productive to unproductive mRNA isoforms (Garcia-Moreno and Romao, 2020). With the development of sequencing and bioinformatic analysis, further studies can explore evolution and interaction with other post-transcriptional mechanisms of AS to better understand the regulation of biological processes.

CONCLUSION

In this study, AS events were observed to be highly specific to tissues and strains, and events in the brain were found to be relatively well-conserved. Meanwhile, inheritance pattern analysis of AS events showed that dominant pattern was predominant excluding conserved pattern, and indicated that a large proportion of inheritance patterns exhibited enhanced parent-specific dominance related to heterosis in chickens. These findings provide new insights into the genetic mechanisms underlying hybridization.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available in NCBI under accession number PRJNA591354.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Care and Use Committee of China Agricultural University.

AUTHOR CONTRIBUTIONS

XQ analyzed and interpreted the sequencing data and drafted the manuscript. HG participated in sample collection, performing reciprocal cross experiment and improved the manuscript. LQ conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

REFERENCES

- Abasht, B., and Lamont, S. J. (2007). Genome-wide Association Analysis Reveals Cryptic Alleles as an Important Factor in Heterosis for Fatness in Chicken F2 Population. *Anim. Genet.* 38, 491–498. doi:10.1111/j.1365-2052.2007.01642.x
- Barbazuk, W. B., Fu, Y., and McGinnis, K. M. (2008). Genome-wide Analyses of Alternative Splicing in Plants: Opportunities and Challenges. *Genome Res.* 18, 1381–1392. doi:10.1101/gr.053678.106
- Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* 338, 1587–1593. doi:10.1126/science.1230612
- Bateson, W. (1910). Mendels Principles of Heredity. *Z. Vererbungslehre* 3, 108–109. doi:10.1007/bf02047719
- Bell, G. D., Kane, N. C., Rieseberg, L. H., and Adams, K. L. (2013). RNA-seq Analysis of Allele-specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. *Genome Biol. Evol.* 5, 1309–1323. doi:10.1093/gbe/evt072
- Bland, C. S., Wang, E. T., Vu, A., David, M. P., Castle, J. C., Johnson, J. M., et al. (2010). Global Regulation of Alternative Splicing during Myogenic Differentiation. *Nucleic Acids Res.* 38, 7651–7664. doi:10.1093/nar/gkq614
- Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: Intelligent Prioritization and Exploratory Visualization of Biological Functions for Gene Enrichment Analysis. *Nucleic Acids Res.* 49, W317–W325. doi:10.1093/nar/gkab447
- Cardoso, T. F., Quintanilla, R., Castelló, A., González-Prendes, R., Amills, M., and Cánovas, Á. (2018). Differential Expression of mRNA Isoforms in the Skeletal Muscle of Pigs with Distinct Growth and Fatness Profiles. *Bmc Genomics* 19, 145. doi:10.1186/s12864-018-4515-2
- Chen, C., Chen, H., Shan, J.-X., Zhu, M.-Z., Shi, M., Gao, J.-P., et al. (2013). Genetic and Physiological Analysis of a Novel Type of Interspecific Hybrid Weakness in Rice. *Mol. Plant* 6, 716–728. doi:10.1093/mp/sss146
- Chen, P. R., Suh, Y., Shin, S., Woodfint, R. M., Hwang, S., and Lee, K. (2019a). Exogenous Expression of an Alternative Splicing Variant of Myostatin Prompts Leg Muscle Fiber Hyperplasia in Japanese Quail. *Ijms* 20, 4617. doi:10.3390/ijms20184617
- Chen, S.-Y., Li, C., Jia, X., and Lai, S.-J. (2019b). Sequence and Evolutionary Features for the Alternatively Spliced Exons of Eukaryotic Genes. *Ijms* 20, 3834. doi:10.3390/ijms20153834
- Chen, Z. J. (2010). Molecular Mechanisms of Polyploidy and Hybrid Vigor. *Trends Plant Sci.* 15, 57–71. doi:10.1016/j.tplants.2009.12.003

FUNDING

This work was supported by the Beijing Innovation Team of the Modern Agro-industry Technology Research System (BAIC04-2021).

ACKNOWLEDGMENTS

We gratefully acknowledge our colleagues in the Poultry Team at the National Engineering Laboratory for Animal Breeding of China Agricultural University, for their assistance on sample collection and helpful comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.774240/full#supplementary-material>

- Clasen, J. B., Norberg, E., Madsen, P., Pedersen, J., and Kargo, M. (2017). Estimation of Genetic Parameters and Heterosis for Longevity in Crossbred Danish Dairy Cattle. *J. Dairy Sci.* 100, 6337–6342. doi:10.3168/jds.2017-12627
- Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and Mode of Regulatory Evolution in Drosophila. *Genome Res.* 24, 797–808. doi:10.1101/gr.163014.113
- Dlamini, Z., Hull, R., Mbatha, S. Z., Alaouna, M., Qiao, Y.-L., Yu, H., et al. (2021). Prognostic Alternative Splicing Signatures in Esophageal Carcinoma. *Cmar* 13, 4509–4527. doi:10.2147/cmar.s305464
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635
- Donner, K., Sandbacka, M., Lehtokari, V.-L., Wallgren-Pettersson, C., and Pelin, K. (2004). Complete Genomic Structure of the Human Nebulin Gene and Identification of Alternatively Spliced Transcripts. *Eur. J. Hum. Genet.* 12, 744–751. doi:10.1038/sj.ejhg.5201242
- Dozmorov, M. G., Adrianto, L., Giles, C. B., Glass, E., Glenn, S. B., Montgomery, C., et al. (2015). Detrimental Effects of Duplicate Reads and Low Complexity Regions on RNA- and ChIP-Seq Data. *Bmc Bioinformatics* 16, S10. doi:10.1186/1471-2105-16-S13-S10
- Elsa, C., and Shoba, R. (2009). Comprehensive Splicing Graph Analysis of Alternative Splicing Patterns in Chicken, Compared to Human and Mouse. *Bmc Genomics* 10, S5. doi:10.1186/1471-2164-10-S1-S5
- Gao, C., Zhai, J., Dang, S., and Zheng, S. (2018). Analysis of Alternative Splicing in Chicken Embryo Fibroblasts in Response to Reticuloendotheliosis Virus Infection. *Avian Pathol.* 47, 585–594. doi:10.1080/03079457.2018.1511047
- Gao, Q., Sun, W., Balleger, M., Libert, C., and Chen, W. (2015). Predominant Contribution of Cis-Regulatory Divergence in the Evolution of Mouse Alternative Splicing. *Mol. Syst. Biol.* 11, 816. doi:10.15252/msb.20145970
- García-Moreno, J. F., and Romão, L. (2020). Perspective in Alternative Splicing Coupled to Nonsense-Mediated mRNA Decay. *Int. J. Mol. Sci.* 21, 9424. doi:10.3390/ijms21249424
- Gu, H., Qi, X., Jia, Y., Zhang, Z., and Qu, L. (2020). Publisher Correction: Inheritance Patterns of the Transcriptome in Hybrid Chickens and Their Parents Revealed by Expression Analysis. *Scientific Rep.* 10. doi:10.1038/s41598-020-63873-0
- Han, H., Sun, X., Xie, Y., Feng, J., and Zhang, S. (2014). Transcriptome and Proteome Profiling of Adventitious Root Development in Hybrid Larch (*Larix Kaempferi* X *Larix Olgensis*). *BMC Plant Biol.* 14. doi:10.1186/s12870-014-0305-4
- Hegarty, M. J., Ba Rker, G. L., Wilson, I. D., Abbott, R. J., Edwards, K. J., and Hiscock, S. J. (2006). Transcriptome Shock after Interspecific Hybridization in

- senecio Is Ameliorated by Genome Duplication. *Curr. Biol.* 16, 1652–1659. doi:10.1016/j.cub.2006.06.071
- Jones, D. F. (1917). Dominance of Linked Factors as a Means of Accounting for Heterosis. *Proc. Natl. Acad. Sci. United States America* 3, 310–312. doi:10.1073/pnas.3.4.310
- Katara, J. L., Verma, R. L., Parida, M., Ngangkham, U., and Mohapatra, T. (2020). Differential Expression of Genes at Panicle Initiation and Grain Filling Stages Implied in Heterosis of Rice Hybrids. *Int. J. Mol. Sci.* 21, 1–17. doi:10.3390/ijms21031080
- Lemos, B., Araripe, L. O., Fontanillas, P., and Hartl, D. L. (2008). Dominance and the Evolutionary Accumulation of Cis- and Trans-effects on Gene Expression. *Proc. Natl. Acad. Sci. United States America* 105, 14471–14476. doi:10.1073/pnas.0805160105
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352
- Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., and Li, X. (2008). Dominance, Overdominance and Epistasis Condition the Heterosis in Two Heterotic Rice Hybrids. *Genetics* 180, 1725–1742. doi:10.1534/genetics.108.091942
- Li, Z., Xu, Y., and Lin, Y. (2018). Transcriptome Analyses Reveal Genes of Alternative Splicing Associated with Muscle Development in Chickens. *Gene* 676, 146–155. doi:10.1016/j.gene.2018.07.027
- Lopez-Maestre, H., Carmelossi, E., Lacroix, V., Burlet, N., Mugat, B., Chambeyron, S., et al. (2017). Identification of Misexpressed Genetic Elements in Hybrids between *Drosophila*-Related Species. *Scientific Rep.* 7, 40618. doi:10.1038/srep40618
- Mai, C., Wen, C., Sun, C., Xu, Z., and Yang, S. (2019). Implications of Gene Inheritance Patterns on the Heterosis of Abdominal Fat Deposition in Chickens. *Genes* 10, 824. doi:10.3390/genes10100824
- Mai, C., Wen, C., Xu, Z., Xu, G., Chen, S., Zheng, J., et al. (2021). Genetic Basis of Negative Heterosis for Growth Traits in Chickens Revealed by Genome-wide Gene Expression Pattern Analysis. *J. Anim. Sci. Biotechnol.* 12, 52. doi:10.1186/s40104-021-00574-2
- Marie-Christine, C., Yann, H., Alexis, D., Stéphanie, R., Juan-Carlos, H., and Philippe, L. (2015). Regulatory Divergence between Parental Alleles Determines Gene Expression Patterns in Hybrids. *Genome Biol. Evol.* 7, 1110–1121. doi:10.1093/gbe/evv057
- Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome Survey Reveals Increased Complexity of the Alternative Splicing Landscape in Arabidopsis. *Genome Res.* 22, 1184–1195. doi:10.1101/gr.134106.111
- Mcdaniel, R. G., and Grimwood, B. G. (1971). Hybrid Vigor in *Drosophila*: Respiration and Mitochondrial Energy Conservation. *Comp. Biochem. Physiol. B Comp. Biochem.* 38, 309–314. doi:10.1016/0305-0491(71)90009-5
- Mcmanus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory Divergence In *Drosophila* Revealed By Mrna-Seq. *Genome Res.* 20, 816–825.
- Mehmood, A., Laiho, A., Veninen, M. S., Mcglinchey, A. J., and Elo, L. L. (2019). Systematic Evaluation of Differential Splicing Tools for RNA-Seq Studies. *Brief. Bioinform.* 21, 2052–2065. doi:10.1093/bib/bbz126
- Merkin, J., Russell, C., Ping, C., and Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* 338, 1593–1599. doi:10.1126/science.1228186
- Montes, M., Sanford, B. L., Comiskey, D. F., and Chandler, D. S. (2019). RNA Splicing and Disease: Animal Models to Therapies. *Trends Genet.* 35, 68–87. doi:10.1016/j.tig.2018.10.002
- Nikonova, E., Kao, S. Y., and Spletter, M. L. (2020). Contributions of Alternative Splicing to Muscle Type Development and Function. *Semin. Cell Dev Biol* 104, 65–80. doi:10.1016/j.semcdb.2020.02.003
- Oldřich, B., Jan, R., Jan, K., Jan, P., Ladislav, P., Miloslav, P., et al. (2019). The Legacy of Sexual Ancestors in Phenotypic Variability, Gene Expression, and Homeolog Regulation of Asexual Hybrids and Polyploids. *Mol. Biol. Evol.* 36, 1902–1920. doi:10.1093/molbev/msz114
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing. *Nat. Genet.* 40, 1413–1415. doi:10.1038/ng.259
- Patel, R. K., Mukesh, J., and Liu, Z. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE* 7, e30619. doi:10.1371/journal.pone.0030619
- Rapp, R. A., Udall, J. A., and Wendel, J. F. (2009). Genomic Expression Dominance in Allopolyploids. *BMC Biol.* 7, 18. doi:10.1186/1741-7007-7-18
- Rodriguez, J. M., Pozo, F., Di Domenico, T., Vazquez, J., and Tress, M. L. (2020). An Analysis of Tissue-specific Alternative Splicing at the Protein Level. *Plos Comput. Biol.* 16, e1008287. doi:10.1371/journal.pcbi.1008287
- Rogers, T. F., Palmer, D. H., and Wright, A. E. (2021). Sex-Specific Selection Drives the Evolution of Alternative Splicing in Birds. *Mol. Biol. Evol.* 38, 519–530. doi:10.1093/molbev/msaa242
- Shen, S., Won, P. J., Huang, J., Dittmar, K. A., Zhi-Xiang, L., Zhou, Q., et al. (2012). MATS: a Bayesian Framework for Flexible Detection of Differential Alternative Splicing from RNA-Seq Data. *Nucleic Acids Res.* 40, e61. doi:10.1093/nar/gkr1291
- Shull, G. H. (1908). The Composition of a Field of Maize. 4 296–301. doi:10.1093/jhered/os-4.1.296
- Sun, C., Jin, K., Zuo, Q., Sun, H., Song, J., Zhang, Y., et al. (2021). Characterization of Alternative Splicing (AS) Events during Chicken (*Gallus gallus*) Male Germ-Line Stem Cell Differentiation with Single-Cell RNA-Seq. *Animals (Basel)* 11, 1469. doi:10.3390/ani11051469
- Sun, X., Li, M., Sun, Y., Cai, H., and Chen, H. (2015). The Developmental Transcriptome Landscape of Bovine Skeletal Muscle Defined by Ribo-Zero Ribonucleic Acid Sequencing. *J. Anim. Sci.* 93, 5648–5658. doi:10.2527/jas.2015-9562
- Swanson-Wagner, R., Jia, Y., Decook, R., Borsuk, L., Nettleton, D., and Schnable, P. (2006). All Possible Modes of Gene Action Are Observed in a Global Comparison of Gene Expression in a maize F1 Hybrid and its Inbred Parents. *Proc. Natl. Acad. Sci. United States America* 103, 6805–6810. doi:10.1073/pnas.0510430103
- Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H. Y., Karlak, B., Daverman, R., et al. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* 13, 2129–2141. doi:10.1101/gr.772403
- Thomas, S., Underwood, J. G., Tseng, E., Holloway, A. K., and Bench to Basinet Cv, D. C. I. S. (2014). Long-read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms. *PLoS One* 9, e94650. doi:10.1371/journal.pone.0094650
- Vuong, C. K., Black, D. L., and Zheng, S. (2016). The Neurogenetics of Alternative Splicing. *Nat. Rev. Neurosci.* 17, 265–281. doi:10.1038/nrn.2016.27
- Wang, J., Pan, Y., Shen, S., Lin, L., and Xing, Y. (2017). rMATS-DVR: rMATS Discovery of Differential Variants in RNA. *Bioinformatics* 14, 2216–2217. doi:10.1093/bioinformatics/btx128
- Weatheritt, R. J., Sterne-Weiler, T., and Blencowe, B. J. (2016). The Ribosome-Engaged Landscape of Alternative Splicing. *Nat. Struct. Mol. Biol.* 23, 1117–1123. doi:10.1038/nsmb.3317
- Whitney, K. D., Ahern, J. R., Campbell, L. G., Albert, L. P., and King, M. S. (2010). Patterns of Hybridization in Plants. *Perspect. Plant Ecol. Evol. Syst.* 12, 175–182. doi:10.1016/j.ppees.2010.02.002
- Wu, X., Li, R., Li, Q., Bao, H., and Wu, C. (2016). Comparative Transcriptome Analysis Among Parental Inbred and Crosses Reveals the Role of Dominance Gene Expression in Heterosis in *Drosophila melanogaster*. *Scientific Rep.* 6. doi:10.1038/srep21124
- Zhang, C., Wang, G., Wang, J., Ji, Z., Liu, Z., Pi, X., et al. (2013). Characterization and Comparative Analyses of Muscle Transcriptomes in Dorper and Small-Tailed Han Sheep Using RNA-Seq Technique. *PLoS One* 8, e72686. doi:10.1371/journal.pone.0072686
- Zhang, J., Cheng, C., Xiaobin, P., Guangrong, L., Liang, C., Yang, Z., et al. (2015). Transcriptome Analysis of Interspecific Hybrid between Brassica Napus and B. Rapa Reveals Heterosis for Oil Rape Improvement. *Int. J. Genomics* 2015, 230985. doi:10.1155/2015/230985
- Zhang, M., Tang, Y. W., Qi, J., Liu, X. K., and Zhang, W. J. (2019). Effects of Parental Genetic Divergence on Gene Expression Patterns in Interspecific Hybrids of Camellia. *BMC Genomics* 2, 120. doi:10.1186/s12864-019-6222-z
- Zhang, Y., Li, D., Han, R., Wang, Y., Li, G., Liu, X., et al. (2017). Transcriptome Analysis of the Pectoral Muscles of Local Chickens and Commercial Broilers Using Ribo-Zero Ribonucleic Acid Sequencing. *PLoS One* 12, e0184115. doi:10.1371/journal.pone.0184115
- Zhao, D., Kogut, M. H., Genovese, K. J., Hsu, C. Y., and Farnell, Y. Z. (2019). Altered Expression of Lactate Dehydrogenase and Monocarboxylate Transporter Involved in Lactate Metabolism in Broiler Wooden Breast. *Poult. Sci.* 99. doi:10.3382/ps/pez572

- Zhao, S. (2019). Alternative Splicing, RNA-Seq and Drug Discovery. *Drug Discov. Today* 24, 1258–1267. doi:10.1016/j.drudis.2019.03.030
- Zhuang, Y., and Tripp, E. A. (2017). Genome-scale Transcriptional Study of Hybrid Effects and Regulatory Divergence in an F1 Hybrid *Ruellia* (Wild Petunias: Acanthaceae) and its Parents. *BMC Plant Biol.* 17, 1–13. doi:10.1186/s12870-016-0962-6
- Zhuo, Z., Lamont, S. J., and Abasht, B. (2019). RNA-seq Analyses Identify Additivity as the Predominant Gene Expression Pattern in F1 Chicken Embryonic Brain and Liver. *Genes* 10. doi:10.3390/genes10010027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Qi, Gu and Qu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Pleiotropic Loci Associated With Foot Disorders and Common Periparturient Diseases in Holstein Cattle

Ellen Lai, Alexa L. Danner, Thomas R. Famula and Anita M. Oberbauer*

Animal Science Department, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Peipei Ma,
Shanghai Jiao Tong University, China
Sayed Haidar Abbas Raza,
Northwest A and F University, China

*Correspondence:

Anita M. Oberbauer
amoberbauer@ucdavis.edu

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 17 July 2021

Accepted: 01 November 2021

Published: 06 December 2021

Citation:

Lai E, Danner AL, Famula TR and
Oberbauer AM (2021) Pleiotropic Loci
Associated With Foot Disorders and
Common Periparturient Diseases in
Holstein Cattle.
Front. Genet. 12:742934.
doi: 10.3389/fgene.2021.742934

Lameness is an animal welfare issue that incurs substantial financial and environmental costs. This condition is commonly caused by digital dermatitis (DD), sole ulcers (SU), and white line disease (WLD). Susceptibility to these three foot disorders is due in part to genetics, indicating that genomic selection against these foot lesions can be used to reduce lameness prevalence. It is unclear whether selection against foot lesions will lead to increased susceptibility to other common diseases such as mastitis and metritis. Thus, the aim of this study was to determine the genetic correlation between causes of lameness and other common health disorders to identify loci contributing to the correlation. Genetic correlation estimates between SU and DD and between SU and WLD were significantly different from zero ($p < 0.05$), whereas estimates between DD and mastitis, DD and milk fever, and SU and metritis were suggestive ($p < 0.1$). All five of these genetic correlation estimates were positive. Two-trait genome-wide association studies (GWAS) for each of these five pairs of traits revealed common regions of association on BTA1 and BTA8 for pairs that included DD or SU as one of the traits, respectively. Other regions of association were unique to the pair of traits and not observed in GWAS for other pairs of traits. The positive genetic correlation estimates between foot disorders and other health disorders imply that selection against foot disorders may also decrease susceptibility to other health disorders. Linkage disequilibrium blocks defined around significant and suggestive SNPs from the two-trait GWAS included genes and QTL that were functionally relevant, supporting that these regions included pleiotropic loci.

Keywords: pleiotropy, multivariate, genome-wide association study, dairy cattle, lameness, disease, genetic correlation

INTRODUCTION

Abnormal gait or posture in a cow are considered indicators of lameness and signifies pain and discomfort. Lameness is the second most prevalent disease after mastitis and the third most common reason for culling after mastitis and infertility (USDA, 2018). Lameness is commonly caused by foot lesions classified as infectious [e.g., digital dermatitis (DD), heel horn erosion, and foot rot] or noninfectious lesions [e.g., sole hemorrhage, sole ulcer (SU), white line disease (WLD), and

Abbreviations: *Bos taurus* autosome (BTA), digital dermatitis (DD), genetic relatedness matrix (GRM), Genetic Type I error calculator (GEC), genome-wide association study (GWAS), likelihood ratio test (LRT), linkage disequilibrium (LD), minor allele frequency (MAF), quantitative trait loci (QTL), single nucleotide polymorphism (SNP), sole ulcers (SU), white line disease (WLD).

laminitis]. Lameness not only raises welfare concerns, but also has economic and environmental consequences. Financial costs associated with lameness include direct costs for treatment and increased labor and indirect costs from reduced milk production and fertility; together, these costs range from \$64 per case of DD to \$178 per case of SU (Cha et al., 2010; Dolecheck and Bewley, 2018; Dolecheck et al., 2019). Reduced fertility, premature culling, and reduced milk production associated with lameness reduces the efficiency of resource use, as resources used for the cow are invested over a less productive and shorter lifetime, inflating the environmental costs per unit of milk by 14 (1.5%) kg CO₂ equivalents per ton of fat-and-protein-corrected milk, on average (of DD, SU, and WLD combined) (Mostert et al., 2018).

Prevention of lameness is achieved through routine claw trimming, foot baths (for prevention of infectious lesions), maintaining floor hygiene, and nutrition. Despite these prevention efforts, lameness remains highly prevalent in the United States, affecting 16.8% of cows and 3.2% of bred heifers (USDA, 2018). These non-genetic methods of prevention can be aided by genetic selection, as implied by the low to moderate estimates of heritability for foot lesions, ranging from 0.01 to 0.4 for DD, 0.01 to 0.3 for SU, and 0.017 to 0.26 for WLD (Van der Waaij et al., 2005; Onyiro et al., 2008; van der Linde et al., 2010; Häggman and Juga, 2013; Oberbauer et al., 2013; van der Spek et al., 2013, 2015; Malchiodi et al., 2015; Biemans et al., 2018). Genetic selection uses prior knowledge about the contribution of certain genetic markers to traits of interest and creating a selection index reflecting a weighted average of multiple traits that is used to rank animals. Selective breeding programs utilize both genetic correlation among traits that are included in the selection index and specific susceptibility loci associated with the traits. Accordingly, selection against foot disorders would likely account for correlated lesion traits because some foot disorders are genetically correlated with each other, particularly within the infectious (strongest between DD and heel erosion) and noninfectious (strongest among sole hemorrhage, SU, and WLD) groupings of lesions (Koenig et al., 2005; Van der Waaij et al., 2005; van der Linde et al., 2010; Buch et al., 2011; Gernand et al., 2012; Häggman and Juga, 2013; van der Spek et al., 2013; Pérez-Cabal and Charfeddine, 2015; Malchiodi et al., 2017).

Additionally, certain foot lesions are genetically correlated with mastitis or indicator traits of mastitis. For example, the genetic correlations between clinical mastitis and sole hemorrhage or SU were estimated at 0.35 and 0.32, respectively, in Swedish Red cows (Buch et al., 2011). For Holstein cows, the genetic correlation between somatic cell score and individual foot lesions or lameness in general ranged from 0.15 to 0.24 (Koenig et al., 2005) and 0.23 (Gernand et al., 2012), respectively, although other studies failed to identify significant genetic correlations between DD or interdigital hyperplasia and clinical mastitis (Buch et al., 2011; Gernand et al., 2013). Nevertheless, the genetic correlation among foot disorders and between individual foot disorders and mastitis traits implies that common loci may coordinately influence these traits (Koenig et al., 2005; Buch et al., 2011). Such pleiotropic loci

have not been identified as of yet. The values for genetic correlation between foot and other health disorders that have been reported were estimated using pedigree information. To our knowledge, no DNA-based studies have been performed to estimate the genetic correlation between foot disorders and disease traits other than mastitis. Using genomic data from individual cows to estimate relationships may be more accurate than using pedigree data (Goddard, 2009; Hayes et al., 2009) because using genomic data reduces the standard error of the genetic correlation estimate (Visscher et al., 2014). Therefore, the aim of this study was to identify loci associated with susceptibility to multiple foot disorders and other common diseases, which could be coordinately used to inform breeding programs.

METHODS

All procedures were conducted in accordance with ethical standards set by the University of California, Davis and approved by the Institutional Animal Care and Use Committee (protocol #22099).

Phenotypes

Five large commercial dairies (Dairies A–E, each with >1,000 cows) in Northern and Central California participated in this study. Phenotypes were derived from hoof trimming and other health records provided by the dairies beginning from the cows' first lactation. Three hoof trimmers recorded the foot lesions used for phenotyping foot lesions: one who serviced Dairies A, B, and C; one who serviced Dairy D; and another who serviced Dairy E. Hoof trimmer experience and hoof trimming regimens were described previously (Lai et al., 2020, Lai et al., 2021). Foot disorders recorded included DD, foot rot, sole hemorrhage, SU, WLD, wall abscess, sole abscess, heel abscess, and laminitis. Other health events were also recorded by dairy personnel, which included diarrhea, displaced abomasum, ketosis, mastitis, metritis, milk fever, pneumonia, and retained placenta. For each foot or other health disorder, cases were defined as cows with at least one record of the disorder and controls were defined as cows that did not have records of the given foot or health disorder. Consequently, for each trait, controls included cows with disorders other than the disorder the cases had.

Genotypes

Whole blood samples were obtained and the buffy coat was used to extract genomic DNA using the QIAGEN QIAamp DNA Blood Mini Kit (QIAGEN Inc., Valencia, CA). DNA samples were quantified using the NanoDrop (ND-2000 v3.2.1) spectrophotometer (Thermo Scientific, Wilmington, DE) and sent to GeneSeek (Lincoln, NE) for SNP genotyping on the high-density BovineHD BeadChip (777K SNPs, Illumina Inc., San Diego, CA). Genotype calls were made using Illumina's GenCall algorithm. SNP genotypes from a subset of the cows used in this study were used in our past studies (Lai et al., 2020, Lai et al., 2021) and are publicly available at the NCBI Gene

Expression Omnibus database (GEO series record GSE159157 and GSE165945), along with the additional samples from this study (GSE to be added when received from GEO). SNP genotypes were updated to the ARS-UCD1.2 assembly positions (Rosen et al., 2020) and quality-filtered in PLINK 1.9 (Purcell and Chang, 2015) by removing from further analyses SNPs and cows with <95% genotyping rate, SNPs with significant deviation from Hardy-Weinberg equilibrium ($p < 1E-6$) to exclude systematic genotyping errors, and SNPs with minor allele frequency <5% to exclude rare variants. Missing genotypes for each cow were imputed using BEAGLE 5.1 (Browning et al., 2018) using the other cows in the sample population as the reference population, an effective sample size of 58 for the United States Holstein cattle population (Makanjuola et al., 2020), and default parameters. Genetic similarity among cows was visualized in a multidimensional scaling (MDS) plot depicting the first two dimensions.

Estimation of Genetic Correlation

Genetic correlation was estimated between each foot lesion and other health traits, including other foot lesions (e.g., genetic correlation was estimated between SU and WLD, SU and DD, SU and mastitis, and SU and metritis) using cows that had phenotypes for both traits and at least 40 case cows for each disease. PLINK 2.0 was used to filter cows by requiring phenotypes for both traits (Chang et al., 2015; Purcell and Chang, 2021). GCTA was used to calculate the genetic relatedness matrix (GRM), which was used with farm as a fixed effect to estimate the additive genetic variance and covariance between the two traits using two-trait GREML (Yang et al., 2011; Lee et al., 2012). Specifically, the phenotype for trait 1 of the k th cow from the i th farm at the j th SNP was modeled as

$$y_{1ijk} = \mu_1 + F_{1i} + S_{1j} + a_{1ik} + \varepsilon_{1ijk}$$

where μ_1 was an unknown constant common to all cows for trait 1, F_{1i} was contribution of i th farm to the risk of disease, S_{1j} was the contribution of the j th SNP genotype to risk of the disorder, and a_{1ik} were the additive genetic effects assumed to be drawn from the multivariate normal density $N(0, A\sigma_a^2)$, where A was the GRM. ε_{1ijk} was the residual term for trait 1. Similarly, the phenotype for trait 2 of the k th cow from the i th farm at the j th SNP was modeled using the same components for trait 2 as

$$y_{2ijk} = \mu_2 + F_{2i} + S_{2j} + a_{2ik} + \varepsilon_{2ijk}$$

The covariance of additive genetic effects was computed as a function of the numerator relationship matrix, and genetic correlation was calculated as the covariance of additive genetic effects divided by the product of the standard deviations of the genetic effect of traits 1 and 2. All genetic correlation estimates were transformed from the observed scale (0/1) to the underlying liability scale to account for case ascertainment using the prevalence of each disorder obtained from the literature (Oberbauer et al., 2013; USDA, 2018). Genetic correlation estimates were considered significantly different from zero if the estimate had $p < 0.05$ from the likelihood ratio test, and suggestive genetic correlation estimates were those with $p < 0.1$.

Two-Trait Genome-Wide Association Study

Pairs of traits that had significant or suggestive genetic correlation estimates using the frequentist approach were evaluated further in two-trait GWAS to identify regions potentially contributing to both traits. Multi-trait association testing can improve the power to detect associations while accounting for population stratification (Banerjee et al., 2008; Korte et al., 2012; Zhou and Stephens, 2012, Zhou and Stephens, 2014) because the additional information from the covariance of traits is still informative, even if only one of the traits is associated with the genotype (Stephens, 2013). Two-trait genome-wide association analysis was performed to test for association of each SNP with at least one of the traits. A standardized GRM was constructed and included in the linear mixed model to account for relatedness and population stratification, and farm was included as a fixed effect to adjust for differences among farms. The linear mixed model association testing was conducted using the multivariate association testing function in GEMMA (Zhou and Stephens, 2012, Zhou and Stephens, 2014) using the same models for estimating genetic correlation. Bonferroni correction for multiple testing assumes that each test for SNP association with phenotype(s) is independent. However, because SNPs are not independent due to linkage disequilibrium (LD) between SNPs, the Genetic Type I error calculator (GEC) was used to calculate the effective number of markers after accounting for linkage disequilibrium between SNPs for use as the denominator in Bonferroni-corrected thresholds of significance (Li et al., 2012). Genome-wide significant SNPs were thus defined as those with likelihood ratio test (LRT) $p < 0.05/M_e$ and suggestive SNPs, as those with LRT $p < 1/M_e$ (Lander and Kruglyak, 1995). Manhattan and quantile-quantile plots were generated using the qqman package in R (R Development Core Team, 2010; Turner, 2014).

Because SNPs may not be causal for the traits but rather in LD with causal variants due in part to the long range linkage disequilibrium in cattle (Cai et al., 2019), SNPs were used to define LD blocks that were then mined for overlap with genes and previously defined QTL. SNPs in LD with significant and suggestive SNPs were used to define the start and end of LD blocks using a method similar to Richardson et al. (2016) and Twomey et al. (2019). SNPs that were within 5 Mb (upstream or downstream) and in LD ($R^2 > 0.5$) with significant or suggestive SNPs were considered belonging to the same LD block. LD blocks were queried in the region search of FAANGMine (FAANG, 2019) to identify genes within or overlapping with the LD block. LD blocks were also queried for overlap with previously defined QTL and associations related to feet and legs conformation traits and disease traits in the Cattle QTLdb (Hu et al., 2019) (version 46, accessed 4/30/2021).

RESULTS

Descriptive Data

Hoof trimming records were available for 21,044 cows across the five dairies (distribution of records for each type of foot lesion is described in detail by Lai et al. (2021), of which 417 cows were

TABLE 1 | Count of genotyped cows after quality filtering, split by cases for each foot disorder or other health condition and controls across the five dairies.

	Farm					Total
	A	B	C	D	E	
<i>Datasets for foot disorders</i>						
Sole ulcer						
Cases	44	8	4	71	25	152
Controls	138	70	26	23	0	257
White line disease						
Cases	48	13	7	33	16	117
Controls	134	65	23	61	9	292
Digital dermatitis						
Cases	19	22	30	30	5	106
Controls	163	56	0	64	20	303
<i>Datasets for other disorders</i>						
Mastitis						
Cases	89	66	NR	77	17	249
Controls	93	12	NR	17	8	130
Metritis						
Cases	57	51	NR	8	15	131
Controls	125	27	NR	86	10	248
Ketosis						
Cases	13	17	NR	NR	0	30
Controls	169	61	NR	NR	25	255
Retained placenta						
Cases	16	35	NR	NR	0	234
Controls	166	43	NR	NR	25	51
Diarrhea						
Cases	19	0	NR	NR	1	20
Controls	163	78	NR	NR	24	265
Milk fever						
Cases	61	9	NR	NR	0	70
Controls	121	69	NR	NR	25	215
Displaced abomasum						
Cases	1	17	NR	NR	2	20
Controls	181	61	NR	NR	23	265
Pneumonia						
Cases	2	4	NR	22	13	41
Controls	180	74	NR	72	12	338

NR, No records available; cows were excluded from analyses.

selected for SNP genotyping as controls or cases for a certain foot lesion(s). Traits that were recorded at multiple dairies were used for genetic correlation estimation and two-trait GWAS, and the distribution of case/control phenotypes for each trait is listed in **Table 1**. All five dairies recorded SU, WLD, and DD foot disorders. All dairies except Dairy C also had health records available for phenotyping other health traits. These four dairies (Dairies A, B, D, and E) recorded mastitis, metritis, and pneumonia. Dairies A, B, and E also recorded ketosis, retained placenta, diarrhea, milk fever, and displaced abomasum. After excluding traits that have ≤ 40 cases, genetic correlation was estimated between each pair of foot disorders (SU, WLD, and DD) as well as each foot disorder with another health disorder (mastitis, metritis, retained placenta, milk fever, and pneumonia).

Quality filtering removed eight cows and 218,306 SNPs, leaving 409 cows with 559,656 SNPs for analyses with case/control phenotypes presented in **Table 1**. The MDS plot indicated slight population stratification with a prominent center cluster, though cows were not strongly stratified by farm (**Figure 1**).

Genetic Correlation Estimates

Of the pairs of traits for which genetic correlation was estimated, genetic correlation estimates between SU and WLD and between SU and DD were significantly different from zero ($p < 0.05$), and estimates between DD and mastitis, DD and milk fever, and SU and metritis were suggestive ($p < 0.1$, **Table 2**). Consequently, each pair of these traits was analyzed in two-trait GWAS.

Two-Trait Genome-Wide Association Analysis

The effective number of markers after accounting for LD was 162,435 SNPs, corresponding to a suggestive threshold of 6.2×10^{-6} [5.2 on the $-\log_{10}(p)$ scale] for genome-wide suggestive significance and 3.1×10^{-7} [6.5 on the $-\log_{10}(p)$ scale] for genome-wide significance. Manhattan plots from the two-trait GWAS are shown in **Figure 2**. Genomic inflation factors ranged from 1.02 to 1.06 and, combined with the qqplots (**Supplementary Figure S1**), indicated that population stratification had been sufficiently accounted for (Price et al., 2010).

Significant and suggestive SNPs and the LD blocks they defined are shown in **Table 3**. Supplemental materials report the genes and QTL LD blocks (**Supplementary Tables S1–S3**). The GWAS that included DD as one of the traits (DD and mastitis, SU and DD, and DD and milk fever) identified significant and suggestive SNPs belonging to the same LD block at BTA1:125550933–125822143. For the DD and mastitis and DD and milk fever GWAS, the peak on BTA1 reached or approached genome-wide significance despite the genetic correlation estimate only reaching suggestive significance (**Table 1** and **Supplementary Table S1**). GWAS that included SU as one of the traits (that is, between SU and WLD, SU and DD, and SU and metritis) all identified suggestive SNPs in an LD block at BTA8:42926603–44642925. Other SNP associations were unique to the pair of traits for which the GWAS was performed such that SNPs that were associated in a certain GWAS for a pair of traits were not associated in other GWAS for other pairs of traits. For instance, the LD block on BTA14 detected from the GWAS for SU and DD was only detected in the SU and DD GWAS and not detected in any other of the comparisons such as that for SU and WLD, DD and mastitis, DD and milk fever, and SU and metritis. Although SU and WLD were strongly genetically correlated (0.92), the suggestive SNPs identified in the two-trait GWAS had opposite effect signs between the two traits: if the effect of the SNP was positive for SU, it was negative for WLD and *vice versa* (**Table 3** and **Supplementary Table S1**). The LD blocks defined from all the two-trait GWAS overlapped with 83 protein-coding genes, some functionally relevant to the etiology of the disorders (**Supplementary Table S2**).

DISCUSSION

We estimated the genetic correlation between common foot disorders (DD, SU, and WLD) and other health traits

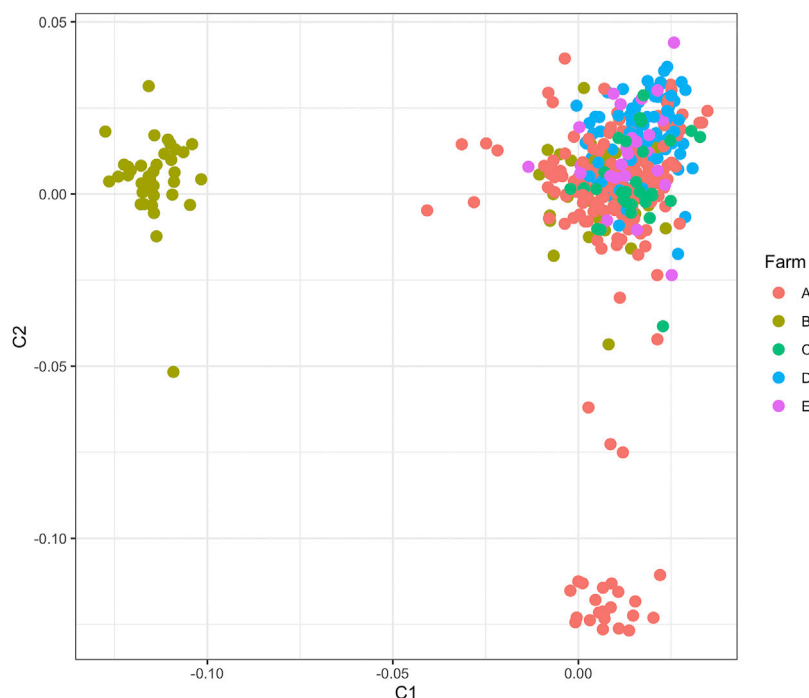


FIGURE 1 | Multidimensional scaling plot showing the first two dimensions for 409 cows from the five dairies used in the estimation of genetic correlation and genome-wide association analyses.

TABLE 2 | Genetic correlation estimates (and standard error, SE) between sole ulcer (SU), white line disease (WLD), digital dermatitis (DD), and other foot or health traits that were significantly or suggestively different from zero.

Trait 1	Trait 2	Genetic correlation (SE)	<i>p</i>	Significance
SU	DD	0.46 (0.25)	4.81E-02	^a
SU	WLD	0.92 (0.46)	2.54E-02	^a
DD	Mastitis	0.49 (0.36)	7.77E-02	^b
DD	Milk fever	0.49 (0.39)	9.46E-02	^b
SU	Metritis	0.7 (0.46)	5.22E-02	^b

^aSignificant.

^bSuggestive significance.

(mastitis, metritis, milk fever, retained placenta, and pneumonia). For pairs of traits having significant or suggestive genetic correlation, the loci that were contributing to the correlation were examined using two-trait GWAS. To our knowledge, this is the first study to estimate genetic correlation between foot disorders and diseases other than mastitis from individual-level genotype data rather than pedigree data and identify loci potentially contributing to the correlation. Genetic correlation estimates that were significant or suggestive included SU or DD as one of the traits and estimates were positive, indicating a favorable genetic correlation between pairs of disease traits such that genetic selection against one disease will lead to selection against the other disease. Significant and suggestive SNPs were detected in the same regions on BTA1 and BTA8 for two-trait GWAS datasets that had DD and SU as one of the traits, respectively, suggesting that DD and SU were driving the

association in these genomic regions. Other significant and suggestive SNPs were specific to the dataset from which they were detected and not detected in GWAS for other pairs of traits.

Compared to previous estimates of genetic correlation between foot disorders and other health traits, estimates from this study were higher and had larger standard errors. Previous estimates of genetic correlation between foot disorders and mastitis or somatic cell count were significantly different from zero (0.15–0.35) (Koenig et al., 2005; Buch et al., 2011) or close to zero (Gernand et al., 2012), whereas we estimated the genetic correlation between DD and mastitis at 0.49 (SE = 0.36). The genetic correlation between SU and WLD was 0.92 (SE = 0.46) and substantially higher than previous estimates, which ranged from 0.41 to 0.60 (van der Linde et al., 2010). The estimates of genetic correlation from this study were higher likely because controls were shared between the two traits and the proportion of cows with DD and/or SU was higher than for other disorders. Because case cows were sampled primarily for DD and SU and other disorders were phenotyped after sampling DD and SU cases, cases for other disorders frequently also had DD and/or SU. This overrepresentation of cases with DD and/or SU in addition to the disorder of interest likely inflated genetic correlation estimates, which the correction for case ascertainment was unable to overcome. The strong genetic correlation between SU and WLD in this study implied that whichever other traits SU is correlated with, WLD will also be correlated with and *vice versa*; however, SU was correlated with metritis and DD whereas WLD was not correlated with either disorder. This divergence would suggest that although SU and WLD share a genetic component, differences exist in the location or direction of the effect for susceptibility loci between

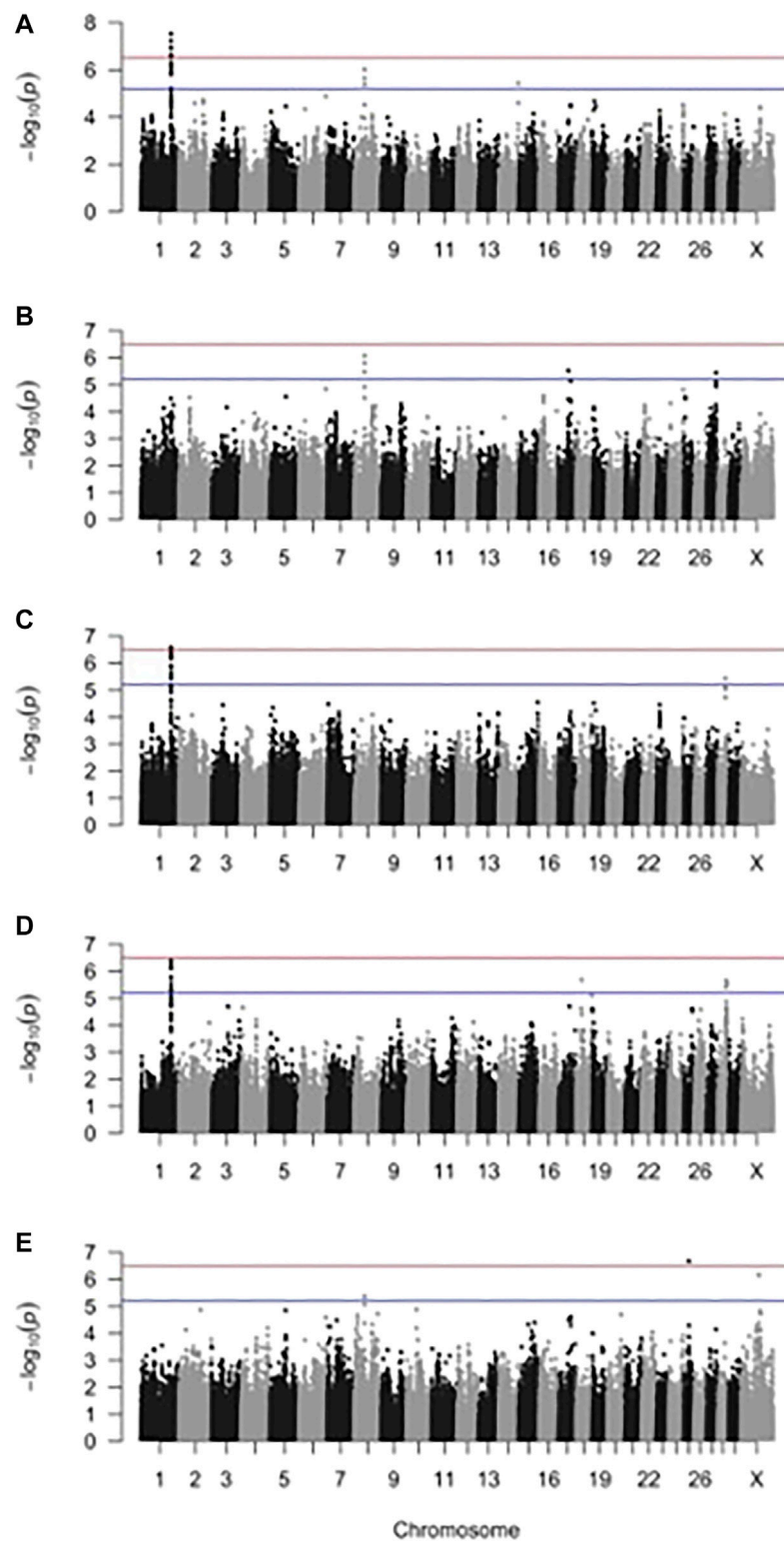


FIGURE 2 | Manhattan plot for two-trait genome-wide association analysis of **(A)** sole ulcer (SU) and digital dermatitis (DD), **(B)** SU and white line disease (WLD), **(C)** DD and mastitis, **(D)** DD and milk fever, and **(E)** SU and metritis. The blue line indicates genome-wide suggestive significance, and the red line indicates genome-wide significance.

TABLE 3 | Linkage disequilibrium (LD) blocks and the most significant SNP within the LD block defined from the two-trait genome-wide association analyses of sole ulcer (SU), white line disease (WLD), and digital dermatitis (DD) paired with each other and other health disorders.

Dataset (trait 1 and trait 2)	BTA	LD block start (bp)	LD block end (bp)	LD block length (kb)	Most significant SNP	SNP position (bp)	Minor/ Major allele	Effect size for trait 1	Effect size for trait 2	Variance matrix for beta effects			p
										Variance of effect size for trait 1	Covariance between effect sizes of trait 1 and 2	Variance of effect size for trait 2	
SU and DD	1	125550933	125822143	271.21	BovineHD0100035768	125563251	A/G	1.85E-01	4.64E-03	1.09E-03	3.13E-05	1.46E-03	2.58E-07 ^a
	8	42926603	44642925	1716.322	BovineHD0800013406	44628587	T/C	2.05E-01	-8.65E-02	2.30E-03	6.91E-04	2.99E-03	3.06E-06 ^b
	14	81655298	81664096	8.798	BovineHD1400023802	81655298	G/T	-3.00E-02	1.58E-01	1.05E-03	3.17E-04	1.21E-03	3.68E-06 ^b
SU and WLD	8	42926603	44642925	1716.322	BovineHD0800013408	44632844	G/T	4.90E-02	1.91E-01	9.97E-04	-9.75E-06	1.14E-03	2.11E-07 ^a
	17	41328134	41328134	0	BovineHD1700011766	41328134	C/T	-1.21E-01	1.32E-01	1.12E-03	8.92E-05	1.23E-03	7.07E-07 ^b
	27	37518206	38922466	1,404.26	BovineHD2700011209	38898651	T/C	-1.21E-01	1.32E-01	1.12E-03	8.92E-05	1.23E-03	7.07E-07 ^b
DD and mastitis	27	37518206	38922466	1,404.26	BovineHD2700011210	38901656	G/A	-1.21E-01	1.32E-01	1.12E-03	8.92E-05	1.23E-03	7.07E-07 ^b
	1	125550933	125822143	271.21	BovineHD0100035835	125691064	A/G	1.02E-01	1.72E-01	1.06E-03	1.61E-04	9.47E-04	2.98E-08 ^a
	28	33357088	33385923	28.835	BovineHD2800009006	33385923	C/T	9.39E-02	1.62E-01	1.01E-03	1.59E-04	9.07E-04	1.20E-07 ^a
DD and milk fever	1	125550933	125822143	271.21	BovineHD0100035785	125585828	C/T	8.88E-02	1.64E-01	1.01E-03	1.61E-04	9.08E-04	1.19E-07 ^a
	1	125550933	125822143	271.21	BovineHD0100035800	125624770	A/C	8.88E-02	1.64E-01	1.01E-03	1.61E-04	9.08E-04	1.19E-07 ^a
	18	24087895	24329676	241.781	BovineHD1800007458	24087895	C/T	6.46E-02	1.50E-01	9.06E-04	1.51E-04	8.12E-04	8.13E-07 ^b
	28	34935232	35093950	158.718	BTB-00987935	35093950	G/T	6.46E-02	1.50E-01	9.06E-04	1.51E-04	8.12E-04	8.13E-07 ^b
	28	35837718	36740498	902.78	BovineHD2800010153	36916301	C/T	6.52E-02	1.49E-01	8.93E-04	1.47E-04	8.00E-04	7.86E-07 ^b
	28	35837718	36740498	902.78	BovineHD2800010156	36926419	C/T	6.52E-02	1.49E-01	8.93E-04	1.47E-04	8.00E-04	7.86E-07 ^b
	28	38776483	42482917	3706.434	BovineHD2800011177	40061500	G/A	8.15E-02	1.45E-01	9.42E-04	1.53E-04	8.54E-04	1.40E-06 ^b
SU and metritis	8	42926603	44642925	1716.322	BovineHD0800013412	44642925	A/C	6.16E-02	1.57E-01	9.06E-04	1.50E-04	8.03E-04	2.22E-07 ^a
	25	22127459	22966511	839.052	BovineHD2500006264	22127459	A/G	1.86E-01	-2.77E-02	1.52E-03	2.99E-04	1.49E-03	4.09E-06 ^b
	X	75319558	75610976	291.418	BovineHD3000022324	75393744	A/G	2.01E-01	-1.21E-02	1.49E-03	2.90E-04	1.46E-03	9.61E-07 ^b

^aGenome-wide significance.^bGenome-wide suggestive significance.

SU and WLD, as indicated by the opposite signs of suggestive SNP effects between the traits and the lack of association of WLD to metritis or DD in the two-trait GWAS.

Compared to our previous one-trait GWAS for DD and SU (Lai et al., 2020, 2021), the two-trait GWAS detected the same LD block on BTA1 for DD and a different LD block on BTA8 for SU. Specifically, the LD block at BTA1:125550933–125822143 common to all datasets that had DD as one of the traits (DD and mastitis, SU and DD, and DD and milk fever) was the same LD block detected in our previous single-trait DD GWAS (Lai et al., 2020). The increase in significance of association also suggests that this region may play a role in both infectious (mastitis) and metabolic (SU and milk fever) disorders. Infectious and metabolic disorders have been observed to coincide and happen most frequently during the early lactation period (USDA, 2018), potentially due to a common cause. Some have attributed the cause of higher incidence of infectious and noninfectious foot disorders during early lactation to the extreme negative energy balance during this period (Collard et al., 2000; Gernand et al., 2013). Accordingly, it is thought that cows that are better able to cope with the energy requirements during this period are consequently less susceptible to metabolic and infectious disorders, a hypothesis supported by the association of a more robust adaptive immune response with lower incidence of metabolic disease during the periparturient period (Thompson-Crispi et al., 2012). Another common LD block at BTA8:42926603–44642925 was detected from the two-trait GWAS with SU as one of the traits (SU and WLD, SU and DD, and SU and metritis). This LD block was 30 Mb upstream of the LD block on BTA8 observed in our previous one-trait SU GWAS (Lai et al., 2021). Our previous GWAS used the same SU cases but only sound, older (>6.0 years old) cows as controls, whereas the present GWAS included controls with foot disorders other than the foot disorder the cases had. Consequently, the present GWAS controlled for other foot disorders that the cases had such that associated regions were more likely for SU specifically and not for other foot disorders correlated with SU, whereas the single-trait GWAS used the most phenotypically divergent cows as controls to maximize the power to detect genetic differences.

In addition to detecting the same regions as the one-trait GWAS, the two-trait GWAS detected other regions that were not detected in the one-trait GWAS for DD, SU, and/or WLD. The DD and mastitis GWAS detected a region on BTA28 that the DD GWAS did not find (Lai et al., 2020). The SU and DD two-trait GWAS found a region on BTA14 that was not identified in either the DD or SU one-trait GWAS (Lai et al., 2020, Lai et al., 2021). The suggestive SNP on BTA17 from the two-trait GWAS of SU and WLD was not in LD with other SNPs and not detected in the one-trait GWAS for SU or WLD (Lai et al., 2021). The two-trait GWAS for DD and milk fever detected regions on BTA18 and 28 that were not detected in the one-trait DD GWAS (Lai et al., 2020). Finally, the two-trait GWAS for DD and milk fever detected regions on BTA18 and 28 that were not detected in the one-trait DD GWAS (Lai et al., 2020). The genetic correlation between the two traits may have provided additional power to detect these associations that were underpowered in the one-trait GWAS.

The LD blocks defined from each dataset overlapped with genes and/or QTL that were functionally relevant to both traits. Genes

having functions that were considered relevant to the etiology of each disorder were defined for each trait (**Supplementary Table S4**) and included those with a role in immune function, hair follicle morphology, hair density, skin integrity, fibroblast proliferation, bone growth and mineralization, adipose and body fat, and glucose metabolism. The LD block at BTA1:125550933–125822143 from the GWAS that included DD as one of the traits contained *SLC9A9* (solute carrier family 9 member A9) (Lai et al., 2020), which has been implicated in multiple sclerosis in humans through its role in regulating T-cell activation and differentiation to induce a proinflammatory response (Esposito et al., 2015). Notably, the DD and mastitis LD block at 1:125839933–125852054 overlapped with a QTL associated with length of productive life (Cole et al., 2011), corroborating the shorter productive life associated with DD and mastitis susceptibility (Shabalina et al., 2020). Previous estimates of genetic correlation between foot lesion traits and productive life were close to zero (Dhakal et al., 2015), suggesting that uncorrelated traits may still share pleiotropic loci, as observed previously between various production, fertility, and conformation traits (Xiang et al., 2017). This LD block on BTA1 from the DD and mastitis GWAS and the LD block on BTA27 from the SU and WLD GWAS both overlapped with QTL for feet and legs conformation traits (Cole et al., 2011), and could be a pleiotropic locus contributing to the genetic correlation between feet and legs conformation and susceptibility to foot lesions (Häggman and Juga, 2013; Malchiodi et al., 2017; Ring et al., 2018), though this genetic correlation is too low to justify indirect selection on lameness using feet and legs conformation traits (Van Raden et al., 2021). The LD blocks from the GWAS for SU and DD, SU and WLD, and SU and metritis overlap with QTL for infectious disease traits (tuberculosis susceptibility, clinical mastitis, and somatic cell score/count) and blood cortisol, which may reflect the interplay of the stress from the negative energy balance during the periparturient period possibly potentiating metabolic and infectious foot disorders (**Supplementary Table S3**). Cows with SU tend to exhibit markers of chronic inflammation compared to cows without SU (O'Driscoll et al., 2015), though it is unclear if SU causes inflammation, *vice versa*, or both are the product of stress.

The main limitations of this study were the small sample size of genotyped cows and the variation in the number of case cows across the various disorders. At the expense of a larger sample size, we minimized the environmental variation by constraining the sample population to cows to a small geographical region under similar management and nutrition practices and minimized the number of hoof trimmers to reduce variation in phenotyping foot lesions. Minimizing environmental and consistent phenotyping improves the power to detect significant genetic correlation; however, the resulting small sample size limited the accuracy of genetic correlation estimates. For instance, one workaround for the inflation of genetic correlation estimates due to shared controls is to randomly partition the controls between the two traits before estimating genetic correlation; however, the small sample size prevented using this approach. The small sample size also limited the benefit of using genomic data instead of pedigree data to estimate genetic correlation. Although using genomic data to estimate relationships may be more accurate than using pedigree data (Goddard, 2009; Hayes et al., 2009)

due to reduced standard error of the genetic correlation estimate (Visscher et al., 2014), the standard error of the genetic correlation estimates in this study was large, reflecting the limited sample size. The reduction in standard error from using genomic data would be more appreciable in larger sample sizes. Ascertainment bias for cows with DD and SU but not the other disorders likely led to an overrepresentation of cows with DD and/or SU in the dataset, resulting in inflated estimates between DD or SU and the other disorders. Despite the inflated and large standard errors of the genetic correlation estimates, some estimates were significantly or suggestively different from zero and provided grounds for further investigation of SNPs contributing to the correlation using the two-trait GWAS. The sample size also provided sufficient power in the two-trait GWAS to detect significant and suggestive SNPs that defined LD blocks overlapping with functionally relevant genes and QTL, similar to previous GWAS using similar small sample sizes (~400 cows) and high-density SNP genotypes (Buzanskas et al., 2017; Lehner et al., 2018).

CONCLUSION

A genomic relatedness matrix calculated from SNP genotypes was used to estimate genetic correlation between individual foot disorders (DD, SU, and WLD) and other health disorders (mastitis, metritis, milk fever, retained placenta, and pneumonia). The positive estimates of genetic correlation between individual foot disorders and other health disorders indicate that direct selection against foot disorders will not increase the incidence of other health disorders and may in fact reduce their prevalence. Genomic assessment for pairs of traits that were genetically correlated revealed multiple associated regions. Whereas some of these chromosomal regions were shared across multiple pairs of traits that included SU or DD as one of the traits, others were unique to the pair of traits, indicating the complexity of genetic contributions within and between traits. The LD blocks defined from associated SNPs included protein-coding genes and QTL that were functionally relevant to both traits, suggesting that selection for markers in these LD blocks would affect susceptibility to both traits.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://www.ncbi.nlm.nih.gov/geo/>, GSE159157, GSE165945, and GSE186266.

REFERENCES

- Banerjee, S., Yandell, B. S., and Yi, N. (2008). Bayesian Quantitative Trait Loci Mapping for Multiple Traits. *Genetics* 179, 2275–2289. doi:10.1534/genetics.108.088427
- Biemans, F., Bijma, P., Boots, N. M., and de Jong, M. C. M. (2018). Digital Dermatitis in Dairy Cattle: The Contribution of Different Disease Classes to Transmission. *Epidemics* 23, 76–84. doi:10.1016/j.epidem.2017.12.007
- Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103, 338–348. doi:10.1016/j.ajhg.2018.07.015
- Buch, L. H., Sørensen, A. C., Lassen, J., Berg, P., Eriksson, J.-A., Jakobsen, J. H., et al. (2011). Hygiene-related and Feed-Related Hoof Diseases Show Different Patterns of Genetic Correlations to Clinical Mastitis and Female Fertility. *J. Dairy Sci.* 94, 1540–1551. doi:10.3168/jds.2010-3137
- Buzanskas, M. E., Grossi, D. d. A., Ventura, R. V., Schenkel, F. S., Chud, T. C. S., Stafuzza, N. B., et al. (2017). Candidate Genes for Male and Female

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

AO, TF, AD, and EL conceptualized the research aims. With supervision from AO and TF, AD and EL arranged blood sample collection and processing, curated hoof trimming records, developed the methodology, and performed the computational analyses. TF was instrumental in developing code for the computational analyses that EL adapted for the dataset. AO provided the funding, lab space, and computing resources for this research. AO, TF, EL, and AD prepared and edited the article.

FUNDING

This research was funded by a Western Sustainable Agriculture Research and Education Graduate Student grant (project number GW18-126); U.S. Department of Agriculture Cooperative State Research, Education, and Extension Service Animal Health Funding (project number 2250-AH); a Department of Animal Science Kellogg Endowment; and Jastro Shields awards from the College of Agricultural and Environmental Sciences at the University of California, Davis.

ACKNOWLEDGMENTS

We thank the dairy producers, dairy personnel, and hoof trimmers who participated in this study. We also gratefully acknowledge the infrastructure support of the Department of Animal Science, College of Agricultural and Environmental Sciences, and the California Agricultural Experiment Station of the University of California, Davis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.742934/full#supplementary-material>

- Reproductive Traits in Canchim Beef Cattle. *J. Anim. Sci. Biotechnol* 8, 1–10. doi:10.1186/s40104-017-0199-8
- Cai, Z., Guldbrandsen, B., Lund, M. S., and Sahana, G. (2019). Dissecting Closely Linked Association Signals in Combination with the Mammalian Phenotype Database Can Identify Candidate Genes in Dairy Cattle. *BMC Genet.* 20, 15. doi:10.1186/s12863-019-0717-0
- Cha, E., Hertl, J. A., Bar, D., and Gröhn, Y. T. (2010). The Cost of Different Types of Lameness in Dairy Cows Calculated by Dynamic Programming. *Prev. Vet. Med.* 97, 1–8. doi:10.1016/j.prevetmed.2010.07.011
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8
- Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., et al. (2011). Genome-wide Association Analysis of Thirty One Production, Health, Reproduction and Body Conformation Traits in Contemporary U.S. Holstein Cows. *BMC Genomics* 12, 408. doi:10.1186/1471-2164-12-408
- Collard, B. L., Boettcher, P. J., Dekkers, J. C. M., Petitclerc, D., and Schaeffer, L. R. (2000). Relationships between Energy Balance and Health Traits of Dairy Cattle in Early Lactation. *J. Dairy Sci.* 83, 2683–2690. doi:10.3168/jds.S0022-0302(00)75162-9
- Dhakal, K., Tiezzi, F., Clay, J. S., and Maltecca, C. (2015). Short Communication: Genomic Selection for Hoof Lesions in First-Parity US Holsteins. *J. Dairy Sci.* 98, 3502–3507. doi:10.3168/jds.2014-8830
- Dolecheck, K. A., Overton, M. W., Mark, T. B., and Bewley, J. M. (2019). Use of a Stochastic Simulation Model to Estimate the Cost Per Case of Digital Dermatitis, Sole Ulcer, and white Line Disease by Parity Group and Incidence Timing. *J. Dairy Sci.* 102, 715–730. doi:10.3168/jds.2018-14901
- Dolecheck, K., and Bewley, J. (2018). Animal Board Invited Review: Dairy Cow Lameness Expenditures, Losses and Total Cost. *Animal* 12, 1462–1474. doi:10.1017/S1751731118000575
- Esposito, F., Sorosina, M., Ottoboni, L., Lim, E. T., Replogle, J. M., Raj, T., et al. (2015). A Pharmacogenetic Study implicates SLC9A9 in Multiple Sclerosis Disease Activity. *Ann. Neurol.* 78, 115–127. doi:10.1002/ana.24429
- FAANG (2019). “Functional Annotation of Animal Genomes (FAANG) Consortium,” in *FAANGMine*. Available at: <http://128.206.116.18:8080/faangmine/begin.do> (Accessed August 3, 2020).
- Gernand, E., Döhne, D. A., and König, S. (2013). Genetic Background of Claw Disorders in the Course of Lactation and Their Relationships with Type Traits. *J. Anim. Breed. Genet.* 130, 435–444. doi:10.1111/jbg.12046
- Gernand, E., Rehbein, P., von Borstel, U. U., and König, S. (2012). Incidences of and Genetic Parameters for Mastitis, Claw Disorders, and Common Health Traits Recorded in Dairy Cattle Contract Herds. *J. Dairy Sci.* 95, 2144–2156. doi:10.3168/jds.2011-4812
- Goddard, M. (2009). Genomic Selection: Prediction of Accuracy and Maximisation of Long Term Response. *Genetica* 136, 245–257. doi:10.1007/s10709-008-9308-0
- Hägman, J., and Juga, J. (2013). Genetic Parameters for Hoof Disorders and Feet and Leg Conformation Traits in Finnish Holstein Cows. *J. Dairy Sci.* 96, 3319–3325. doi:10.3168/jds.2012-6334
- Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased Accuracy of Artificial Selection by Using the Realized Relationship Matrix. *Genet. Res.* 91, 47–60. doi:10.1017/S0016672308009981
- Hu, Z.-L., Park, C. A., and Reecy, J. M. (2019). Building a Livestock Genetic and Genomic Information Knowledgebase through Integrative Developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* 47, D701–D710. doi:10.1093/nar/gky1084
- Koenig, S., Sharifi, A. R., Wentrot, H., Landmann, D., Eise, M., and Simianer, H. (2005). Genetic Parameters of Claw and Foot Disorders Estimated with Logistic Models. *J. Dairy Sci.* 88, 3316–3325. doi:10.3168/jds.S0022-0302(05)73015-0
- Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A Mixed-Model Approach for Genome-wide Association Studies of Correlated Traits in Structured Populations. *Nat. Genet.* 44, 1066–1071. doi:10.1038/ng.2376
- Lai, E., Danner, A. L., Famula, T. R., and Oberbauer, A. M. (2020). Genome-Wide Association Studies Reveal Susceptibility Loci for Digital Dermatitis in Holstein Cattle. *Animals* 10, 2009. doi:10.3390/ani10112009
- Lai, E., Danner, A. L., Famula, T. R., and Oberbauer, A. M. (2021). Genome-Wide Association Studies Reveal Susceptibility Loci for Noninfectious Claw Lesions in Holstein Dairy Cattle. *Front. Genet.* 12, 728. doi:10.3389/FGENE.2021.657375
- Lander, E., and Kruglyak, L. (1995). Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results. *Nat. Genet.* 11, 241–247. doi:10.1038/ng1195-241
- Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012). Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood. *Bioinformatics* 28, 2540–2542. doi:10.1093/bioinformatics/bts474
- Lehner, S., Zerbin, I., Doll, K., Rehage, J., and Distl, O. (2018). A Genome-wide Association Study for Left-Sided Displacement of the Abomasum Using a High-Density Single Nucleotide Polymorphism Array. *J. Dairy Sci.* 101, 1258–1266. doi:10.3168/jds.2017-13216
- Li, M.-X., Yeung, J. M. Y., Cherny, S. S., and Sham, P. C. (2012). Evaluating the Effective Numbers of Independent Tests and Significant P-Value Thresholds in Commercial Genotyping Arrays and Public Imputation Reference Datasets. *Hum. Genet.* 131, 747–756. doi:10.1007/s00439-011-1118-2
- Makanjuola, B. O., Miglior, F., Abdalla, E. A., Maltecca, C., Schenkel, F. S., and Baes, C. F. (2020). Effect of Genomic Selection on Rate of Inbreeding and Coancestry and Effective Population Size of Holstein and Jersey Cattle Populations. *J. Dairy Sci.* 103, 5183–5199. doi:10.3168/jds.2019-18013
- Malchiodi, F., Koeck, A., Chapinal, N., Sargolzaei, M., Fleming, A., Kelton, D. F., et al. (2015). Genetic Analyses of Hoof Lesions in Canadian Holsteins Using an Alternative Contemporary Group. *Interbull Bull.*
- Malchiodi, F., Koeck, A., Mason, S., Christen, A. M., Kelton, D. F., Schenkel, F. S., et al. (2017). Genetic Parameters for Hoof Health Traits Estimated with Linear and Threshold Models Using Alternative Cohorts. *J. Dairy Sci.* 100, 2828–2836. doi:10.3168/jds.2016-11558
- Mostert, P. F., van Middelaar, C. E., de Boer, I. J. M., and Bokkers, E. A. M. (2018). The Impact of Foot Lesions in Dairy Cows on Greenhouse Gas Emissions of Milk Production. *Agric. Syst.* 167, 206–212. doi:10.1016/j.agry.2018.09.006
- Oberbauer, A. M., Berry, S. L., Belanger, J. M., McGoldrick, R. M., Pinos-Rodriguez, J. M., and Famula, T. R. (2013). Determining the Heritable Component of Dairy Cattle Foot Lesions. *J. Dairy Sci.* 96, 605–613. doi:10.3168/jds.2012-5485
- O’Driscoll, K., McCabe, M., and Earley, B. (2015). Differences in Leukocyte Profile, Gene Expression, and Metabolite Status of Dairy Cows with or without Sole Ulcers. *J. Dairy Sci.* 98, 1685–1695. doi:10.3168/jds.2014-8199
- Onyiro, O. M., Andrews, L. J., and Brotherstone, S. (2008). Genetic Parameters for Digital Dermatitis and Correlations with Locomotion, Production, Fertility Traits, and Longevity in Holstein-Friesian Dairy Cows. *J. Dairy Sci.* 91, 4037–4046. doi:10.3168/jds.2008-1190
- Pérez-Cabal, M. A., and Charfeddine, N. (2015). Models for Genetic Evaluations of Claw Health Traits in Spanish Dairy Cattle. *J. Dairy Sci.* 98, 8186–8194. doi:10.3168/jds.2015-9562
- Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New Approaches to Population Stratification in Genome-wide Association Studies. *Nat. Rev. Genet.* 11, 459–463. doi:10.1038/nrg2813
- Purcell, S., and Chang, C. (2021). PLINK 2.0. Available at: <https://www.cog-genomics.org/plink/2.0/> (Accessed April 20, 2021).
- Purcell, S. M., and Chang, C. C. (2015). PLINK 1.9. Available at: <https://www.cog-genomics.org/plink2> (Accessed October 19, 2021).
- R Development Core Team (2010). R: A Language and Environment for Statistical Computing. Available at: <https://www.r-project.org/> (Accessed April 14, 2021).
- Richardson, I. W., Berry, D. P., Wiencko, H. L., Higgins, I. M., More, S. J., McClure, J., et al. (2016). A Genome-wide Association Study for Genetic Susceptibility to Mycobacterium Bovis Infection in Dairy Cattle Identifies a Susceptibility QTL on Chromosome 23. *Genet. Sel. Evol.* 48, 19. doi:10.1186/s12711-016-0197-x
- Ring, S. C., Twomey, A. J., Byrne, N., Kelleher, M. M., Pabiou, T., Doherty, M. L., et al. (2018). Genetic Selection for Hoof Health Traits and Cow Mobility Scores Can Accelerate the Rate of Genetic Gain in Producer-Scored Lameness in Dairy Cows. *J. Dairy Sci.* 101, 10034–10047. doi:10.3168/jds.2018-15009
- Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). De Novo assembly of the Cattle Reference Genome with Single-Molecule Sequencing. *Gigascience* 9, 1–9. doi:10.1093/gigascience/giaa021

- Shabalina, T., Yin, T., and König, S. (2020). Influence of Common Health Disorders on the Length of Productive Life and Stayability in German Holstein Cows. *J. Dairy Sci.* 103, 583–596. doi:10.3168/jds.2019-16985
- Stephens, M. (2013). A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS One* 8, e65245. doi:10.1371/journal.pone.0065245
- Thompson-Crispi, K. A., Hine, B., Quinton, M., Miglior, F., and Mallard, B. A. (2012). Short Communication: Association of Disease Incidence and Adaptive Immune Response in Holstein Dairy Cows. *J. Dairy Sci.* 95, 3888–3893. doi:10.3168/jds.2011-5201
- Turner, S. D. (2014). Qqman: an R Package for Visualizing GWAS Results Using Q-Q and manhattan Plots. *bioRxiv*, 5165. doi:10.1101/005165
- Twomey, A. J., Berry, D. P., Evans, R. D., Doherty, M. L., Graham, D. A., and Purfield, D. C. (2019). Genome-wide Association Study of Endo-Parasite Phenotypes Using Imputed Whole-Genome Sequence Data in Dairy and Beef Cattle. *Genet. Sel. Evol.* 51, 1–17. doi:10.1186/s12711-019-0457-7
- USDA (2018). *Dairy 2014, - Health and Management Practices on U.S. Dairy Operations, 2014*. Fort Collins, CO: USDA–APHIS–VS–CEAH–NAHMS.
- van der Linde, C., de Jong, G., Koenen, E. P. C., and Eding, H. (2010). Claw Health index for Dutch Dairy Cattle Based on Claw Trimming and Conformation Data. *J. Dairy Sci.* 93, 4883–4891. doi:10.3168/jds.2010-3183
- van der Spek, D., van Arendonk, J. A. M., and Bovenhuis, H. (2015). Genetic Relationships between Claw Health Traits of Dairy Cows in Different Parities, Lactation Stages, and Herds with Different Claw Disorder Frequencies. *J. Dairy Sci.* 98, 6564–6571. doi:10.3168/jds.2015-9561
- van der Spek, D., van Arendonk, J. A. M., Vallée, A. A. A., and Bovenhuis, H. (2013). Genetic Parameters for Claw Disorders and the Effect of Preselecting Cows for Trimming. *J. Dairy Sci.* 96, 6070–6078. doi:10.3168/jds.2013-6833
- Van der Waaij, E. H., Holzhauer, M., Ellen, E., Kamphuis, C., and De Jong, G. (2005). Genetic Parameters for Claw Disorders in Dutch Dairy Cattle and Correlations with Conformation Traits. *J. Dairy Sci.* 88, 3672–3678. doi:10.3168/jds.s0022-0302(05)73053-8
- Van Raden, P. M., Cole, J. J., and Parker Gaddis, K. L. (2021). Net merit as a Measure of Lifetime Profit: 2021 Revision. Available at: https://www.ars.usda.gov/ARUserFiles/80420530/Publications/ARR/nmcalc-2021_ARR-NM8.pdf.
- Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G.-B., Lee, S. H., Wray, N. R., et al. (2014). Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *Plos Genet.* 10, e1004269. doi:10.1371/journal.pgen.1004269
- Xiang, R., MacLeod, I. M., Bolormaa, S., and Goddard, M. E. (2017). Genome-wide Comparative Analyses of Correlated and Uncorrelated Phenotypes Identify Major Pleiotropic Variants in Dairy Cattle. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-09788-9
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011
- Zhou, X., and Stephens, M. (2014). Efficient Multivariate Linear Mixed Model Algorithms for Genome-wide Association Studies. *Nat. Methods* 11, 407–409. doi:10.1038/nmeth.2848
- Zhou, X., and Stephens, M. (2012). Genome-wide Efficient Mixed-Model Analysis for Association Studies. *Nat. Genet.* 44, 821–824. doi:10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lai, Danner, Famula and Oberbauer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Significance and Relevance of Spermatozoal RNAs to Male Fertility in Livestock

Bijayalaxmi Sahoo¹, Ratan K. Choudhary², Paramajeet Sharma², Shanti Choudhary² and Mukesh Kumar Gupta^{1*}

¹Department of Biotechnology and Medical Engineering, National Institute of Technology Rourkela, Rourkela, India, ²College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, India

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Guillermo Giovambattista,
CONICET Institute of Veterinary
Genetics (IGEVE), Argentina
İsmaıl Kudret Sağlam,
Koç University, Turkey

*Correspondence:

Mukesh Kumar Gupta
guptam@nitrkl.ac.in

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 31 August 2021

Accepted: 15 November 2021

Published: 09 December 2021

Citation:

Sahoo B, Choudhary RK, Sharma P,
Choudhary S and Gupta MK (2021)
Significance and Relevance of
Spermatozoal RNAs to Male Fertility
in Livestock.
Front. Genet. 12:768196.
doi: 10.3389/fgene.2021.768196

Livestock production contributes to a significant part of the economy in developing countries. Although artificial insemination techniques brought substantial improvements in reproductive efficiency, male infertility remains a leading challenge in livestock. Current strategies for the diagnosis of male infertility largely depend on the evaluation of semen parameters and fail to diagnose idiopathic infertility in most cases. Recent evidences show that spermatozoa contains a suit of RNA population whose profile differs between fertile and infertile males. Studies have also demonstrated the crucial roles of spermatozoal RNA (spRNA) in spermatogenesis, fertilization, and early embryonic development. Thus, the spRNA profile may serve as unique molecular signatures of fertile sperm and may play pivotal roles in the diagnosis and treatment of male fertility. This manuscript provides an update on various spRNA populations, including protein-coding and non-coding RNAs, in livestock species and their potential role in semen quality, particularly sperm motility, freezability, and fertility. The contribution of seminal plasma to the spRNA population is also discussed. Furthermore, we discussed the significance of rare non-coding RNAs (ncRNAs) such as long ncRNAs (lncRNAs) and circular RNAs (circRNAs) in spermatogenic events.

Keywords: circRNA, lncRNA, ncRNA, piRNA, spermatozoa, spermatozoal RNA, spRNA, transcriptome

INTRODUCTION

Spermatogenic defects and sperm abnormalities are responsible for high incidence of male infertility cases in both animals and human. The diagnosis and treatment of spermatogenic failure remain to be a thrilling challenge to veterinarians and medical practitioners despite significant research progress in the investigation and treatment of infertility. Male infertility in livestock is commonly evaluated from semen quality assessment parameters such as sperm concentration, forward progressive motility, morphological defects, acrosomal abnormalities, hypo-osmotic swelling test (HOST) for membrane integrity, etc. Numerous functional assays have also been developed to evaluate the competence of spermatozoa within the female genital tract and include assessment of *in vitro* capacitation, acrosomal reaction and hypermotility, cervical mucus penetration test, zona-free hamster egg penetration assay, and *in vitro* fertilization (IVF) (Saacke et al., 2000; Foxcroft et al., 2008). Assessment of DNA fragmentation in the sperm nucleus has been a relatively recent addition to the semen evaluation parameters in farm animals (Kumaresan et al., 2020). However, these procedures fail to diagnose spermatogenic failure and, several cases remain undiagnosed and declared as idiopathic. Heterogeneous sperm population within a single

ejaculate, seasonal variation, influence of semen freezing protocol, etc., further adds to the variability of results (Yang et al., 2010; Varona et al., 2019). Thus, there is a need to develop newer molecular and biochemical methods for rapid and reliable diagnosis of spermatogenic failure and male infertility.

Spermatozoa contain a suite of RNA species, including both coding RNAs and non-coding RNAs such as microRNAs (miRNAs), intranuclear RNAs, small nucleolar RNAs (snRNAs), etc. The concentration of spermatozoal RNA (spRNA) may be as much as 0.015 pg per sperm (in humans), which is conspicuously high considering the high nuclear:cytoplasmic ratio and very low volume of cytoplasm in sperm (Miller et al., 2005). Microarray analysis has revealed that more than 3,500 unique mRNAs were present in ejaculated human spermatozoa (Ostermeier et al., 2002). However, these spRNAs have traditionally been considered a residual of the spermatogenesis process due to transcriptionally dormant nuclear genome (Hecht, 1998) and lack of 28S rRNA and 18S rRNA in the cytoplasm of mammalian spermatozoa (Miller et al., 1999). The lack of rRNAs eliminates the possibility of translation within sperm and hence, *de novo* protein synthesis (Miller et al., 1999).

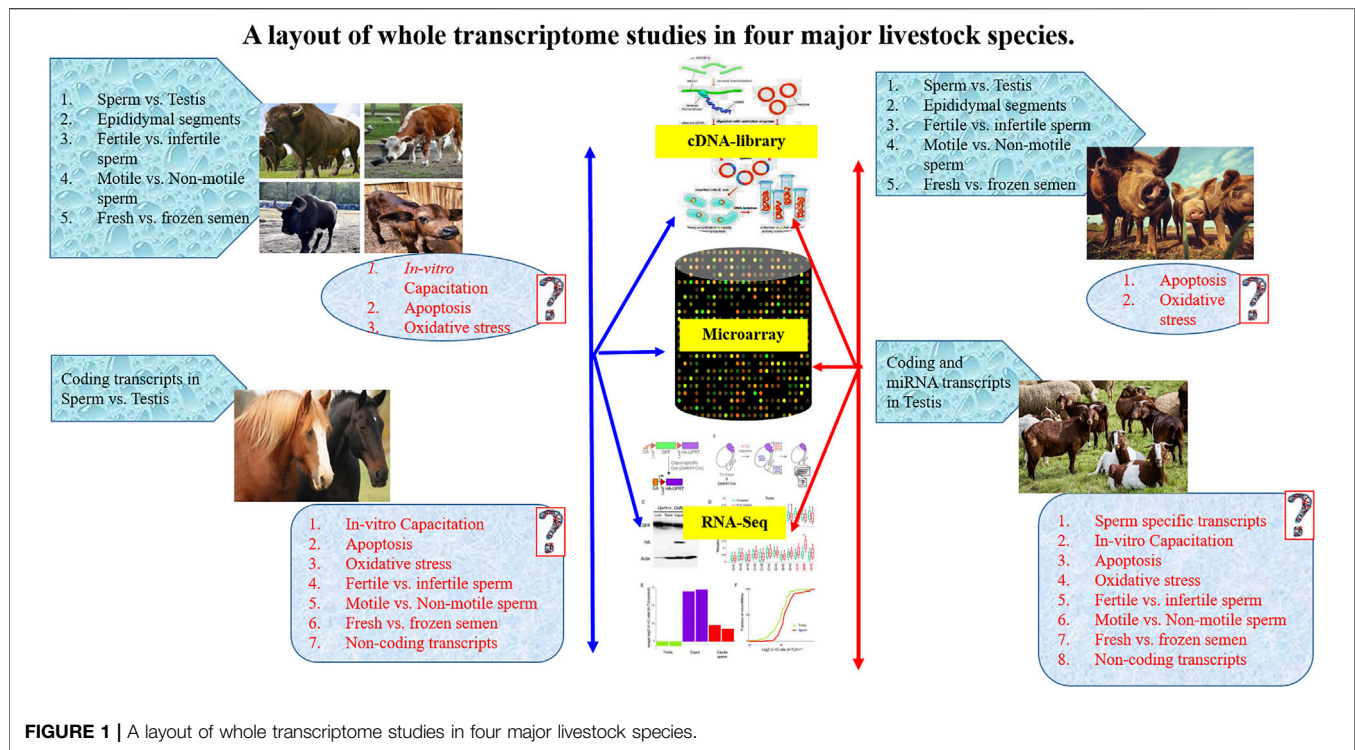
More recent studies have shown that, during fertilization, at least six sperm-specific, developmentally related mRNAs are delivered to the zygote (Ostermeier et al., 2004). Studies have further shown that protein-coding spRNAs undergo spatio-temporally regulated degradation during early embryonic development (Ziyyat, 2001; Hayashi, 2003; Ostermeier et al., 2004). Thus, spRNAs may have specific functions during early embryonic development, although their exact role or mechanism of action remains elusive. (Wykes et al., 1997) observed that spRNA were localized at the periphery of the sperm nucleus and co-localized with spermatozoal histone. Since imprinted genes such as *IGF2* are bound with histone protein in the sperm nucleus, it has been suggested that spRNAs may have a role in mediating imprinting mechanism or chromatin repackaging following fertilization (Miller et al., 2005). Microarray profiling of spRNAs in infertile men revealed a distinct RNA profile that could be used as a molecular signature of infertile man's sperm (Ostermeier et al., 2002) (Wang et al., 2004). Thus, knowing the identity, functions, and regulation of spRNAs may have direct relevance to the diagnosis of male infertility and developing RNA-based therapies or contraceptives. The aberrance or absence of spRNA may be a contributing factor for idiopathic infertility and may bear on the poor performance of livestock. However, while many studies have analyzed the spRNA population in human spermatozoa, there are extremely limited results from livestock species. This review article provides a detailed review of known spRNA populations in various livestock species, their possible role in male fertility and early embryonic development, and their significance in the diagnosis and/or treatment of idiopathic infertility. A brief on RNA populations in seminal plasma is also discussed. Owing to very limited data on spRNAs in livestock, suitable relevance is also taken from humans.

Spermatogenesis and Spermatozoa

Spermatogenesis is a complex set of events initiated at the germinal epithelium of seminiferous tubules. The spermatogonial stem cells (SSCs), the stem cells at the basement of germinal epithelium, are responsible for the maintenance of spermatogenesis throughout the adulthood of males. They possess the ability to self-renew themselves and differentiate into haploid spermatozoa through the well-orchestrated spermatogenesis process. Upon molecular cues to differentiate, the SSCs initially divide mitotically to produce spermatogonia and subsequently undergo a meiotic process to produce spermatocytes, followed by the production of haploid spermatids (Gomes et al., 2013). The spermatids undergo sequential morphological changes by spermiogenesis to develop into haploid spermatozoa. The spermatozoa in seminiferous tubules are immature and undergo a maturation process in the epididymis. The epididymal maturation of spermatozoa is characterized by several physiological and molecular changes in the plasma membrane (e.g., lipid composition, surface proteins, etc.) and nucleus (e.g., DNA condensation, protamine formation, chromatin rearrangement, etc.) (Légaré et al., 2017), formation of acrosome and development of flagella, etc. before spermiation (Griswold, 2016; Neto et al., 2016). Due to cytoplasmic expulsion, mature spermatozoa are produced with very little-to-no cytoplasm. During sperm maturation, sperm transition proteins (TNP1 and TNP2) replaces the majority of the histone proteins of the spermatids, followed by the deposition of highly basic protamines (PRM1 and PRM2) in elongated spermatids and spermatozoa (Jambor et al., 2017). The complex chromatin packaging makes the sperm epigenome highly stable and makes it transcriptionally inactive (Lambard et al., 2004). However, transcriptional and translational activities have been observed during the early stages of spermiogenesis (Miller, 2007; Yang et al., 2010) until the development of round spermatids. Translation of specific mRNAs also continues for several days after discontinuing the transcription process and completing spermiogenesis. The latter is believed to occur from those spRNAs that are associated with leftover histone protein in the nucleus and are potentiated for transcription (Miller, 2007).

The spRNAs Population in Spermatozoa

In last 1 decade, a number of studies have prepared cDNA libraries and performed high-throughput RNA sequencing (RNA-seq) or microarray analysis of spermatozoal samples from epididymis and ejaculates. The spermatozoal samples have been compared between fertile and infertile males for spRNA profiling (**Figure 1**) and reported to have several coding and non-coding mRNA transcripts (Li and Zhou, 2012). The concentration of spRNA varied from picograms to femtograms per cell ranging from 100 fg in mice (Miller, 2014), 10–20 fg in humans (Zhang et al., 2017), two fg in bull (Sellem et al., 2020), 5 fg in swine (Hamatani, 2012), to 2–20 fg in stallions (Das et al., 2010). However, earlier reports on spRNAs have suggested their inertness for transcriptional and translational activities in spermatozoa (Zhang et al., 2017). The mature spermatozoa contain insufficient 28S or 18S rRNAs to support



translation (Miller et al., 1999). The absence of essential components of translational machinery raised scepticism on *de novo* translation or any possible alternative roles of spRNAs in spermatozoa beyond the delivery of a paternal genome. Thus, the presence of spRNA not only aroused controversies but is intriguing to date from the discovery of sperm-borne RNAs in zygotes (Krawetz et al., 2011). The presence of RNAs in the male gamete is traditionally assumed to be either degraded leftover or spermatogenic expulsion of the residual body (Zhang et al., 2017). It was believed that *de novo* gene expression could not occur in mature spermatozoa due to highly compacted DNA by protamine that substitutes histone proteins during spermiogenesis (Balhorn, 2007). Conversely, spermatozoa rely upon pre-formed spRNAs and proteins for chromatin re-compaction, completion of the spermatogenesis process, and subsequent fertilization events (Miller, 2014).

Although sperm RNA population consists of various classes of coding RNAs, non-coding RNAs (miRNAs, piRNAs, siRNAs, lncRNAs), mitochondrial RNAs, ribosomal RNAs (rRNAs), and some intronic retained elements (Miller, 2007), the accuracy in determining its quantity and functional significance has been challenging (Dadoune, 2009). It was suggested that the variable amount of RNA in spermatozoa in different species might be due to differences in RNA isolation protocol or contamination of somatic cells in the semen. Thus, an optimized protocol for sperm purification and RNA extraction is required before any conclusion being drawn on their relevancy to male fertility (Gòdia et al., 2018). Nevertheless, evidence of both coding and non-coding spRNAs were proclaimed in different livestock species, including cattle, pigs, and stallion (Kempisty et al., 2008). Notably, full-length mRNA transcripts have also been

reported (Sun et al., 2021) that had the potential to be translated *de novo* under certain circumstances (Miller and Ostermeier, 2006).

The spRNA was also suggested to have the potential to modulate phenotype through epigenetic alternations in gene expression (Pantano et al., 2015) and imprinting of IGF2 expression (Boerke et al., 2007). It was shown that spRNAs were located at the periphery of the nucleus at the boundary of histone-bound and protamine-bound DNA in the sperm. Thus, they may have an association with potentiated DNAs for gene expression (Jodar et al., 2013). Further, the sperm transcriptome profiling using Microarray (Das et al., 2013; Zhang Y. et al., 2017), and RNA-seq (Gòdia et al., 2018) have identified a suite of RNAs in spermatozoa (Sandler et al., 2013) that included both coding and non-coding transcripts and were associated with regulation of various biological functions such as chromatin repackaging, genomic imprinting, early embryonic development (Das et al., 2013), and post-fertilization events (Prakash et al., 2021). The spRNA profile also differed between low-motile and high-motile spermatozoa (Lambard et al., 2004) and between fertile and infertile men (Wang et al., 2004; Ostermeier et al., 2005). Yatsenko and co-workers detected abnormal *UBE2B* (Yatsenko et al., 2013), *ZBP1* (Yatsenko et al., 2012) and *KLHL10* (Yatsenko et al., 2006) genes in spRNA and associated it with impaired fertility. Wu et al. (Wu et al., 2012) reported that miR-19b and miR-let7a could serve as specific and sensitive biomarkers for spermatogenic status in idiopathic infertile males with oligozoospermia and non-obstructive azoospermia. Altered miRNAs expression has also been found in the testis with non-obstructive azoospermia (Lian et al., 2009). Thus, the expression level of spRNAs and

TABLE 1 | Important protein coding transcripts reported in cattle spermatozoa by transcriptome analysis.

Species	Study type	Groups	Transcripts (Gene symbol)	Function	References
Cattle	RNA-Seq	Whole transcriptome	<i>ZBTB20</i>	Spermatogenesis	Raval et al. (2019)
	RNA-Seq	Whole transcriptome <i>KIF5C</i> and <i>KCNJ6</i>	<i>YWHAZ</i>	Spermatogenesis and acrosomal reactions	Selvaraju et al. (2017)
			Microtubule function		
			<i>TNP1</i> , <i>RAD21</i> , <i>UBE2B</i> and <i>RAN</i>	Spermatid development	
			<i>TEKT1</i>	Acrosome and on flagella and could play a key role during fertilization	
			<i>CAPZA3</i>	Spermatozoa capacitation and spermatozoa egg fusion	
			<i>MAP7</i> , <i>PTK2</i> , <i>PLK1S1</i> , <i>MYH9</i> and <i>PRKCZ</i>	Spermatogenesis and sperm function	
			<i>PLCZ1</i> and <i>PLCB1</i>	Calcium signalling and the spermatozoon-induced activation of an oocyte leading to its final maturation	
			<i>ADAM1B</i> , <i>ADAM2</i> and <i>ADAM32</i>	Spermatozoal motility	
			<i>ZP3</i> and <i>FOXG1</i>	Embryogenesis-associated transcript	
			<i>PAG5</i> , <i>PAG7</i> and <i>PAG10</i>	Implantation and placentation	
	RNA-Seq	High vs Low fertile	<i>PFN1</i>	Oocyte maturation, fertilization, embryo development, spermatogenesis	Prakash et al. (2021)
			<i>PRKA 1</i>	Acrosome formation	
			<i>ODF1</i> , <i>BCL2L11</i> , <i>PRM2</i> , <i>TNP2</i> , <i>ODF2</i> , <i>SPEM1</i> , and <i>MEA1</i>	Spermatogenesis	
			<i>YBX1</i> , <i>UBE2B</i> , <i>BCL2L11</i> , <i>MYH10</i> , and <i>RBBP6</i>	<i>In-vitro</i> embryonic development	

seminal RNAs may have potential use in studying the post-spermatogenesis events and detecting defects in sperm function (Prakash et al., 2021). A layout of whole transcriptome studies in major livestock species is presented in **Figure 1**.

Coding RNAs in Spermatozoa of Livestock

The spRNA population has been described in various farm animals such as cattle, horses, and pigs (Card et al., 2013; Das et al., 2013). The coding RNAs were limited mainly to mRNA transcripts, whereas ncRNAs belonged to different types such as miRNAs, lncRNA, rRNA, tRNA, tsRNA, circRNA, etc. The presence of fragmented rRNAs (Cappallo-Obermann et al., 2011), 5,000–6,000 mRNAs, and other classes of RNAs such as tRNA, small RNA, and miRNA have also been described (Parthipan et al., 2017). Among protein-coding RNAs, Hydrolases, SP-40, Sulfated glycoprotein 2, Calmegin, Heat shock proteins (HSPs) were highly abundant. At the same time, non-coding RNA (ncRNA) mainly included miRNAs and siRNAs whose targets were signaling pathways such as the Wnt signaling pathway that is known to be regulators of early embryonic development and differentiation (Peifer, 2000).

Coding spRNAs in Cattle

The whole transcriptome profiling of bulls' spermatozoa revealed a wide range of potential transcripts (**Table 1**) that regulate DNA packaging, cytoskeletal organization, acrosomal reactions (e.g., *PLCB1*, *YWHAZ*) (Selvaraju et al., 2017), oocyte activation, embryogenesis, placental development (e.g., *PAG5*, *PAG7*, and *PAG10*), embryonic morphogenesis, microtubule

function (e.g., *KIF5C* and *KCNJ6*), mitochondrial function (e.g., *COX5A* and *COX11*), calcium signaling (e.g., *PLCZ1* and *PLCB1*), and centrosome organization (e.g., *MAP7*, *MYH9*, *PLK1S1*, *PRKCZ*, and *PTK2*). Bovine cDNA microarray analysis of spermatozoa from different segments of epididymis further revealed segmental differences in the transcriptome. Among the differentially expressed genes (DEGs), the top 10 DEGs were related to reproductive function (*ADAM28*, *AKAP4*, *CTCFL*, *FAM161A*, *ODF1*, *SMCP*, *SORD*, *SPATA3*, *SPATA18*, and *TCP11*) and five DEGs (*DEAD*, *CYST11*, *DEFB119*, *DEFB124*, and *MX1*) were related to the immune response and cellular defense (Légaré et al., 2017). Notably, most up-regulated transcripts of the caput epididymis are known to regulate spermatogenesis and sperm morphology and included *AKAP4*, *ODF1*, *CTCFL*, *SMCP*, *SPATA3*, *SPZ1*, and *SPATA18*. These genes in spermatozoa have been associated with sub-fertility of bulls with incomplete spermiogenesis (Hermo et al., 2010). The sperm transcripts related to sperm maturation were *ADAM28*, *CRISP2*, *CST11*, *LCN9S*, and *TCP11*. On the other hand, dysregulated immune defense-related genes were *CATGL4* and *GSTA2* in the caput epididymis; *GPX5*, *MX1*, and *DEFB124* in the corpus epididymis and; *DEFB7* and *DEFB119* in the cauda epididymis (Selvaraju et al., 2017).

Various intact rRNA (e.g., *RPL23*, *RPL27A*, and *RPS18*) and degraded rRNAs (*RPL6*, *RPL36AL*, and *RPL37*) have also been reported in bull spermatozoa (Montjean et al., 2012), which might suggest their potential role in spermatogenesis. The comparative RNA profiles of semen from high- and low-fertility bulls using microarray revealed 415 DEGs (out of 24,000 analyzed genes) with a significant number of fertility-

associated markers (Feugang et al., 2010). The study showed higher expression of membrane and extracellular matrix genes in spermatozoa from high-fertility bulls, while transcripts of transcriptional and translational factors were lower in the low-fertility bulls. Increased expression of *CSN2* and *PRM1* in spermatozoa of high-fertility bull and lower expression of *CD36* in low-fertility bull were suggested as possible fertility markers (Feugang et al., 2010).

The role of coding spRNA transcripts was also indicated in the motility of sperm flagellum (*AKAP*) and DNA packaging (*PRM2*) (Singh et al., 2019). Differences in spRNA transcripts among high- and low-fertility dairy bulls have been reported with 805 unique transcripts in high-fertility and 2,944 unique transcripts in the low-fertility bulls (Card et al., 2017). Expression of cytochrome oxidase subunit (*COX7C*) was negatively correlates with male fertility (Card et al., 2017). In another study, a combination of five genes (*AK1*, *ITGB5*, *TIMP*, *SNRPN2*, and *PLCz1*) accounted for 97.4% of the variation of conception rate from frozen-thawed Holstein bulls, which was indicative of the sire fertility index (Kasimanickam et al., 2012). The upregulated sperm transcripts (e.g., *RPL3*, *PABPC1*, *TPT1*, *RPL14*, *RPS8*, *PFN1*, *DDX39B*, and *CD74*) in low-fertile crossbred bulls annotated to biological functions like apoptosis, cellular differentiation, apoptosis of germ cells, spermatogenesis, fertilization, and early embryo development (Arcuri et al., 2004; Rawe et al., 2006; Selvaraju et al., 2018) whereas downregulated sperm transcripts (e.g., *RUNDC3A*, *LYRM4*, *FAM71F1*, *ZFN706*, *PICK1*, *LUZP1*, *ANKRD9*, and *EPOP*) were found to be modulating the cytoskeletal organization and acrosome formation (Selvaraju et al., 2017). The *TSSK6*, *C12H13orf46*, *FABP3*, and *IQCF1* genes were the top spRNA transcripts that were unique to high-fertility bull spermatozoa and were associated with biological functions of protein phosphorylation, sperm chromatin condensation during spermiogenesis, sperm motility, acrosome reaction, and gamete fusion during fertilization (Bissonnette et al., 2009; Sosnik et al., 2009; Fang et al., 2015; Selvaraju et al., 2018). The upregulated sperm transcripts that were unique to the low-fertility bulls included ribosomal proteins and thymosin beta 10 (*TMSB10*), which are involved in sperm capacitation, fertilization, and cellular remodeling during trophoblast adhesion (Cammass et al., 2005; Selvaraju et al., 2017). These studies suggest an association between coding spRNAs in bulls sperm and their fertility.

Coding spRNAs in Pigs

The significance of spRNA functions has also been recognized in pigs (Table 2). Using RNA-Seq, several genes linked to spermatogenesis (e.g., *FGF-14* and *BAMBI*), energy metabolism (e.g., *ND6* and *ACADM*), protein phosphorylation (e.g., *PTPRU* and *PTPN2*), autophagy (e.g., *RAB33B*), inflammation, and apoptosis (e.g., *EAF2*, *FOS*, *ITGAL*, *NFATC3*, and *ZDHHC14*) have been reported to be significantly up-regulated in poor freezable ejaculates of boar semen (Fraser and Adeoya-Osiguwa, 2001). Thus, the spRNA population may serve as a molecular signature or marker of

sperm freezability. Comparison of spRNAs in fresh vs. frozen-thawed boar sperm resulted differential expression of 567 protein coding mRNA and 135 non-coding miRNA (Card et al., 2017; Dai et al., 2019). The KEGG pathway analysis of DEGs revealed several signaling pathways governing the spermatogenesis process, such as chemokines signaling, PI3K-AKT, cGMP-PKG signaling, JAK-STAT signaling, calcium signaling, TNF signaling, MAPK, calcium signaling, NF-kappa B signaling, and AMPK signaling pathways. Similarly, microarray analysis of spRNA population in differentially fertile pigs revealed significant involvement of major signaling pathways associated with JAK2 and STAT3 pathways, cytokine receptor activity, activation of B- and T-lymphocytes (Yang et al., 2009). High fertile groups were also marked with downregulation of cell adhesion and proliferation genes (e.g., *CDH10*, *CDSN*, *ITGB8*, *ANGPTL1*, and *CTNNA3*) and upregulation of genes associated with the cellular component organization (e.g., *KRI1* and *ZNHIT6*), endocytic receptor activity (e.g., *CXCL16*), membrane channels (e.g., *KCNA3*, *KCNIP3*, *KCNH4*, and *KCTD9*) and translation regulator activity (e.g., *CPEB3*). Transcripts regulating sperm motility during *in vitro* capacitation, such as *CATSPERG* (CatSper channel auxiliary subunit gamma) was upregulated, and *CATSPERB* (CatSper channel auxiliary subunit beta) was down-regulated in high-fertility pigs. The DEGs annotating to zinc finger nucleases (ZFNs) such as *LOC100739821*, *ZNRF4*, *PLAGL2*, *ZFN25*, and *ZDHHC7* were primarily upregulated in high-fertility boars (Alvarez-Rodriguez et al., 2020). Some of the ZFN transcripts (e.g., *ZFN283*, *FEZF2*, and *GLI1*) were downregulated in high fertile groups and were involved with sonic hedgehog signaling pathway (Kroft et al., 2001). Apart from this, transcripts of *NFYA*, *TCF21*, *MBTPS2*, *MTF2*, *IRF3*, and *ISG20L2* were significantly overexpressed, and *UBTFL1*, *ADAM7*, *ADAM29*, *IL23R*, *IFN-DELTA-4*, *IFNK*, and *IFIT3* were under-expressed in high fertile boars (Turner et al., 2006; Wei et al., 2011).

The spRNAs enriched in cell-cell adhesion, endometrial epithelial cell receptivity, lipid and glucose metabolism, inflammation, autophagy, matrix metalloproteases, mitochondrial apoptosis, and immune-related signaling pathways have also been reported (Alvarez-Rodriguez et al., 2020). The EST library of ejaculated spermatozoa from Landrace pig deciphered some of the putative homologs of transcripts regulating embryogenesis (e.g., *HSP70.2*, *SSFA2*, and *SESN1*). These transcripts were also marked as potential regulators of spermiogenesis, oocyte fertilization, cellular growth, and cleavage (Yang et al., 2009). Seasonal variation in RNA-seq profiling of spRNAs was also reported in pigs (Yang et al., 2010) and included transcripts having functional similarity with those previously reported in cattle (Selvaraju et al., 2017). The mRNA transcripts regulating spermatogenesis (e.g., *ODF2* and *SPATA18*), nuclear genome structure (e.g., *PRM1*, *OAZ3*, *HSPB9*, and *NDUFS4*), mitochondrial function (e.g., *COX1ATP8*), and fertilization (e.g., *HSPA1L* and *PRSS37*) were observed to be associated with compaction of sperm chromatin, energy metabolism, oxidative stress, apoptosis, and early embryo

TABLE 2 | Important protein coding transcripts reported in pig spermatozoa by transcriptome analysis.

Species	Study type	Groups	Transcripts (Gene symbol)	Function	References
Pigs	cDNA-library	Ejaculated spermatozoa	<i>Hsp70.2</i> , <i>SSFA2</i> , <i>SESN1</i>	Embryogenesis	Yang et al. (2009)
	RNA-Seq	Capacitated sperm	<i>MAPK1</i> , <i>PGK1</i> , <i>PPM1B</i> , and <i>PGAM1</i>	Capacitation	Li et al. (2018)
	RNA-Seq	Fresh vs frozen semen	<i>VEGFA</i>	Self-renewal and maintenance of male spermatogonial stem cells, activate the AKT signalling, improve sperm motility	Dai et al. (2019)
			<i>DNMT3A</i> , <i>DNMT3B</i> , <i>JHDM2A</i> , <i>KAT8</i> , and <i>PRM1</i>	Epigenetic regulation	
			<i>PLCZ1</i> and <i>CRISP2</i>	Calcium ion pump, sperm-egg interaction, regulation of sperm intake, and fertilization process	
			<i>MTHFR</i> , <i>PAX8</i> , <i>IGF2</i> , <i>LIT1</i> , and <i>SNRPN</i>	Epigenetic regulation	
	Microarray	High fertile vs Low fertile	<i>AKAPs</i> <i>CATSPER</i> Zinc finger proteins	Capacitation, motility Mammalian fertilization by Calcium signalling Transcriptional regulation, ubiquitin-mediated protein degradation, signal transduction, actin targeting, DNA repair and cell migration	Alvarez-Rodriguez et al. (2020)
			Matrix metallo-proteases	Activation of TNF-alfa and generation of Epidermal Growth Factor Receptor	

development (Sendler et al., 2013; Sakurai et al., 2016; Selvaraju et al., 2017). It was also suggested that the *TSARG1* transcript in spermatozoa might be involved in inhibiting apoptosis (Yang et al., 2005). The study reported that the expression level of *TSARG1* and testis-specific kinase 1 (*TESK1*), which are members of the DnaJ-like protein family and serine/threonine kinases, respectively, were significantly higher in the sperm pools collected in winter than in summer. The transcriptome analysis of pig testis and epididymis revealed several transcripts associated with different regions of the epididymis (Guyonnet et al., 2009). Several lipocalins were observed in different epididymal segments such as *LCN6*, *LCN8*, *LCN9*, *LCN10*, *PTGDS*, but their role is not yet evident in response to epididymal functionalities.

Coding spRNAs in Stallion

In order to establish the significance of spRNAs as potential biomarkers of stallion fertility, global transcriptome profiling of semen has been explored (Das et al., 2013; Suliman et al., 2018). The abundance of spermatozoal mRNAs (e.g., *ADIPO*, *AK1*, *CRISP2*, *DOPPEL*, *ITGB5*, *NGF*, *PEBP1*, *PLCZ1*, and *TIMP2*) were found to be positively correlated with fertility, while mRNA transcripts of *CCT8* and *PRM2* were found to be negatively correlated (Kasimanickam et al., 2012). The stallion sperm transcripts were predicted to regulate critical biological functions, viz., the integrity of plasma membrane, RNA processing, transcription regulation, mitochondrial ribosomal protein, ion binding, chemokine receptor DNA packaging, protein folding, cytoskeleton, GTPase activator, chromatin assembly complex, and protein transport (Suliman et al., 2018). The cysteine-rich secretory protein (CRISP), found in seminal plasma proteins, was involved in gamete fusion (Töpfer-Petersen et al., 2005). Some ribosomal binding proteins (e.g., *GRTH*, *SAM68*, *MSY2*, and *DAZAP1*) were also observed that are known to promote the translation of

mRNAs in germ cells. The latter may thus, suggest the importance of spRNAs in regulating protein translation during spermatogenesis (Bettegowda and Wilkinson, 2010). Comparative transcriptomics of stallion sperm identified 149 DE miRNAs between dense and less-dense spermatozoa (Ing et al., 2020) wherein *BCO1*, *GLRA4*, *OTOL1*, *PRM1*, *SCP2D1*, and *SPATA31D1* were highly expressed in dense spermatozoa compared to those of less-dense spermatozoa. This study also found that expression of *PRM1* transcripts was significantly higher in morphologically normal spermatozoa from sub-fertile stallions and in spermatozoa with abnormal morphology (Paradowska-Dogan et al., 2014). The association of protamines expression levels in frozen-thawed semen had implications in stallion fertilization (Kadivar et al., 2020).

Non-Coding RNAs in Spermatozoa of Livestock

Two groups of ncRNAs are the short and long ncRNAs (Watson et al., 2019). Transcripts <200 nucleotides are the small ncRNA (sncRNA) that include piwi-interacting RNA (piRNAs), small interfering RNA (siRNAs), miRNA, rRNA, tRNA, snoRNAs, and small nuclear RNA (snRNAs) (Grivna et al., 2006). Transcripts >200 nucleotides are generally categorized as long ncRNA (lncRNA) (Fort et al., 2021). Deep sequencing of cattle spRNA revealed that sncRNA population comprises the most abundant rRNA followed by piRNAs, miRNAs, and tRNA fragments (tsRNA), contradictory to sperm-based data where tRNAs represented most of the reads (Sellem et al., 2020). Microarray analysis of spRNAs in humans revealed that spRNA contains an array of ncRNAs that included miRNAs and siRNAs. Interestingly, 68 of these siRNAs had protein targets which were regulators of development and differentiation in *C. elegans* (Martins and Krawetz, 2005). The function of these

sperm-derived miRNA was further demonstrated by (Amanai et al., 2006). Using anti-miRNA microinjection to oocytes, (Amanai et al., 2006), showed that sprNA has a limited function during early embryonic development in pigs. Nevertheless, direct proof of the function of these RNAs is still elusive. The presence of ncRNAs in livestock has been associated with spermatogenesis, fertilization, early embryogenesis (Prakash et al., 2020), and epigenetic modification of the sperm genome (Jodar et al., 2013; Prakash et al., 2020).

Spermatozoal miRNA

The miRNAs are short single-stranded RNAs of ~18–22 nucleotide length that are found abundantly in mammals and are known to post-transcriptionally regulate the function of mRNAs through inhibition or suppression of translation or degradation of mRNA themselves. The significance of miRNAs in male reproduction has been documented in the maintenance of self-renewal and differentiation of SSCs into spermatozoa as well as during the spermiogenesis process (Chen X. et al., 2017). We previously showed that the let-7 family of miRNA is involved in regulating germ cell proliferation and may be used as a marker of male germ cells (Jung et al., 2010) in addition to imprinted miRNAs (Shin et al., 2011). Similarly, miR-21, miR-34c, miR-183, miR-465a, etc., were also found to be vital for SSC self-renewal through regulation of *Etv5*, *Cdnd1*, and *Stat3* expression in mice (Niu et al., 2011; He et al., 2013). On the other hand, miR-224 and miR-322 regulate the Wnt signaling pathway for self-renewal of mice SSCs by increasing the expression of *Gfra1*, *Plzf*, and *Rassf8* (Cui et al., 2016; Wang Y. et al., 2019) whereas miR-100 and miR-10b promoted the proliferation of SSCs through *Stat3* and *Klf4*, respectively (Huang et al., 2017; Li et al., 2017). The miRNAs were also shown to play crucial roles in the differentiation of mice SSCs and spermatogenesis by targeting the *Dmrt1* (Cui et al., 2016) and through the retinoic acid pathway (Bellvé et al., 1977).

The complexity of the miRNA population has also been seen in spermatozoa of several mammalian species, including cattle, pigs, stallion, mice, and humans (Capra et al., 2017; Castillo et al., 2018; Gódia et al., 2018). However, the precise functions of these sperm miRNAs in spermatogenesis or post-fertilization events are yet to be established. Spermatozoal miRNAs are believed to play essential roles during fertilization of oocytes and early embryonic development (Castillo et al., 2018). They may also affect the abundance and epigenetic status of maternal mRNAs in oocytes and zygotes (Capra et al., 2017).

Spermatozoal miRNAs in Cattle

Spermatozoal miRNAs have been reported in healthy bulls and those differing in their spermatozoal motility or fertility (Table 3). The whole miRNA profiling of bull spermatozoa through deep sequencing identified 2022 miRNAs and included a large number of isomiRs across different species (Turri et al., 2021). RNA-seq analysis of high fertility and low fertility bulls reported nine DE miRNAs (miR-2285n, miR-378, miR-423-3p, miR-19, miR-2904, miR-378c, miR-431, miR-486, and miR-2478) in which miR-2285n, miR-378, and miR-486 were observed to have altered expression levels between low and high

motile sperm. The highest expression of miR-2285n and miR-486 was seen in low motile spermatozoa and the semen samples from low fertility bulls. The miR-486 plays a crucial role in regulating stemness and cell proliferation of SSCs. The miR-378 targets mRNAs involved in metabolism and sperm motility (Liu et al., 2019). Differential expression of spermatozoal miRNAs was also observed upon microarray analysis of spermatozoal miRNA in low vs. high motility spermatozoal and low vs. high fertility bull (Sellem et al., 2020; Keles et al., 2021).

The miRNA profiling of bull sperm using microarray presented an abundant quantity of miRNA along with DE miRNA target genes from high vs. low fertility groups (Dai et al., 2019) suggests the roles of miRNA in regulatory mechanisms. Seven important miRNAs (hsa-aga-3155, 8197, 6,125, 6,727, 11,796, 13,659 and 14,189) were DE (using human microarray probe set) and validated by qRT-PCR. The target genes of these miRNAs included *AQP7P1*, *CHN2-EPHA1-EFNA2*, *DNM2*, *IFT80*, *TOB2*, *CHN2*, *CLUL1*, *BC035897*, *BTBD2*, possibly regulating gametogenesis, acrosome integrity, and fertilization. An interactome model of miRNA target genes has also been proposed to identify the master regulator of a set of genes with a single miRNA molecule (Govindaraju et al., 2012). The miR-26a and miR-455-5p were significantly up-regulated in highly motile spermatozoa, whereas levels of miR-10a and miR-1 were significantly down-regulated. The miR-26a also participates in PTEN and PI3K/AKT signaling pathways, affecting sperm viability and motility (Dai et al., 2019). In another study, miR-34b-3p and miR-100-5p were also significantly over-expressed in spermatozoa of high fertility bull compared to those of low fertility bull (Keles et al., 2021). The miRNAs, previously known to be expressed testis, were also detected in bull spermatozoa and included miR-34b/c miRNA cluster, which plays vital roles in chromatin condensation during spermiogenesis (Capra et al., 2017). Thus, miRNA profile could be a useful indicator for high fertility spermatozoa.

The miRNA profiling has also found its application in assessing the success of semen cryopreservation protocols. Increased expression of miR-34c was seen in the highly motile cryopreserved bull spermatozoa purified through Percoll density gradient centrifugation. The miRNA profiling of differentially motile cryopreserved spermatozoa revealed that spermatozoal miRNAs target STAT3, PI3K/AKT, and PTEN signaling pathways that play essential roles during mammalian spermatogenesis, mitochondrial membrane potential, sperm maturation, and fertilization. PTEN, the target of multiple miRNAs (e.g., miR-17-5p, miR-26a-5p, and miR-486-5p), inhibits AKT signaling and activates RAF1/ERK signaling to initiate sperm maturation. Two miRNA viz. miR-122 and miR-184 were upregulated in low motile fractions of semen, which targets the AKT signaling pathway to cause apoptosis. In addition, miR-17-5p and miR-20a-5p, which were found to be under-expressed in the low motile fraction of semen, targets PTEN and STAT signaling pathways to trigger apoptosis (Capra et al., 2017).

The spermatozoal miRNA have also been shown to differ in gonadotoxicity, heat stress, and impaired spermatogenesis. Differentially expressed miRNAs have been reported between

TABLE 3 | Important non-coding microRNAs reported in cattle spermatozoa by transcriptome analysis.

Species	Study type	Groups	MicroRNAs	Function	References
Bovine	cDNA library	High fertile vs Low fertile	miR-34b/c miR-17-92 and miR-106b-25 miR-10a-5p	Sperm chromatin condensation Spermatogenesis Motility	Lian et al. (2021)
	MicroRNA microarray RNA-Seq	High and low fertile Semen sample	miR-3155, miR-8197, miR-6727, miR-11796, miR-14189, miR-6125, and miR-13659 miR-365-2	Sperm fertilization Oocyte fertilization and cleavage	Govindaraju et al. (2012) Selvaraju et al. (2017)
	RNA-Seq	Small RNA sequencing	bta-miR-103, bta-miR-30b-5p, bta-miR-17-5p, bta-miR-106b, bta-miR-142-3p, bta-miR-34b, bta-miR-18a, bta-miR-34c, bta-miR-455-5p, bta-miR-10b, bta-miR-99b, bta-miR-1246, bta-miR-99a-5p, bta-miR-1388-5p	Motility	Capra et al. (2017)

normal and abnormal spermatozoa of bulls affected by Fescue toxicosis (Stowe et al., 2014). An increased expression of let-7a and miR-22 was shown in spermatozoa with abnormal morphology due to toxicosis. Further, among the most abundant miRNAs, the let-7 family and miRNA-146 were found to regulate the differentiation of spermatogonia and impaired blastocyst implantation in mice whereas, miRNA-16 is known to coordinate cell-cycle (Zhang et al., 2012). A recent study suggests miRNA (e.g., miR-126-5p) accumulates during the final stage of spermatogenesis, and epididymal transit and their expression is affected during heat stress, which might explain reduced fertility in heat-stressed bulls due to impaired spermatogenesis and maturation (Dos Santos da Silva et al., 2021). The miR-17-92 and miR-106b-25 are two miRNA clusters whose target genes were reported to be essential for the progression of spermatogenesis (Tong et al., 2012). It was reported that spermatozoa that exhibited normal morphology had mutant miR-92a, implying that miR-92a might affect male fertility. The miR-10a and miR-9-5p have been frequently reported in cattle and pig spermatozoa. Both of these miRNAs were observed to be upregulated in cryopreserved spermatozoa. On the other hand, miR-10a-5p was found to be overexpressed in low motility spermatozoa and was associated with spermatogenesis and DNA repair capacity.

Spermatozoal miRNAs in Pig

The miRNAome of boar spermatozoa has been described in several studies (Table 4) and may vary with season (Varona et al., 2019). The miR-34c, miR-191, miR-30d, miR-10b, and let-7a were the most abundant spermatozoal miRNAs associated with spermatogenesis in pigs. The differential expression of several miRNAs is seen in low vs. high fertility boars, low vs. high motility, and fresh vs. cryopreserved spermatozoa. The small RNA libraries of fresh and frozen-thawed boar spermatozoa revealed several potential cell signaling pathways governed by spermatozoal miRNAs (Dai et al., 2019). For example, miR-17-5p, miR-20a-5p, miR-26a-5p, miR-122-5p, miR-184, and miR-486-5p were involved in regulating PTEN, PI3K/AKT, and STAT signaling pathways that influence sperm motility, viability, and apoptosis in frozen-thawed spermatozoa (Zhang Y. et al., 2017). The increased expression of miRNAs such as let-7a, 7d, 7e, miR-22, let-7d, and let-7e were seen in frozen-thawed spermatozoa with morphological abnormalities or low motility. The

cryopreservation of boar sperm also led to reduced expression of let-7c, miR-22, miR-26a, miR-186, and miR-450b-5p in frozen-thawed boar spermatozoa (Capra et al., 2017). The differential expression of these spermatozoal miRNAs may, thus, have implications in sperm fertility as they are associated with motility and apoptosis in frozen-thawed boar spermatozoa.

The miRNA microarray profile of high and low fertile boars found 326 pig-specific miRNAs (Alvarez-Rodriguez et al., 2020), among which miR-1285 was found to be related to sperm production, promoting AMPK phosphorylation and regulating oxidative stress (Jiao et al., 2015). Interestingly, miR-15/miR-16 was the most abundant miRNA and is known to suppress the TGF- β signaling pathway (Ayaz and Dinç, 2018) and enhance spermatogonial proliferation and spermatogenesis in gonadotoxic patients (Moraveji et al., 2019). The mRNA targets of these miRNA are involved in various essential cellular functions such as lipid metabolism (e.g., miR-4332), cell proliferation, and apoptosis (e.g., miR-671-5p, miR-425-5p) (Qiu et al., 2018). Moreover, miR-425 and IL-23 were downregulated in high fertile bulls, whereas its receptor (IL-23R) was down-regulated in high-fertility boars (Lu et al., 2019). The overexpressed miRNAs in high fertile boars included miR-191, miR-42, which play significant role in regulating the motility of spermatozoa inside the female reproductive tract (Zhang et al., 2018). It has also been indicated as a key activator of the NF- κ B signaling pathway, the regulator of estrogen receptors, and an inhibitor of the Wnt pathway (Alvarez-Rodriguez et al., 2020). Two down-regulated miRNAs, viz. miR-615, miR-221 are known to act as potential targets for EGFR and regulate acrosome reaction (Michailov et al., 2014) and PI3K-AKT and the estrogen signaling pathways (Alvarez-Rodriguez et al., 2019).

The pioneer report on sperm capacitation-specific miRNA profiling in boar was elucidated by (Li et al., 2018). This study identified DE miRNAs in fresh non-capacitated and capacitated spermatozoa. Some of the upregulated miRNAs in capacitated spermatozoa included, miR-148a-3p, miR-151-3p, miR-425-5p, miR-132, miR-451, miR7136-5p, miR-489, miR-1343, miR-1306-3p and fresh spermatozoa were enriched for miR-378b-3p, miR493-5p, miR-133a-3p, miR-362, and miR-214. Pathway analysis of miRNA targets revealed their biological significance in PI3K-AKT, MAPK, cAMP-PKA, and calcium signaling pathways, which are

TABLE 4 | Important non-coding microRNAs reported in pig spermatozoa by transcriptome analysis.

Species	Study type	Groups	MicroRNAs	Function	References
Pigs	qRT-PCR	Ejaculated spermatozoa	let-7a, -7d, and -7e miR-22	Spermatogenesis Sperm structure	Curry et al. (2011)
		Epididymal and ejaculated sperm	let-7a and miR-92a	Calcium and camp signalling	Chang et al. (2016)
			ssc-let-7a, ssc-let-7d, ssc-let-7e and ssc-miR-98 regulate miR-19 and miR-26a miR-224, miR-19b, miR-504 and miR-676 miR-26a, and miR-455-5p	Sperm apoptosis AKT/PKB signalling pathway Cell apoptosis and cell proliferation PTEN, and PI3K/AKT signalling pathway	Dai et al. (2019)
	RNA-Seq	Fresh vs frozen spermatozoa			
	RNA-Seq	Capacitated sperm	miR-127, miR-1343, miR-151-3p	Calcium signalling pathway and MAPK signalling pathway	Li et al. (2018)
	RNA-Seq	Fresh vs frozen semen	miR-17-5p, miR-26a-5p, miR-486-5p, miR-122-5p, miR-184, and miR-20a-5p let-7a, -7d, -7e, miR-22, let-7d, and let-7e miR-3155, miR-8197, miR-6727, miR-11796, miR-14189, miR-6125, and miR-13659 miRNAs, miR-26a, and miR-455-5p miR-10a and miR-1 were miR-615 miR-221	Regulate PTEN, PI3K/AKT, and STAT signalling influence motility, viability, and sperm apoptosis Sperm motility Bovine sperm fertilization Sperm motility Sperm motility Capacitation Wnt2, BDNF, CREB-related genes, PI3K-Akt and the estrogen signalling pathway	Dai et al. (2019) Alvarez-Rodriguez et al. (2020)
	Microarray	High fertile vs Low fertile			

considered necessary for protein tyrosine phosphorylation and sperm capacitation. These miRNA targets also play a significant role in cell proliferation, differentiation, sperm motility, hyper-activation, and acrosome reaction, primarily *via* the ERK (Ras/Raf/MEK/ERK) signaling pathway (Li et al., 2018). The cAMP-PKA signaling activates the Ca^{2+} channel and is considered a crucial step for sperm capacitation. The miR-134 was up-regulated in capacitated spermatozoa and targets *COL11A1* and *PDE4A* genes in PI3K-AKT and cAMP-PKA signaling pathways. Some other miRNA targets such as *AKAP3* for miR-1285, *VDAC1* and *HSPA2* for miR-127, *CATSPER4* for miR-151-3p regulate sperm motility, ATPase activity, sperm capacitation (Alvarez-rodriguez, 2017). *CABYR* and *ACRBP* regulate tyrosine phosphorylation during capacitation processes (Dong et al., 2015).

The abundance of miRNAs has also been documented in epididymal spermatozoa and seminal plasma of boar semen (Chen C. et al., 2017). A total of 221, 259, and 136 miRNAs were DE between ejaculated spermatozoa and epididymal spermatozoa; seminal plasma and epididymal spermatozoa; and seminal plasma and ejaculated spermatozoa, respectively. Further, three miRNAs *viz.* let-7a, miR-26a, and miR-10b were among the top ten most abundant miRNAs in ejaculated spermatozoa, epididymal spermatozoa, and seminal plasma. The most abundant spermatozoal miRNAs targeted mRNA transcripts of binding proteins, including metal ion binding, ATP binding, and nucleotide-binding. However, caution is to be exercised in analyzing spermatozoal miRNAs in boar spermatozoa collected at different seasons. Studies have shown that spermatozoal miRNAome may show seasonal variations in boars (Varona et al., 2019). For example, miR-34c, miR-221-3p, miR-362, miR-378, miR-106a, and miR-34c were down-regulated, whereas miR-1306-5p and miR-1249 were

upregulated in boar spermatozoa collected during the winter seasons. These miRNA targets primarily regulate fatty acid metabolism and oxidative stress (Wu et al., 2017), suggesting their role during the winter season. On the other hand, miR-106b, miR-378, miR-221 were previously reported to regulate autophagy machinery in different cell lines (Zhai et al., 2013; Tan et al., 2018).

Spermatozoal miRNAs in Stallion

A few reports have also documented the spermatozoal miRNAs in stallion spermatozoa. Direct sequencing of spermatozoal miRNAs in stallion sperm revealed 82 sperm-specific miRNAs (Das et al., 2013), out of which 68 miRNAs were previously reported in human spermatozoa (He et al., 2009). Several spermatozoal miRNAs of stallion were the same as identified in the sperm of men (Krawetz et al., 2011), boars (Curry et al., 2009), and mice (Liu et al., 2012). These spermatozoal miRNAs are noteworthy because they were absent in oocytes but were present in zygotes and were involved in regulating first cleavage division in mice (Krawetz et al., 2011; Liu et al., 2012). Three highly abundant miRNAs, *viz.* miR34B, miR34C, and miR449A, regulate early embryonic development either by direct interaction with mRNAs or *via* epigenetic mechanisms in humans (Liu et al., 2012). A total of 66 new miRNAs were reported in epididymal (cauda epididymis) spermatozoa, whose predicted pathways suggested their role in sperm motility, sperm viability, and early embryonic development (Das et al., 2013).

Other Spermatozoal ncRNAs Spermatozoal sncRNAs and piRNAs

Recent studies have documented the presence of thousands of sncRNAs in spermatozoa of several animal species. These

sncRNAs comprise rRNAs, tRNAs, tsRNAs, snoRNAs, piRNA, etc. (Sellem et al., 2020) and are believed to play essential regulatory roles in spermatogenesis and determining male fertility. The piRNAs are 26–32 nucleotide-long sncRNAs that exclusively represent the germline population and are considered “guardian of germline” by transposon surveillance to protect genome integrity (Krawetz et al., 2011). At tissue-specific levels, piRNA modulates key signaling pathways both at transcription and translational levels. These sncRNAs are characterized to maintain transposon silencing and genome stability during spermatogenesis (Luo et al., 2016). The piRNAs represent the most abundant sncRNA population in human sperm (Pantano et al., 2015) associated with sperm concentration and fertilization rate. Some reports evaluated their significance as semen quality parameters for cattle spermatozoa (Capra et al., 2017).

Several growing pieces of evidences suggest that piRNAs may regulate protein-coding genes in germ cells and engage in determining sperm fertility and early embryonic development (Suh and Blelloch, 2011). Although piRNAs are generally produced from repeat-associated regions or transposons by PIWIL2/PIWIL4-directed pathways, the PIWIL1-directed pathway is the primary production pathway in mature bull spermatozoa. In pig spermatozoa, piRNAs have been annotated to sperm morphology, spermatogenesis, acrosomal reaction, sperm hyperactivation, and male fertility (Ablondi et al., 2021). Some potential coding transcripts falling within the 5 kb region of centered piRNA are *CATSPER2*, *CATSPERG*, *OAZ3*, *ODF1*, *ODF2*, *PRM1*, *TEX14*, *TSSK2*, *TSSK3*, and *TSSK6*. The DE piRNAs in boar spermatozoa regulate spermatogenesis-related cell signaling pathways such as cAMP, cGMP, MAPK, and PI3K–AKT signaling pathways (Wang et al., 2021). Thus, taken together, the involvement of piRNA in the maintenance of genomic integrity during spermatogenesis and fertility seems to be a potential molecular parameter for evaluating male fertility in livestock animals.

Spermatozoal lncRNAs

Preliminary reports on lncRNAs considered these classes of RNAs as noise in the RNA content of a species due to lack of any protein-coding open coding frames (ORF) (Brownmiller et al., 2020). Later on, collective evidences supported their roles in epigenetic regulation, controlling transcription, and post-transcriptional mechanism (Peris-Frau et al., 2019). The role of lncRNA on spermatogenesis has been predicted by several researchers (Wallrapp et al., 2001; Bianchi et al., 2006; Forsberg et al., 2014; Falchi et al., 2018; Peris-Frau et al., 2019; Xu X. et al., 2020), but functional studies exploring their mechanism of action and relevance to spermatozoal fertility are lacking. Many targets of spermatozoal lncRNAs have been predicted (Gao et al., 2019) and were found to be enriched in apoptosis (e.g., PI3K–AKT, p53) (Xiong et al., 2019) and capacitation-related pathways (e.g., Calcium, cAMP, and MAPK signaling) (Gao et al., 2017). In stallion, comparative transcriptomics of dense and less-dense spermatozoa revealed 1,492 bp lncRNA as the most prevalent RNA (Ing et al., 2020). Further, the expression of 159 RNAs was higher in dense spermatozoa than in less-dense spermatozoa. Importantly, the dense spermatozoa resulted in a higher

pregnancy rate than those achieved by unfractionated stallion spermatozoa (Morrell et al., 2011).

Spermatozoal circRNAs

Circular RNAs (circRNAs) are a class of ncRNAs having a closed-loop structure formed by alternative back splicing of pre-mRNA in which the 3'-end of an exon is spliced to the 5'-end of an upstream exon (Gòdia et al., 2020). Depending on their genomic locations, they can be of exonic, intronic, and intergenic types (Zhou et al., 2019). The exonic circRNAs are preferentially located in the cytoplasm, whereas intronic and intergenic circRNAs are mainly found in the nucleus. Key circRNAs can participate in testis development or spermatogenesis (Xiong et al., 2019). CircRNAs Sry (circSry) is the first reported testicular circRNA in mice (Capel et al., 1993), whereas the first report on circRNA of spermatozoa was reported in boars (Gòdia et al., 2020). Gene Ontology of circRNA revealed their epigenetic functions such as histone modification, histone H3-K36 methylation, and chromatin organization during spermatogenesis, embryonic development. Four genes *viz.* *ATP6V0A2*, *PPA2*, *PAIP2*, and *PAXIP1* have been directly implicated in spermatozoal function and male fertility. Among these, the *PPA2* is a pyrophosphatase enzyme located at the mitochondrial membrane and is involved in ATP production during sperm capacitation and motility (Asghari et al., 2017). On the other hand, *PAXIP1* plays crucial functions in genomic stability and chromatin condensation during spermiogenesis, and its knockout resulted in testicular atrophy and male infertility in mice (Schwab et al., 2013). The *TESK2* is a protein kinase that is primarily expressed in round spermatids and is predicted to function during the early stages of spermatogenesis (Røsok et al., 1999). The *SPATA19* is vital for mitochondrial function and ATP production for sperm motility and fertilization (Luo et al., 2014). Genes related to early embryonic development (e.g., *ANGPT1*, *CDC73*, *DHX36*, *IPMK*, *RICTOR*, etc.) have also been found to be influenced by circRNAs. Remarkably, four circRNAs host genes (*DENND1B*, *PTK2*, *SLC5A10*, and *CAMSAP1*) showed a significant correlation with the motility of spermatozoa.

The circRNAs reported in adult vs. piglet testis (Zhang et al., 2021) were involved in regulating spermatogenic events (e.g., circRNA 10979 derived from *POC1A* gene) and germ cell development (e.g., circRNA 18456 derived from the *TDRD1* gene). Due to their stability and spatiotemporal specificity, circRNAs such as circRNA 10187, circRNA 6,682, circRNA 10979, and circRNA 18456 could be used as biomarkers of boar sexual maturity. The circRNA 1774 (derived from *CDC42* gene) and circRNA 18184 (derived from *PTEN* gene) were significantly downregulated whereas circRNA 40370 (derived from the *RICTOR* gene) was significantly upregulated in the testis of the sexually mature boars. The circRNAs were involved in signaling pathways that regulate stem cell pluripotency (e.g., *AKT3*, *AVCVR2A*, *FGFR1*, *ACVR1*, *FZD3*, *SMAD4*), tight junction (*MYH15*, *PRKCA*, *AMOTL1*, *PPP2CB*, *CDC42*, *PTEN*), and adhesion connections (e.g., *LEF1*, *NECTIN3*, *SMAD2*, *AFDN*), hedgehog signaling pathways (e.g., *HHIP*,

GSK3B, *PTCH1*, *SMO*, *RAB23*), cAMP signaling pathways (e.g., *PLD1*, *TIAM1*, *GNAI1*, *ADCY1*), mTOR signaling pathways (e.g., *BRAF*, *RICTOR*, *HIF1A*), and phosphatidylinositol signaling systems (e.g., *CDS1*, *DGKH*, *PLCB1*, *DGKK*). The expression profile of circRNAs in neonatal and adult cattle testis revealed some of the potential target genes of circRNAs mainly involved in cell-to-cell junction, oocyte maturation, and TGF β signaling pathway (Gao et al., 2018). These transcripts included *PIWIL1*, *DPY19L2*, *SLC26A8*, *IFT81*, *SMC1B*, *IQCG*, and *TTL5*. The host genes were associated with spermatogenesis and included *PIWIL1*, *SPATA6*, *TGF β 2*, *TGFBR2*, *ACVR2A*, and *SMAD2*.

RNA Population in Seminal Plasma and Their Contribution to spRNA

Seminal plasma is the secretions of accessory glands containing proteins and RNAs and has been overlooked for decades. Seminal plasma components have no role in assisted reproductive technologies (ART) and, therefore, are excluded from IVF and intracytoplasmic injection (ICSI) of ejaculated sperm. The ARTs have resulted in successful fertilization, pregnancy, and live birth of offspring in humans (Marzano et al., 2020) and animals (Morgan et al., 2020) without the use of seminal plasma. However, recent studies have found that the spRNA population may be contributed by seminal plasma via exosome. Exosomes of seminal plasma are released by epididymis and other accessory sex organs, which interact with spermatozoa to unload RNA cargo to the sperm. Therefore, analyzing the exosome of the epididymis (epididymosomes) or seminal plasma may be used as a marker of reproductive disease and male infertility [reviewed in (Vickram et al., 2021)]. Further, each ejaculation contain trillions of exosomes that aids in fertilization in two ways, 1) exosome exerts immunosuppressive effects on cells of mucosa layer and, 2) transfer of RNAs to the spermatozoa (Jodar, 2019). In humans, exosomes contained miRNAs (21.7% of the total RNA), Y-RNA, mRNAs, mature piwi-RNAs, and tRNAs, suggesting that exosome delivers regulatory signals to the spermatozoa (Vojtech et al., 2014).

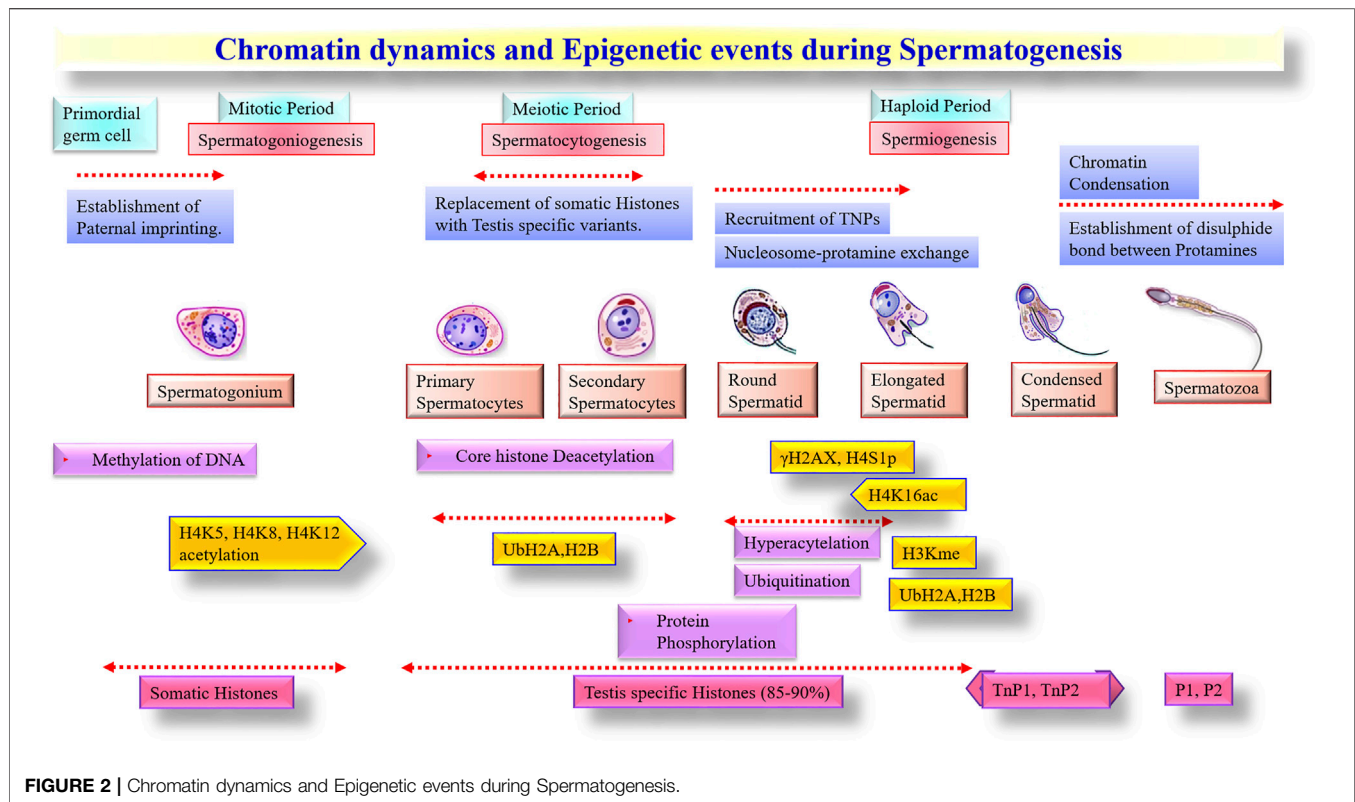
In pigs, sequencing of seminal plasma-derived extracellular vesicles showed diverse small RNAs, including mRNA (25% of the total reads), tsRNA (0.01% of the total reads), miRNA, and piRNA. There were 325 miRNAs, of which 37 novel miRNAs were identified in boar (Xu Z. et al., 2020). The study identified the adverse effects of miR-21-5p on boar sperm fertility. Likewise, miRNA and piRNA have also been identified in beef bulls' seminal plasma. Of 617 small RNA, nine miRNA were DE between high and low fertility beef bulls (Stephanie, 2016). The role of seminal plasma RNA in determining male fertility seems to be undeniable. In another study, expression of *PRM1* in seminal plasma was positively correlated with mitochondrial membrane potential, and lateral movement of sperm head was associated with expression of *BMP2*, *UBE2D3*, *TRADD*, and *CASP3* (Shilpa et al., 2017). In this study, investigators have found a positive association of high expression of nerve growth factor (NGF) in the maintenance of post-thaw integrity of

membrane of bull's spermatozoa. More studies need to be conducted in other livestock animals as well to ascertain the effectiveness of seminal plasma RNAs as markers of the reproductive performance of male animals.

Epigenetic Control of spRNA Expression

The terminal stages of spermatogenesis are accompanied by chromatin condensation and the replacement of histones by protamines. These changes are primarily driven by transition proteins and cleavage of rRNAs by nuclease activities and lead to the progressive shutdown of transcriptional and translational activities in mature spermatozoa (Jodar, 2019). Interestingly, however, some studies have documented post-translational modifications of histones in mature spermatozoa and included acetylation (Johnson et al., 2011), ubiquitination (Wang T. et al., 2019), methylation (Johnson et al., 2011), and phosphorylation (Dada et al., 2012) that are known to be involved in regulating chromatin remodeling. Thus, some canonical structures of histone variants appear to have remained unchanged throughout the final stages of spermatogenesis to maintain a hierarchical layer of genomic organization in paternal chromatin. Indeed, studies have shown that about 15% of the paternal histone remains associated with the genome in mature spermatozoa (Ward, 2009). Initially, these segments were considered a mark of partial/incomplete transition of histone-protamine exchange, but gradually these histone segments were noted with a significant regulatory network by post-translational modifications (Barrachina et al., 2018). Notably, the histone retention in spermatozoa was not randomly positioned on chromosomes, instead distinctly localized within the nucleus (Johnson et al., 2011). It is, therefore, believed that specific regions in sperm chromatin are differentially marked with modified histones to have the potential for transcription through epigenetic regulations (Krawetz et al., 2011).

A substantial proportion of coding spRNAs detected in mature sperm are also considered as a spermatogenic leftover from transcription events during the spermatogenesis process (Jodar, 2019) and were under epigenetic control. During spermatid elongation, protamines undergo phosphorylation by serine/arginine protein-specific kinase 1 (*SRPK1*) and calcium/calmodulin-dependent protein kinase 4 (*CAMK4*), followed by rapid dephosphorylation to establish disulfide bonds between the unmasked cysteine residues of dephosphorylated protamines (Carrell et al., 2007). Histone acetylation was shown to be essential for the initiation of chromatin remodeling in spermatids (Stiavnicka et al., 2016). The hyperacetylated testicular histones were gradually replaced upon relaxation of the nucleosome complex. Testicular histone variants (*H2B* and *TH2B*) incorporate into spermatids with the help of transition proteins (*TNP1*, *TNP2*), which supersede with protamines (*PRM1*, *PRM2*) (Carrell et al., 2007). Histone methylation also has a significant role in the differentiation of spermatogonia, especially for H3me (H3 methylation) and H4me (H4 methylation) variants. It was reported that reduced H4me and hyperacetylation of H4 during spermatid elongation collectively mediated the histone-protamine replacement (SHIRAKATA et al., 2014). The significance of ubiquitination has also been



implicated in spermatogenesis and sperm maturation to eliminate dead and defective spermatozoa and epididymal cells (Røsok et al., 1999). The stage-specific significance of epigenetic alternations during spermatogenesis is summarised (Figure 2).

Relevance of spRNAs in Sperm Function in Livestock Species

While recent literature on Microarray and high-throughput RNAseq data have provided novel information on the repertoire of RNA population in spermatozoa, their functional validation is largely missing in the literature. Moreover, common genes and pathways across different species are yet to be fully deciphered. Nevertheless, the involvement of spRNAs in the ontology of capacitation, motility, and fertilization are apparent in various species. Among all spRNAs, protamines were the most prominent sperm-specific transcript, which is known to affect fertility in cattle (Chen et al., 2015), pigs (Gòdia et al., 2018), horses (Kadivar et al., 2020), and human (Jodar et al., 2013). Similarly, the phospholipase C (PLC)-mediated pathway was the most common pathway affected by spRNA in cattle (Selvaraju et al., 2017), pigs (Kasimanickam and Kastelic, 2016), and horses (Varner et al., 1993). The PLC is an essential regulator of intracellular Ca^{2+} oscillations that play a critical role during oocyte activation. Several studies also identified Calcium ion channels such as *CatSper* in the spRNA population, which is essential in regulating capacitation and sperm motility in livestock species as well as humans (Jan et al., 2017; Li et al., 2018; Singh et al., 2019; Nicolas et al.,

2020). Besides these common sperm transcripts across various livestock species, there were flagellar specific spRNAs and sperm motility proteins such as *BSP* (de Souza et al., 2017; Pardede et al., 2020), *AKAP* (Gilbert et al., 2007; Chatterjee et al., 2010; Kasimanickam and Kastelic, 2016; Ing et al., 2020), *ODF* (Chen et al., 2015; Li et al., 2018; Pardede et al., 2020; Özbek et al., 2021), zinc finger nucleases (Card et al., 2013; Kasimanickam and Kastelic, 2016; Corral-Vazquez et al., 2021) and heat shock proteins. The heat shock proteins are a group of tyrosine regulatory elements that participate in hyperactivation and nitric oxide synthesis during fertilization (Feugang et al., 2010; Mohamad et al., 2018; Nicolas et al., 2020; Lian et al., 2021; Sun et al., 2021). The spRNAs such as *calmegin* (Guyonnet et al., 2009; Card et al., 2013), *clusterin* (Singh et al., 2019; Lian et al., 2021; Zhao et al., 2021), *TSSK* (Bissonnette et al., 2009; Li et al., 2020; Ablondi et al., 2021), and *CABYR* (Selvaraju et al., 2017) were also reported in cattle and pigs, are known to be essential for male fertility.

A few transcripts related to sperm capacitation were also reported in the spRNA population of cattle, pig, and horse spermatozoa. The important spRNAs related to sperm capacitation included *CABYR* (Bailey, 2010) and *AKAP* (Maciel et al., 2018), which regulate calcium-binding and tyrosine phosphorylation. On the other hand, *CatSper*, *VDAC1*, and *HSPA2* are involved in calcium signaling whereas, *COL11A1*, *PDE4A* can participate in PI3K-Akt and cAMP-PKA signaling pathways (Li et al., 2018). The *HSP70* (Mohamad et al., 2018) and *HSP90* (Jin and Yang, 2017), involved in calcium signaling, were also reported in spRNA

population. Thus, it appears that calcium signaling and tyrosine phosphorylation are important events of capacitation that spRNAs may mediate. Functional studies such as gain or loss-of function research are warranted to verify these findings further.

CONCLUSION

In conclusion, various types of protein-coding and non-coding spRNAs have been documented in the literature with their potential roles in regulating male reproduction and fertility. These spRNAs may be exploited *via* transcriptome analysis of spermatozoa to improve the conception rate of livestock by crossbreeding, artificial insemination, or ARTs. Microarray and RNA-seq have been extensively used as high throughput transcriptome analysis tools and have established sperm transcriptome as a subset yet distinct from testicular transcriptome with some uniquely expressed transcripts in spermatozoa. The spRNA profile, generated by microarray or RNA-seq, may serve as a molecular signature to identify semen with superior fertilizability, freezability, and fertility. Alternatively, exosome analysis of seminal plasma, which

contributes to the spRNA population, may be used as a proxy to spRNAs profiling. It is also emphasized that analysis of spRNA population by high sensitive high throughput sequencing technologies should involve stringent quality control measures to avoid somatic cell contamination and batch-to-batch variation in spRNA isolation protocols. Further, loss- and gain-of function studies are required to validate the function of spRNAs in spermatogenesis, fertilization, and early embryonic development. Profiling of spRNA may also prove helpful in understanding the mechanism of action of genotoxic agents, drugs, capacitation agents, motility enhancers, etc. The next generation sequencing (NGS) of spRNA may find its application in semen evaluation for diagnosing idiopathic male infertility and devising newer methods for treatment.

AUTHOR CONTRIBUTIONS

MG and RC designed the study and proofread the draft manuscript. BS, MG, PS, RC, and SC wrote the manuscript. All authors reviewed and approved the manuscript for submission.

REFERENCES

- Ablondi, M., Gòdia, M., Rodríguez-Gil, J. E., Sánchez, A., and Clop, A. (2021). Characterisation of Sperm piRNAs and Their Correlation with Semen Quality Traits in Swine. *Anim. Genet.* 52, 114–120. doi:10.1111/age.13022
- Alvarez-Rodriguez, M., Atikuzzaman, M., Venhoranta, H., Wright, D., and Rodriguez-Martinez, H. (2019). Expression of Immune Regulatory Genes in the Porcine Internal Genital Tract Is Differentially Triggered by Spermatozoa and Seminal Plasma. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20030513
- Alvarez-rodriguez, M. (2017). *iMedPub Journals Exogenous Individual Lecithin-Phospholipids (Phosphatidylcholine and Phosphatidylglycerol) Cannot Prevent the Oxidative Stress Imposed by Cryopreservation of Boar Sperm Abstract.*
- Alvarez-Rodriguez, M., Martinez, C., Wright, D., Barranco, I., Roca, J., and Rodriguez-Martinez, H. (2020). The Transcriptome of Pig Spermatozoa, and its Role in Fertility. *Int. J. Mol. Sci.* 21, 1572. doi:10.3390/ijms21051572
- Amanai, M., Brahmajosyula, M., and Perry, A. C. F. (2006). A Restricted Role for Sperm-Borne MicroRNAs in Mammalian Fertilization. *Biol. Reprod.* 75, 877–884. doi:10.1095/biolreprod.106.056499
- Arcuri, F., Papa, S., Carducci, A., Romagnoli, R., Liberatori, S., Riparbelli, M. G., et al. (2004). Translationally Controlled Tumor Protein (TCTP) in the Human Prostate and Prostate Cancer Cells: Expression, Distribution, and Calcium Binding Activity. *Prostate* 60, 130–140. doi:10.1002/pros.20054
- Asghari, A., Marashi, S.-A., and Ansari-Pour, N. (2017). A Sperm-specific Proteome-Scale Metabolic Network Model Identifies Non-glycolytic Genes for Energy Deficiency in Asthenozoospermia. *Syst. Biol. Reprod. Med.* 63, 100–112. doi:10.1080/19396368.2016.1263367
- Ayaz, L., and Dinç, E. (2018). Evaluation of microRNA Responses in ARPE-19 Cells against the Oxidative Stress. *Cutan. Ocul. Toxicol.* 37, 121–126. doi:10.1080/15569527.2017.1355314
- Bailey, J. L. (2010). Factors Regulating Sperm Capacitation. *Syst. Biol. Reprod. Med.* 56, 334–348. doi:10.3109/19396368.2010.512377
- Balhorn, R. (2007). The Protamine Family of Sperm Nuclear Proteins. *Genome Biol.* 8, 227. doi:10.1186/gb-2007-8-9-227
- Barrachina, F., Soler-Ventura, A., Oliva, R., and Jodar, M. (2018). Sperm Nucleoproteins (Histones and Protamines). A. *Clin. Guid. Sperm DNA Chromatin Damage*, 31–51. doi:10.1007/978-3-319-71815-6_2
- Bellvé, A. R., Cavicchia, J. C., Millette, C. F., O'Brien, D. A., Bhatnagar, Y. M., and Dym, M. (1977). Spermatogenic Cells of the Prepuberal Mouse. Isolation and Morphological Characterization. *J. Cel Biol.* 74, 68–85. doi:10.1083/jcb.74.1.68
- Bettgowda, A., and Wilkinson, M. F. (2010). Transcription and post-transcriptional Regulation of Spermatogenesis. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 1637–1651. doi:10.1098/rstb.2009.0196
- Bianchi, N. O., Richard, S. M., and Pavicic, W. (2006). Y Chromosome Instability in Testicular Cancer. *Mutat. Res.* 612, 172–188. doi:10.1016/j.mrrev.2005.12.001
- Bissonnette, N., Lévesque-Sergerie, J. P., Thibault, C., and Boissonneault, G. (2009). Spermatozoal Transcriptome Profiling for Bull Sperm Motility: A Potential Tool to Evaluate Semen Quality. *Reproduction* 138, 65–80. doi:10.1530/REP-08-0503
- Boerke, A., Dieleman, S. J., and Gadella, B. M. (2007). A Possible Role for Sperm RNA in Early Embryo Development. *Theriogenology* 68, 147–155. doi:10.1016/j.theriogenology.2007.05.058
- Brownmiller, T., Juric, J. A., Ivey, A. D., Harvey, B. M., Westemeier, E. S., Winters, M. T., et al. (2020). Y Chromosome LncRNA Are Involved in Radiation Response of Male Non-small Cell Lung Cancer Cells. *Cancer Res.* 80, 4046–4057. doi:10.1158/0008-5472.CAN-19-4032
- Cammas, L., Reinaud, P., Dubois, O., Bordas, N., Germain, G., and Charpigny, G. (2005). Identification of Differentially Regulated Genes during Elongation and Early Implantation in the Ovine Trophoblast Using Complementary DNA Array Screening1. *Biol. Reprod.* 72, 960–967. doi:10.1095/biolreprod.104.034801
- Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., et al. (1993). Circular Transcripts of the Testis-Determining Gene Sry in Adult Mouse Testis. *Cell* 73, 1019–1030. doi:10.1016/0092-8674(93)90279-Y
- Cappallo-Obermann, H., Schulze, W., Jastrow, H., Baukloh, V., and Spiess, A. N. (2011). Highly Purified Spermatozoal RNA Obtained by a Novel Method Indicates an Unusual 28S/18S rRNA Ratio and Suggests Impaired Ribosome Assembly. *Mol. Hum. Reprod.* 17, 669–678. doi:10.1093/molehr/gar037
- Capra, E., Turri, F., Lazzari, B., Cremonesi, P., Gliozzi, T. M., Fojadelli, I., et al. (2017). Small RNA Sequencing of Cryopreserved Semen from Single Bull Revealed Altered miRNAs and piRNAs Expression between High- and Low-Motile Sperm Populations. *BMC Genomics* 18, 14. doi:10.1186/s12864-016-3394-7
- Card, C. J., Anderson, E. J., Zamberlan, S., Krieger, K. E., Kaproth, M., and Sartini, B. L. (2013). Cryopreserved Bovine Spermatozoal Transcript Profile as Revealed by High-Throughput Ribonucleic Acid Sequencing1. *Biol. Reprod.* 88, 1–9. doi:10.1095/biolreprod.112.103788
- Card, C. J., Krieger, K. E., Kaproth, M., and Sartini, B. L. (2017). Oligo-dT Selected Spermatozoal Transcript Profiles Differ Among Higher and Lower Fertility Dairy Sires. *Anim. Reprod. Sci.* 177, 105–123. doi:10.1016/j.anireprosci.2016.12.011

- Carrell, D. T., Emery, B. R., and Hammoud, S. (2007). Altered Protamine Expression and Diminished Spermatogenesis: what Is the Link? *Hum. Reprod. Update* 13, 313–327. doi:10.1093/humupd/dml057
- Castillo, J., Jodar, M., and Oliva, R. (2018). The Contribution of Human Sperm Proteins to the Development and Epigenome of the Preimplantation Embryo. *Hum. Reprod. Update* 24, 535–555. doi:10.1093/humupd/dmy017
- Chang, Y., Dai, D. H., Li, Y., Zhang, Y., Zhang, M., Zhou, G. B., et al. (2016). Differences in the Expression of microRNAs and Their Predicted Gene Targets between Cauda Epididymal and Ejaculated Boar Sperm. *Theriogenology* 86, 2162–2171. doi:10.1016/j.theriogenology.2016.07.012
- Chatterjee, M., Nandi, P., Ghosh, S., and Sen, P. C. (2010). Regulation of Tyrosine Kinase Activity during Capacitation in Goat Sperm. *Mol. Cel. Biochem.* 336, 39–48. doi:10.1007/s11010-009-0261-8
- Chen, C., Wu, H., Shen, D., Wang, S., Zhang, L., Wang, X., et al. (2017a). Comparative Profiling of Small RNAs of Pig Seminal Plasma and Ejaculated and Epididymal Sperm. *Reproduction* 153, 785–796. doi:10.1530/REP-17-0014
- Chen, X., Li, X., Guo, J., Zhang, P., and Zeng, W. (2017b). The Roles of microRNAs in Regulation of Mammalian Spermatogenesis. *J. Anim. Sci. Biotechnol.* 8, 1–8. doi:10.1186/s40104-017-0166-4
- Chen, X., Wang, Y., Zhu, H., Hao, H., Zhao, X., Qin, T., et al. (2015). Comparative Transcript Profiling of Gene Expression of Fresh and Frozen-Thawed Bull Sperm. *Theriogenology* 83, 504–511. doi:10.1016/j.theriogenology.2014.10.015
- Corral-Vazquez, C., Blanco, J., Aiese Cigliano, R., Sarrate, Z., Rivera-Egea, R., Vidal, F., et al. (2021). The RNA Content of Human Sperm Reflects Prior Events in Spermatogenesis and Potential post-fertilization Effects. *Mol. Hum. Reprod.* 27, 1–15. doi:10.1093/molehr/gaab035
- Cui, N., Hao, G., Zhao, Z., Wang, F., Cao, J., and Yang, A. (2016). MicroRNA-224 Regulates Self-Renewal of Mouse Spermatogonial Stem Cells via Targeting DMRT1. *J. Cel. Mol. Med.* 20, 1503–1512. doi:10.1111/jcmm.12838
- Curry, E., Ellis, S. E., and Pratt, S. L. (2009). Detection of Porcine Sperm MicroRNAs Using a Heterologous MicroRNA Microarray and Reverse Transcriptase Polymerase Chain Reaction. *Mol. Reprod. Dev.* 76, 218–219. doi:10.1002/mrd.20980
- Curry, E., Safranski, T. J., and Pratt, S. L. (2011). Differential Expression of Porcine Sperm microRNAs and Their Association with Sperm Morphology and Motility. *Theriogenology* 76, 1532–1539. doi:10.1016/j.theriogenology.2011.06.025
- Dada, R., Kumar, M., Jesudasan, R., Fernández, J. L., Gosálvez, J., and Agarwal, A. (2012). Epigenetics and its Role in Male Infertility. *J. Assist. Reprod. Genet.* 29, 213–223. doi:10.1007/s10815-012-9715-0
- Dadoue, J. P. (2009). Spermatozoal RNAs: What about Their Functions? *Microsc. Res. Tech.* 72, 536–551. doi:10.1002/jemt.20697
- Dai, D.-H., Qazi, I., Ran, M.-X., Liang, K., Zhang, Y., Zhang, M., et al. (2019). Exploration of miRNA and mRNA Profiles in Fresh and Frozen-Thawed Boar Sperm by Transcriptome and Small RNA Sequencing. *Int. J. Mol. Sci.* 20, 802. doi:10.3390/ijms20040802
- Das, P. J., McCarthy, F., Vishnoi, M., Paria, N., Gresham, C., Li, G., et al. (2013). Stallion Sperm Transcriptome Comprises Functionally Coherent Coding and Regulatory RNAs as Revealed by Microarray Analysis and RNA-Seq. *PLoS One* 8, 1–17. doi:10.1371/journal.pone.0056535
- Das, P. J., Paria, N., Gustafson-Seabury, A., Vishnoi, M., Chaki, S. P., Love, C. C., et al. (2010). Total RNA Isolation from Stallion Sperm and Testis Biopsies. *Theriogenology* 74, 1099–1106. doi:10.1016/j.theriogenology.2010.04.023
- de Souza, A. P. B., Schorr-Lenz, Á. M., Lucca, F., and Cunha Bustamante-Filho, I. (2017). The Epididymis and its Role on Sperm Quality and Male Fertility. *Anim. Reprod.* 14, 1234–1244. doi:10.21451/1984-3143-AR955
- Dong, H. T., Shi, W. S., Tian, Y., Cao, L. P., and Jin, Y. (2015). Expression and Tyrosine Phosphorylation of Sp32 Regulate the Activation of the Boar Proacrosin/acrosin System. *Genet. Mol. Res.* 14, 2374–2383. doi:10.4238/2015.March.27.23
- Dos Santos da Silva, L., Borges Domingues, W., Fagundes Barreto, B., da Silveira Martins, A. W., Dellagostin, E. N., Komninou, E. R., et al. (2021). Capillary Electrophoresis Affects the Expression of miRNA-122-5p from Bull Sperm Cells. *Gene* 768, 145286. doi:10.1016/j.gene.2020.145286
- Falchi, L., Galleri, G., Zedda, M. T., Pau, S., Bogliolo, L., Ariu, F., et al. (2018). Liquid Storage of Ram Semen for 96 H: Effects on Kinematic Parameters, Membranes and DNA Integrity, and ROS Production. *Livest. Sci.* 207, 1–6. doi:10.1016/j.livsci.2017.11.001
- Fang, P., Xu, W., Li, D., Zhao, X., Dai, J., Wang, Z., et al. (2015). A Novel Acrosomal Protein, IQCF1, Involved in Sperm Capacitation and the Acrosome Reaction. *Andrology* 3, 332–344. doi:10.1111/andr.296
- Feugang, J. M., Rodriguez-Osorio, N., Kaya, A., Wang, H., Page, G., Ostermeier, G. C., et al. (2010). Transcriptome Analysis of Bull Spermatozoa: Implications for Male Fertility. *Reprod. Biomed. Online* 21, 312–324. doi:10.1016/j.rbmo.2010.06.022
- Forsberg, L. A., Rasi, C., Malmqvist, N., Davies, H., Pasupulati, S., Pakalapati, G., et al. (2014). Mosaic Loss of Chromosome Y in Peripheral Blood Is Associated with Shorter Survival and Higher Risk of Cancer. *Nat. Genet.* 46, 624–628. doi:10.1038/ng.2966
- Fort, V., Khelifi, G., and Hussein, S. M. I. (2021). Long Non-coding RNAs and Transposable Elements: A Functional Relationship. *Biochim. Biophys. Acta - Mol. Cel. Res.* 1868, 118837. doi:10.1016/j.bbamcr.2020.118837
- Foxcroft, G. R., Dyck, M. K., Ruiz-Sanchez, A., Novak, S., and Dixon, W. T. (2008). Identifying Useable Semen. *Theriogenology* 70, 1324–1336. doi:10.1016/j.theriogenology.2008.07.015
- Fraser, L. R., and Adeoya-Osiguwa, S. A. (2001). Fertilization Promoting Peptide — A Possible Regulator of Sperm Function *In Vivo*. *Vitam. Horm.* 63, 1–28. doi:10.1016/S0083-6729(01)63001-2
- Gao, F., Zhang, P., Zhang, H., Zhang, Y., Zhang, Y., Hao, Q., et al. (2017). Dysregulation of Long Noncoding RNAs in Mouse Testes and Spermatozoa after Exposure to Cadmium. *Biochem. Biophys. Res. Commun.* 484, 8–14. doi:10.1016/j.bbrc.2017.01.091
- Gao, Y., Li, S., Lai, Z., Zhou, Z., Wu, F., Huang, Y., et al. (2019). Analysis of Long Non-coding RNA and mRNA Expression Profiling in Immature and Mature Bovine (*Bos taurus*) Testes. *Front. Genet.* 10, 1–13. doi:10.3389/fgene.2019.00646
- Gao, Y., Wu, M., Fan, Y., Li, S., Lai, Z., Huang, Y., et al. (2018). Identification and Characterization of Circular RNAs in Qinchuan Cattle Testis. *R. Soc. Open Sci.* 5, doi:10.1098/rsos.180413
- Gilbert, I., Bissonnette, N., Boissonneault, G., Vallée, M., and Robert, C. (2007). A Molecular Analysis of the Population of mRNA in Bovine Spermatozoa. *Reproduction* 133, 1073–1086. doi:10.1530/REP-06-0292
- Gòdia, M., Castelló, A., Rocco, M., Cabrera, B., Rodríguez-Gil, J. E., Balasch, S., et al. (2020). Identification of Circular RNAs in Porcine Sperm and Evaluation of Their Relation to Sperm Motility. *Sci. Rep.* 10, 1–11. doi:10.1038/s41598-020-64711-z
- Gòdia, M., Mayer, F. Q., Nafissi, J., Castelló, A., Rodríguez-Gil, J. E., Sánchez, A., et al. (2018). A Technical Assessment of the Porcine Ejaculated Spermatozoa for a Sperm-specific RNA-Seq Analysis. *Syst. Biol. Reprod. Med.* 64, 291–303. doi:10.1080/19396368.2018.1464610
- Gomes, A. Q., Nolasco, S., and Soares, H. (2013). Non-coding RNAs: Multi-Tasking Molecules in the Cell. *Int. J. Mol. Sci.* 14, 16010–16039. doi:10.3390/ijms140816010
- Govindaraju, A., Uzun, A., Robertson, L., Atli, M. O., Kaya, A., Topper, E., et al. (2012). Dynamics of microRNAs in Bull Spermatozoa. *Reprod. Biol. Endocrinol.* 10, 82. doi:10.1186/1477-7827-10-82
- Griswold, M. D. (2016). Spermatogenesis: The Commitment to Meiosis. *Physiol. Rev.* 96, 1–17. doi:10.1152/physrev.00013.2015
- Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A Novel Class of Small RNAs in Mouse Spermatogenic Cells. *Genes Dev.* 20, 1709–1714. doi:10.1101/gad.1434406
- Guyonnet, B., Marot, G., Dacheux, J.-L., Mercat, M.-J., Schwob, S., Jaffrézic, F., et al. (2009). The Adult Boar Testicular and Epididymal Transcriptomes. *BMC Genomics* 10, 369. doi:10.1186/1471-2164-10-369
- Hamatani, T. (2012). Human Spermatozoal RNAs. *Fertil. Steril.* 97, 275–281. doi:10.1016/j.fertnstert.2011.12.035
- Hayashi, S. (2003). Mouse Preimplantation Embryos Developed from Oocytes Injected with Round Spermatids or Spermatozoa Have Similar but Distinct Patterns of Early Messenger RNA Expression. *Biol. Reprod.* 69, 1170–1176. doi:10.1095/biolreprod.103.016832
- He, Z., Jiang, J., Kokkinaki, M., Tang, L., Zeng, W., Gallicano, I., et al. (2013). MiRNA-20 and Mirna-106a Regulate Spermatogonial Stem Cell Renewal at the post-transcriptional Level via Targeting STAT3 and Ccnd1. *Stem Cells* 31, 2205–2217. doi:10.1002/stem.1474
- He, Z., Kokkinaki, M., Pant, D., Gallicano, G. I., and Dym, M. (2009). Small RNA Molecules in the Regulation of Spermatogenesis. *Reproduction* 137, 901–911. doi:10.1530/REP-08-0494

- Hecht, N. B. (1998). Molecular Mechanisms of Male Germ Cell Differentiation. *BioEssays* 20, 555–561. doi:10.1002/(SICI)1521-1878
- Hermo, L., Pelletier, R.-M., Cyr, D. G., and Smith, C. E. (2010). Surfing the Wave, Cycle, Life History, and Genes/proteins Expressed by Testicular Germ Cells. Part 2: Changes in Spermatid Organelles Associated with Development of Spermatozoa. *Microsc. Res. Tech.* 73, 279–319. doi:10.1002/jemt.20787
- Huang, Y.-L., Huang, G.-Y., Lv, J., Pan, L.-N., Luo, X., and Shen, J. (2017). miR-100 Promotes the Proliferation of Spermatogonial Stem Cells via Regulating Stat3. *Mol. Reprod. Dev.* 84, 693–701. doi:10.1002/mrd.22843
- Ing, N. H., Konganti, K., Ghaffari, N., Johnson, C. D., Forrest, D. W., Love, C. C., et al. (2020). Identification and Quantification of Coding and Long Non-coding RNAs in Stallion Spermatozoa Separated by Density. *Andrology* 8, 1409–1418. doi:10.1111/andr.12791
- Jambor, T., Jana, B., Hana, G., Eva, T., and Norbert, L. (2017). “Male Reproduction: One of the Primary Targets of Bisphenol,” in *Bisphenol A Exposure and Health Risks*. Editors P. Erkekoglu and B. Kocer-Gumusel (Intech). doi:10.5772/intechopen.68629
- Jan, S. Z., Vormer, T. L., Jongejan, A., Röling, M. D., Silber, S. J., de Rooij, D. G., et al. (2017). Unraveling Transcriptome Dynamics in Human Spermatogenesis. *Dev* 144, 3659–3673. doi:10.1242/dev.152413
- Jiao, Z. J., Yi, W., Rong, Y. W., Kee, J. D., and Zhong, W. X. (2015). MicroRNA-1285 Regulates 17 β -Estradiol-Inhibited Immature Boar Sertoli Cell Proliferation via Adenosine Monophosphate-Activated Protein Kinase Activation. *Endocrinology* 156, 4059–4070. doi:10.1210/en.2014-1982
- Jin, S. K., and Yang, W. X. (2017). Factors and Pathways Involved in Capacitation: How Are They Regulated? *Oncotarget* 8, 3600–3627. doi:10.18632/oncotarget.12274
- Jodar, M., Selvaraju, S., Sandler, E., Diamond, M. P., and Krawetz, S. A. (2013). The Presence, Role and Clinical Use of Spermatozoal RNAs. *Hum. Reprod. Update* 19, 604–624. doi:10.1093/humupd/dmt031
- Jodar, M. (2019). Sperm and Seminal Plasma RNAs: what Roles Do They Play beyond Fertilization? *Reproduction* 158, R113–R123. doi:10.1530/REP-18-0639
- Johnson, G. D., Lalancette, C., Linnemann, A. K., Leduc, F., Boissonneault, G., and Krawetz, S. A. (2011). The Sperm Nucleus: Chromatin, RNA, and the Nuclear Matrix. *Reproduction* 141, 21–36. doi:10.1530/REP-10-0322
- Jung, Y. H., Gupta, M. K., Shin, J. Y., Uhm, S. J., and Lee, H. T. (2010). MicroRNA Signature in Testes-Derived Male Germ-Line Stem Cells. *MHR Basic Sci. Reprod. Med.* 16, 804–810. doi:10.1093/molehr/gaq058
- Kadivar, A., Shams Esfandabadi, N., Dehghani Nazhvani, E., Shirazi, A., and Ahmadi, E. (2020). Effects of Cryopreservation on Stallion Sperm Protamine Messenger RNAs. *Reprod. Domest. Anim.* 55, 274–282. doi:10.1111/rda.13615
- Kasimanickam, V., Kasimanickam, R., Arangasamy, A., Saberivand, A., Stevenson, J. S., and Kastelic, J. P. (2012). Association between mRNA Abundance of Functional Sperm Function Proteins and Fertility of Holstein Bulls. *Theriogenology* 78, 2007–2019. doi:10.1016/j.theriogenology.2012.07.016
- Kasimanickam, V., and Kastelic, J. (2016). MicroRNA in Sperm from Duroc, Landrace and Yorkshire Boars. *Sci. Rep.* 6, 1–11. doi:10.1038/srep32954
- Keles, E., Malama, E., Bozukova, S., Siuda, M., Wyck, S., Witschi, U., et al. (2021). The Micro-RNA Content of Unsorted Cryopreserved Bovine Sperm and its Relation to the Fertility of Sperm after Sex-Sorting. *BMC Genomics* 22, 30. doi:10.1186/s12864-020-07280-9
- Kempisty, B., Antosik, P., Bukowska, D., Jackowska, M., Lianeri, M., Jaśkowski, J. M., et al. (2008). Analysis of Selected Transcript Levels in Porcine Spermatozoa, Oocytes, Zygotes and Two-Cell Stage Embryos. *Reprod. Fertil. Dev.* 20, 513–518. doi:10.1071/rd07211
- Krawetz, S. A., Kruger, A., Lalancette, C., Tagett, R., Anton, E., Draghici, S., et al. (2011). A Survey of Small RNAs in Human Sperm. *Hum. Reprod.* 26, 3401–3412. doi:10.1093/humrep/der329
- Kroft, T. L., Patterson, J., Won Yoon, J., Doglio, L., Walterhouse, D. O., Iannaccone, P. M., et al. (2001). GLI1 Localization in the Germinal Epithelial Cells Alternates between Cytoplasm and Nucleus: Upregulation in Transgenic Mice Blocks Spermatogenesis in Pachytene1. *Biol. Reprod.* 65, 1663–1671. doi:10.1095/biolreprod65.6.1663
- Kumaresan, A., Das Gupta, M., Datta, T. K., and Morrell, J. M. (2020). Sperm DNA Integrity and Male Fertility in Farm Animals: A Review. *Front. Vet. Sci.* 7, 1–15. doi:10.3389/fvets.2020.00321
- Lambard, S., Galeraud-Denis, I., Martin, G., Levy, R., Chocat, A., and Carreau, S. (2004). Analysis and Significance of mRNA in Human Ejaculated Sperm from Normozoospermic Donors: Relationship to Sperm Motility and Capacitation. *Mol. Hum. Reprod.* 10, 535–541. doi:10.1093/molehr/gah064
- Légaré, C., Akintayo, A., Blondin, P., Calvo, E., and Sullivan, R. (2017). Impact of Male Fertility Status on the Transcriptome of the Bovine Epididymis. *Mol. Hum. Reprod.* 23, 355–369. doi:10.1093/molehr/gax019
- Li, B., He, X., Zhao, Y., Bai, D., Du, M., Song, L., et al. (2020). Transcriptome Profiling of Developing Testes and Spermatogenesis in the Mongolian Horse. *BMC Genet.* 21, 1–10. doi:10.1186/s12863-020-00843-5
- Li, C., and Zhou, X. (2012). Gene Transcripts in Spermatozoa: Markers of Male Infertility. *Clin. Chim. Acta* 413, 1035–1038. doi:10.1016/j.cca.2012.03.002
- Li, J., Liu, X., Hu, X., Tian, G. G., Ma, W., Pei, X., et al. (2017). MicroRNA-10b Regulates the Renewal of Spermatogonial Stem Cells through Kruppel-like Factor 4. *Cell Biochem. Funct.* 35, 184–191. doi:10.1002/cbf.3263
- Li, Y., Li, R.-H., Ran, M.-X., Zhang, Y., Liang, K., Ren, Y.-N., et al. (2018). High Throughput Small RNA and Transcriptome Sequencing Reveal Capacitation-Related microRNAs and mRNA in Boar Sperm. *BMC Genomics* 19, 736. doi:10.1186/s12864-018-5132-9
- Lian, J., Zhang, X., Tian, H., Liang, N., Wang, Y., Liang, C., et al. (2009). Altered microRNA Expression in Patients with Non-obstructive Azoospermia. *Reprod. Biol. Endocrinol.* 7, 1. doi:10.1186/1477-7827-7-13
- Lian, Y., Gódia, M., Castello, A., Rodríguez-Gil, J. E., Balasch, S., Sanchez, A., et al. (2021). Characterization of the Impact of Density Gradient Centrifugation on the Profile of the Pig Sperm Transcriptome by RNA-Seq. *Front. Vet. Sci.* 8, 1–12. doi:10.3389/fvets.2021.668158
- Liu, S.-S., Maguire, E. M., Bai, Y.-S., Huang, L., Liu, Y., Xu, L., et al. (2019). A Novel Regulatory Axis, CHD1L-MicroRNA 486-Matrix Metalloproteinase 2, Controls Spermatogonial Stem Cell Properties. *Mol. Cell. Biol.* 39. doi:10.1128/MCB.00357-18
- Liu, W.-M., Pang, R. T. K., Chiu, P. C. N., Wong, B. P. C., Lao, K., Lee, K.-F., et al. (2012). Sperm-borne microRNA-34c Is Required for the First Cleavage Division in Mouse. *Proc. Natl. Acad. Sci. U. S. A.* 109, 490–494. doi:10.1073/pnas.1110368109
- Lu, Y., Wu, X., and Wang, J. (2019). Correlation of miR-425-5p and IL-23 with Pancreatic Cancer. *Oncol. Lett.* 17, 4595–4599. doi:10.3892/ol.2019.10099
- Luo, J., McGinnis, L. K., Carlton, C., Beggs, H. E., and Kinsey, W. H. (2014). PTK2b Function during Fertilization of the Mouse Oocyte. *Biochem. Biophys. Res. Commun.* 450, 1212–1217. doi:10.1016/j.bbrc.2014.03.083
- Luo, L. F., Hou, C. C., and Yang, W. X. (2016). Small Non-coding RNAs and Their Associated Proteins in Spermatogenesis. *Gene* 578, 141–157. doi:10.1016/j.gene.2015.12.020
- Maciel, V. L., Caldas-Bussiere, M. C., Silveira, V., Reis, R. S., Rios, A. F. L., and Paes de Carvalho, C. S. (2018). L-arginine Alters the Proteome of Frozen-Thawed Bovine Sperm during *In Vitro* Capacitation. *Theriogenology* 119, 1–9. doi:10.1016/j.theriogenology.2018.06.018
- Martins, R. P., and Krawetz, S. A. (2005). RNA in Human Sperm. *Asian J. Androl.* 7, 115–120. doi:10.1111/j.1745-7262.2005.00048.x
- Marzano, G., Chiriacó, M., Primiceri, E., Dell’Aquila, M., Ramalho-Santos, J., Zara, V., et al. (2020). Sperm Selection in Assisted Reproduction: A Review of Established Methods and Cutting-Edge Possibilities. *Biotechnol. Adv.* 40, 107498. doi:10.1016/j.BIOTECHADV.2019.107498
- Michailov, Y., Ickowicz, D., and Breitbart, H. (2014). Zn2+-stimulation of Sperm Capacitation and of the Acrosome Reaction Is Mediated by EGFR Activation. *Dev. Biol.* 396, 246–255. doi:10.1016/j.ydbio.2014.10.009
- Miller, D., Briggs, D., Snowden, H., Hamlington, J., Rollinson, S., Lilford, R., et al. (1999). A Complex Population of RNAs Exists in Human Ejaculate Spermatozoa: Implications for Understanding Molecular Aspects of Spermiogenesis. *Gene* 237, 385–392. doi:10.1016/S0378-1119(99)00324-8
- Miller, D., Ostermeier, G. C., and Krawetz, S. A. (2005). The Controversy, Potential and Roles of Spermatozoal RNA. *Trends Mol. Med.* 11, 156–163. doi:10.1016/j.molmed.2005.02.006
- Miller, D., and Ostermeier, G. C. (2006). Towards a Better Understanding of RNA Carriage by Ejaculate Spermatozoa. *Hum. Reprod. Update* 12, 757–767. doi:10.1093/humupd/dml037
- Miller, D. (2014/2014). Sperm RNA as a Mediator of Genomic Plasticity. *Adv. Biol.* 1–13. doi:10.1155/2014/179701
- Miller, D. (2007). Spermatozoal RNA as Reservoir, Marker and Carrier of Epigenetic Information: Implications for Cloning. *Reprod. Domest. Anim.* 42, 2–9. doi:10.1111/j.1439-0531.2007.00883.x

- Mohamad, S. F. S., Ibrahim, S. F., Ismail, N. H., Osman, K., Jaafar, F. H. F., Nang, C. F., et al. (2018). Quantification of HSP70 Gene Expression and Determination of Capacitation Status of Magnetically Separated Cryopreserved Bovine Spermatozoa at Different Thawing Temperature and Time. *Sains Malaysiana* 47, 1101–1108. doi:10.17576/jsm-2018-4706-04
- Montjean, D., De La Grange, P., Gentien, D., Rapinat, A., Belloc, S., Cohen-Bacrie, P., et al. (2012). Sperm Transcriptome Profiling in Oligozoospermia. *J. Assist. Reprod. Genet.* 29, 3–10. doi:10.1007/s10815-011-9644-3
- Moraveji, S.-F., Esfandiari, F., Sharbatoghli, M., Taleahmad, S., Nikeghbalian, S., Shahverdi, A., et al. (2019). Optimizing Methods for Human Testicular Tissue Cryopreservation and Spermatogonial Stem Cell Isolation. *J. Cel. Biochem.* 120, 613–621. doi:10.1002/jcb.27419
- Morgan, H. L., Eid, N., Khoshkardar, A., and Watkins, A. J. (2020). Defining the Male Contribution to Embryo Quality and Offspring Health in Assisted Reproduction in Farm Animals. *Anim. Reprod.* 17, 1–14. doi:10.1590/1984-3143-AR2020-0018
- Morrell, J., Mari, G., Kútvolgyi, G., Meurling, S., Mislei, B., Iacono, E., et al. (2011). Pregnancies Following Artificial Insemination with Spermatozoa from Problem Stallion Ejaculates Processed by Single Layer Centrifugation with. *Androcoll-e. Reprod. Domest. Anim.* 46, 642–645. doi:10.1111/j.1439-0531.2010.01721.x
- Neto, F. T. L., Bach, P. V., Najari, B. B., Li, P. S., and Goldstein, M. (2016). Spermatogenesis in Humans and its Affecting Factors. *Semin. Cel Dev. Biol.* 59, 10–26. doi:10.1016/j.semcd.2016.04.009
- Nicolas, C. M., Accogli, G., Douet, C., Magistrini, M., Vern, Y. Le., Sausset, A., et al. (2020). Highlight of New Agents Inducing Capacitation-Related Changes in Stallion Spermatozoa to Cite This Version : HAL Id : Hal-02627722 Highlight of New Agents Inducing Capacitation-Related Changes in Stallion Spermatozoa. 2, 61–70.
- Niu, Z., Goodyear, S. M., Rao, S., Wu, X., Tobias, J. W., Avarbock, M. R., et al. (2011). MicroRNA-21 Regulates the Self-Renewal of Mouse Spermatogonial Stem Cells. *Proc. Natl. Acad. Sci.* 108, 12740–12745. doi:10.1073/pnas.1109987108
- Ostermeier, G. C., Dix, D. J., Miller, D., Khatri, P., and Krawetz, S. A. (2002). Spermatozoal RNA Profiles of normal fertile Men. *Lancet* 360, 772–777. doi:10.1016/S0140-6736(02)09899-9
- Ostermeier, G. C., Goodrich, R. J., Moldenhauer, J. S., Diamond, M. P., and Krawetz, S. A. (2005). A Suite of Novel Human Spermatozoal RNAs. *J. Androl.* 26, 70–74. doi:10.1002/j.1939-4640.2005.tb02874.x
- Ostermeier, G. C., Miller, D., Huntriss, J. D., Diamond, M. P., and Krawetz, S. A. (2004). Delivering Spermatozoan RNA to the Oocyte. *Nature* 429, 154. doi:10.1038/429154a
- Özbek, M., Hitit, M., Kaya, A., Jousan, F. D., and Memili, E. (2021). Sperm Functional Genome Associated with Bull Fertility. *Front. Vet. Sci.* 8, 1–17. doi:10.3389/fvets.2021.610888
- Pantano, L., Jodar, M., Bak, M., Ballešcà, J. L. Iuì, Tommerup, N., Oliva, R., et al. (2015). The Small RNA Content of Human Sperm Reveals Pseudogene-Derived piRNAs Complementary to Protein-Coding Genes. *RNA* 21, 1085–1095. doi:10.1261/rna.046482.114
- Paradowska-Dogan, A., Fernandez, A., Bergmann, M., Kretzer, K., Mallidis, C., Vieweg, M., et al. (2014). Protamine mRNA Ratio in Stallion Spermatozoa Correlates with Mare Fecundity. *Andrology* 2, 521–530. doi:10.1111/j.2047-2927.2014.00211.x
- Pardede, B. P., Agil, M., and Supriatna, I. (2020). Protamine and Other Proteins in Sperm and Seminal Plasma as Molecular Markers of Bull Fertility. *Vet. World* 13, 556–562. doi:10.14202/vetworld.2020.556-562
- Parthipan, S., Selvaraju, S., Somashekar, L., Arangasamy, A., Sivaram, M., and Ravindra, J. P. (2017). Spermatozoal Transcripts Expression Levels Are Predictive of Semen Quality and conception Rate in Bulls (*Bos taurus*). *Theriogenology* 98, 41–49. doi:10.1016/j.theriogenology.2017.04.042
- Peifer, M. (2000). Wnt Signaling in Oncogenesis and Embryogenesis-Aa Look outside the Nucleus. *Sci.* (80- 287, 1606–1609. doi:10.1126/science.287.5458.1606
- Peris-Frau, P., Álvarez-Rodríguez, M., Martín-Maestro, A., Iniesta-Cuerda, M., Sánchez-Ajofrín, I., Garde, J. J., et al. (2019). Comparative Evaluation of DNA Integrity Using Sperm Chromatin Structure Assay and Sperm-Ovis-Halomax during *In Vitro* Capacitation of Cryopreserved Ram Spermatozoa. *Reprod. Domest. Anim.* 54, 46–49. doi:10.1111/rda.13519
- Prakash, M. A., Kumaresan, A., Ebenezer Samuel King, J. P., Nag, P., Sharma, A., Sinha, M. K., et al. (2021). Comparative Transcriptomic Analysis of Spermatozoa from High- and Low-Fertile Crossbred Bulls: Implications for Fertility Prediction. *Front. Cel Dev. Biol.* 9, 1–14. doi:10.3389/fcell.2021.647717
- Prakash, M. A., Kumaresan, A., Sinha, M. K., Kamaraj, E., Mohanty, T. K., Datta, T. K., et al. (2020). RNA-seq Analysis Reveals Functionally Relevant Coding and Non-coding RNAs in Crossbred Bull Spermatozoa. *Anim. Reprod. Sci.* 222, 1–25. doi:10.1016/j.anireprosci.2020.106621
- Qiu, T., Wang, K., Li, X., and Jin, J. (2018). miR-671-5p Inhibits Gastric Cancer Cell Proliferation and Promotes Cell Apoptosis by Targeting URGCP. *Exp. Ther. Med.* 16, 4753–4758. doi:10.3892/etm.2018.6813
- Raval, N. P., Shah, T. M., George, L. B., and Joshi, C. G. (2019). Insight into Bovine (*Bos indicus*) Spermatozoal Whole Transcriptome Profile. *Theriogenology* 129, 8–13. doi:10.1016/j.theriogenology.2019.01.037
- Rawe, V. Y., Payne, C., and Schatten, G. (2006). Profilin and Actin-Related Proteins Regulate Microfilament Dynamics during Early Mammalian Embryogenesis. *Hum. Reprod.* 21, 1143–1153. doi:10.1093/humrep/dei480
- Rösok, O., Pedetour, F., Ree, A. H., and Aasheim, H. C. (1999). Identification and Characterization of TESK2, a Novel Member of the LIMK/TESK Family of Protein Kinases, Predominantly Expressed in Testis. *Genomics* 61, 44–54. doi:10.1006/geno.1999.5922
- Saacke, R. G., Dalton, J. C., Nadir, S., Nebel, R. L., and Bame, J. H. (2000). Relationship of Seminal Traits and Insemination Time to Fertilization Rate and Embryo Quality. *Anim. Reprod. Sci.* 60–61, 663–677. doi:10.1016/s0378-4320(00)00137-8
- Sakurai, K., Mikamoto, K., Shirai, M., Iguchi, T., Ito, K., Takasaki, W., et al. (2016). MicroRNA Profiles in a Monkey Testicular Injury Model Induced by Testicular Hyperthermia. *J. Appl. Toxicol.* 36, 1614–1621. doi:10.1002/jat.3326
- Schwab, K. R., Smith, G. D., and Dressler, G. R. (2013). Arrested Spermatogenesis and Evidence for DNA Damage in PTIP Mutant Testes. *Dev. Biol.* 373, 64–71. doi:10.1016/j.ydbio.2012.10.006
- Sellem, E., Marthey, S., Rau, A., Jouneau, L., Bonnet, A., Perrier, J.-P., et al. (2020). A Comprehensive Overview of Bull Sperm-Borne Small Non-coding RNAs and Their Diversity across Breeds. *Epigenetics Chromatin* 13, 19. doi:10.1186/s13072-020-00340-0
- Selvaraju, S., Parthipan, S., Somashekar, L., Binsila, B. K., Kolte, A. P., Arangasamy, A., et al. (2018). Current Status of Sperm Functional Genomics and its Diagnostic Potential of Fertility in Bovine (*Bos taurus*). *Syst. Biol. Reprod. Med.* 64, 484–501. doi:10.1080/19396368.2018.1444816
- Selvaraju, S., Parthipan, S., Somashekar, L., Kolte, A. P., Krishnan Binsila, B., Arangasamy, A., et al. (2017). Occurrence and Functional Significance of the Transcriptome in Bovine (*Bos taurus*) Spermatozoa. *Sci. Rep.* 7, 42392. doi:10.1038/srep42392
- Sendler, E., Johnson, G. D., Mao, S., Goodrich, R. J., Diamond, M. P., Hauser, R., et al. (2013). Stability, Delivery and Functions of Human Sperm RNAs at Fertilization. *Nucleic Acids Res.* 41, 4104–4117. doi:10.1093/nar/gkt132
- Shilpa, M., Selvaraju, S., GirishKumar, V., Parthipan, S., Binsila, K. B., Arangasamy, A., et al. (2017). Novel Insights into the Role of Cell-free Seminal mRNAs on Semen Quality and Cryotolerance of Spermatozoa in Bulls (*Bos taurus*). *Reprod. Fertil. Dev.* 29, 2446. doi:10.1071/RD16290
- Shin, J. Y., Gupta, M. K., Jung, Y. H., Uhm, S. J., and Lee, H. T. (2011). Differential Genomic Imprinting and Expression of Imprinted microRNAs in Testes-Derived Male Germ-Line Stem Cells in Mouse. *PLoS One* 6, e22481. doi:10.1371/journal.pone.0022481
- Shirakata, Y., Hiradate, Y., Inoue, H., Sato, E., and Tanemura, K. (2014). Histone H4 Modification during Mouse Spermatogenesis. *J. Reprod. Dev.* 60, 383–387. doi:10.1262/jrd.2014-018
- Singh, R., Junghare, V., Hazra, S., Singh, U., Sengar, G. S., Raja, T. V., et al. (2019). Database on Spermatozoa Transcriptogram of Catagorised Frieswal Crossbred (Holstein Friesian X Sahiwal) Bulls. *Theriogenology* 129, 130–145. doi:10.1016/j.theriogenology.2019.01.025
- Sosnik, J., Miranda, P. V., Spiridonov, N. A., Yoon, S. Y., Fissore, R. A., Johnson, G. R., et al. (2009). Tssk6 Is Required for Izumo Relocalization and Gamete Fusion in the Mouse. *J. Cel Sci.* 122, 2741–2749. doi:10.1242/jcs.047225
- Stephanie, P. (2016). Characterization of Small Non-coding RNAs in the Seminal Plasma of Beef Bulls with Predicted High and Low Fertility. Electron Theses Diss. Brookings (SD): South Dakota State University.

- Stiavnicka, M., Alvarez, O. G., Nevoral, J., KraliAkova, M., and Sutovsky, P. (2016). Key Features of Genomic Imprinting during Mammalian Spermatogenesis: Perspectives for Human Assisted Reproductive Therapy: A Review. *Anat. Physiol.* 6. doi:10.4172/2161-0940.1000236
- Stowe, H. M., Calcaterra, S. M., Dimmick, M. A., Andrae, J. G., Duckett, S. K., and Pratt, S. L. (2014). The Bull Sperm microRNAome and the Effect of Fescue Toxicosis on Sperm microRNA Expression. *PLoS One* 9, 1–18. doi:10.1371/journal.pone.0113163
- Suh, N., and Billelloch, R. (2011). Small RNAs in Early Mammalian Development: From Gametes to Gastrulation. *Development* 138, 1653–1661. doi:10.1242/dev.056234
- Suliman, Y., Becker, F., and Wimmers, K., (2018). Implication of Transcriptome Profiling of Spermatozoa for Stallion Fertility. *Reprod. Fertil. Dev.* 30, 1087–1098. doi:10.1071/RD17188
- Sun, Y. H., Wang, A., Song, C., Shankar, G., Srivastava, R. K., Au, K. F., et al. (2021). Single-molecule Long-Read Sequencing Reveals a Conserved Intact Long RNA Profile in Sperm. *Nat. Commun.* 12, 1361. doi:10.1038/s41467-021-21524-6
- Tan, D., Zhou, C., Han, S., Hou, X., Kang, S., and Zhang, Y. (2018). MicroRNA-378 Enhances Migration and Invasion in Cervical Cancer by Directly Targeting Autophagy-Related Protein 12. *Mol. Med. Rep.* 17, 6319–6326. doi:10.3892/mmr.2018.8645
- Tong, M.-H., Mitchell, D. A., McGowan, S. D., Evanoff, R., and Griswold, M. D. (2012). Two miRNA Clusters, Mir-17-92 (Mirc1) and Mir-106b-25 (Mirc3), Are Involved in the Regulation of Spermatogonial Differentiation in Mice. *Biol. Reprod.* 86. doi:10.1095/biolreprod.111.096313
- Töpfer-Petersen, E., Eklashi-Hundrieser, M., Kirchhoff, C., Leeb, T., and Sieme, H. (2005). The Role of Stallion Seminal Proteins in Fertilisation. *Anim. Reprod. Sci.* 89, 159–170. doi:10.1016/j.anireprosci.2005.06.018
- Turner, T. T., Bang, H. J., Attipoe, S. A., Johnston, D. S., and Tomsig, J. L. (2006). Sonic Hedgehog Pathway Inhibition Alters Epididymal Function as Assessed by the Development of Sperm Motility. *J. Androl.* 27, 225–232. doi:10.2164/jandrol.05114
- Turri, F., Capra, E., Lazzari, B., Cremonesi, P., Stella, A., and Pizzi, F. (2021). A Combined Flow Cytometric Semen Analysis and miRNA Profiling as a Tool to Discriminate between High- and Low-Fertility Bulls. *Front. Vet. Sci.* 8, 1–12. doi:10.3389/fvets.2021.703101
- Varner, D. D., Bowen, J. A., and Johnson, L. (1993). Effect of Heparin on Capacitation/acrosome Reaction of Equine Sperm. *Syst. Biol. Reprod. Med.* 31, 199–207. doi:10.3109/01485019308988400
- Varona, L., Clop, A., Gòdia, M., Estill, M., Castelló, A., Balasch, S., et al. (2019). A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and its Seasonal Changes. *Front. Genet.* 10, 1–14. doi:10.3389/fgene.2019.00299
- Vickram, A. S., Srikumar, P. S., Srinivasan, S., Jeyanthi, P., Anbarasu, K., Thanigaivel, S., et al. (2021). Seminal Exosomes – an Important Biological Marker for Various Disorders and Syndrome in Human Reproduction. *Saudi J. Biol. Sci.* 28, 3607–3615. doi:10.1016/j.sjbs.2021.03.038
- Vojtech, L., Woo, S., Hughes, S., Levy, C., Ballweber, L., Sauteraud, R. P., et al. (2014). Exosomes in Human Semen Carry a Distinctive Repertoire of Small Non-coding RNAs with Potential Regulatory Functions. *Nucleic Acids Res.* 42, 7290. doi:10.1093/NAR/GKU347
- Wallrapp, C., Hähnel, S., Boeck, W., Soder, A., Mincheva, A., Lichter, P., et al. (2001). Loss of the Y Chromosome Is a Frequent Chromosomal Imbalance in Pancreatic Cancer and Allows Differentiation to Chronic Pancreatitis. *Int. J. Cancer* 91, 340–344. doi:10.1002/1097-0215(200002)9999:9999<aid-ijc1014>3.0.co;2-u
- Wang, H., Zhou, Z., Xu, M., Li, J., Xiao, J., Xu, Z. Y., et al. (2004). A Spermatogenesis-Related Gene Expression Profile in Human Spermatozoa and its Potential Clinical Applications. *J. Mol. Med.* 82, 317–324. doi:10.1007/s00109-004-0526-3
- Wang, T., Gao, H., Li, W., and Liu, C. (2019a). Essential Role of Histone Replacement and Modifications in Male Fertility. *Front. Genet.* 10, 1–15. doi:10.3389/fgene.2019.00962
- Wang, Y., Li, X., Gong, X., Zhao, Y., and Wu, J. (2019b). MicroRNA-322 Regulates Self-Renewal of Mouse Spermatogonial Stem Cells through Rassf8. *Int. J. Biol. Sci.* 15, 857–869. doi:10.7150/ijbs.30611
- Wang, Y., Zhou, Y., Ali, M. A., Zhang, J., Wang, W., Huang, Y., et al. (2021). Comparative Analysis of piRNA Profiles Helps to Elucidate Cryoinjury between Giant Panda and Boar Sperm during Cryopreservation. *Front. Vet. Sci.* 8, 1–11. doi:10.3389/fvets.2021.635013
- Ward, W. S. (2009). Function of Sperm Chromatin Structural Elements in Fertilization and Development. *Mol. Hum. Reprod.* 16, 30–36. doi:10.1093/molehr/gap080
- Watson, C. N., Belli, A., and Di Pietro, V. (2019). Small Non-coding RNAs: New Class of Biomarkers and Potential Therapeutic Targets in Neurodegenerative Disease. *Front. Genet.* 10, 1–14. doi:10.3389/fgene.2019.00364
- Wei, X., Moncada-Pazos, A., Cal, S., Soria-Valles, C., Gartner, J., Rudloff, U., et al. (2011). Analysis of the Disintegrin-Metalloproteinases Family Reveals ADAM29 and ADAM7 Are Often Mutated in Melanoma. *Hum. Mutat.* 32, E2148–E2175. doi:10.1002/humu.21477
- Wu, W., Hu, Z., Qin, Y., Dong, J., Dai, J., Lu, C., et al. (2012). Seminal Plasma microRNAs: Potential Biomarkers for Spermatogenesis Status. *MHR Basic Sci. Reprod. Med.* 18, 489–497. doi:10.1093/molehr/gas022
- Wu, Y., Xu, D., Zhu, X., and Ren*, G. Y. (2017). MiR-106a Associated with Diabetic Peripheral Neuropathy through the Regulation of 12/15-LOX-Mediated Oxidative/Nitritative Stress. *Curr. Neurovasc. Res.* 14, 117–124. doi:10.2174/1567202614666170404115912
- Wykes, S. M., Visscher, D. W., and Krawetz, S. A. (1997). Haploid Transcripts Persist in Mature Human Spermatozoa. *Mol. Hum. Reprod.* 3, 15–19. doi:10.1093/molehr/3.1.15
- Xiong, S., Li, Y., Xiang, Y., Peng, N., Shen, C., Cai, Y., et al. (2019). Dysregulation of lncRNA and circRNA Expression in Mouse Testes after Exposure to Triptolide. *Curr. Drug Metab.* 20, 665–673. doi:10.2174/1389200220666190729130020
- Xu, X., Tan, Y., Mao, H., Liu, H., Dong, X., and Yin, Z. (2020a). Analysis of Long Noncoding RNA and mRNA Expression Profiles of Testes with High and Low Sperm Motility in Domestic Pigeons (*Columba livia*). *Genes (Basel)* 11, 349. doi:10.3390/genes11040349
- Xu, Z., Xie, Y., Zhou, C., Hu, Q., Gu, T., Yang, J., et al. (2020b). Expression Pattern of Seminal Plasma Extracellular Vesicle Small RNAs in Boar Semen. *Front. Vet. Sci.* 0, 929. doi:10.3389/FVETS.2020.585276
- Yang, C. C., Lin, Y. S., Hsu, C. C., Tsai, M. H., Wu, S. C., and Cheng, W. T. K. (2010). Seasonal Effect on Sperm Messenger RNA Profile of Domestic Swine (*Sus Scrofa*). *Anim. Reprod. Sci.* 119, 76–84. doi:10.1016/j.anireprosci.2009.12.002
- Yang, C. C., Lin, Y. S., Hsu, C. C., Wu, S. C., Lin, E. C., and Cheng, W. T. K. (2009). Identification and Sequencing of Remnant Messenger RNAs Found in Domestic Swine (*Sus scrofa*) Fresh Ejaculated Spermatozoa. *Anim. Reprod. Sci.* 113, 143–155. doi:10.1016/j.anireprosci.2008.08.012
- Yang, H.-M., Liu, G., Nie, Z.-Y., Nie, D.-S., Deng, Y., and Lu, G.-X. (2005). Molecular Cloning of a Novel Rat Gene Tsarg1, a Member of the DnaJ/HSP40 Protein Family. *DNA Seq.* 16, 166–172. doi:10.1080/10425170500129736
- Yatsenko, A. N., Georgiadis, A. P., Murthy, L. J., Lamb, D. J., and Matzuk, M. M. (2013). UBE2B mRNA Alterations Are Associated with Severe Oligozoospermia in Infertile Men. *Mol. Hum. Reprod.* 19, 388–394. doi:10.1093/molehr/gat008
- Yatsenko, A. N., O'Neil, D. S., Roy, A., Arias-Mendoza, P. A., Chen, R., Murthy, L. J., et al. (2012). Association of Mutations in the Zona Pellucida Binding Protein 1 (ZBP1) Gene with Abnormal Sperm Head Morphology in Infertile Men. *MHR Basic Sci. Reprod. Med.* 18, 14–21. doi:10.1093/molehr/gar057
- Yatsenko, A. N., Roy, A., Chen, R., Ma, L., Murthy, L. J., Yan, W., et al. (2006). Non-invasive Genetic Diagnosis of Male Infertility Using Spermatozoal RNA: KLHL10 mutations in Oligozoospermic Patients Impair Homodimerization. *Hum. Mol. Genet.* 15, 3411–3419. doi:10.1093/hmg/ddl417
- Zhai, Z., Wu, F., Chuang, A. Y., and Kwon, J. H. (2013). miR-106b Fine Tunes ATG16L1 Expression and Autophagic Activity in Intestinal Epithelial HCT116 Cells. *Inflamm. Bowel Dis.* 19, 2295–2301. doi:10.1097/MIB.0b013e31829e71cf
- Zhang, J., Zhang, X., Luo, H., Wang, X., Cao, Z., and Zhang, Y. (2021). Expression Analysis of Circular RNAs in Young and Sexually Mature Boar Testes. *Animals* 11, 1430. doi:10.3390/ani11051430
- Zhang, W., Yi, K., Chen, C., Hou, X., and Zhou, X. (2012). Application of Antioxidants and Centrifugation for Cryopreservation of Boar Spermatozoa. *Anim. Reprod. Sci.* 132, 123–128. doi:10.1016/j.anireprosci.2012.05.009
- Zhang, X., Gao, F., Fu, J., Zhang, P., Wang, Y., and Zeng, X. (2017). Systematic Identification and Characterization of Long Non-coding RNAs in Mouse Mature Sperm. *PLoS One* 12, e0173402. doi:10.1371/journal.pone.0173402

- Zhang, X., Wu, M., Chong, Q.-Y., Zhang, W., Qian, P., Yan, H., et al. (2018). Amplification of Hsa-miR-191/425 Locus Promotes Breast Cancer Proliferation and Metastasis by Targeting DICER1. *Carcinogenesis* 39, 1506–1516. doi:10.1093/carcin/bgy102
- Zhang, Y., Dai, D., Chang, Y., Li, Y., Zhang, M., Zhou, G., et al. (2017a). Cryopreservation of Boar Sperm Induces Differential microRNAs Expression. *Cryobiology* 76, 24–33. doi:10.1016/j.cryobiol.2017.04.013
- Zhao, W., Ahmed, S. S., Ahmed, S. S., Yangliu, Y., Wang, H., and Cai, X. (2021). Analysis of Long Non-coding RNAs in Epididymis of Cattleyak Associated with Male Infertility. *Theriogenology* 160, 61–71. doi:10.1016/j.theriogenology.2020.10.033
- Zhou, F., Chen, W., Jiang, Y., and He, Z. (2019). Regulation of Long Non-coding RNAs and Circular RNAs in Spermatogonial Stem Cells. *Reproduction* 158, R15–R25. doi:10.1530/REP-18-0517
- Ziyyat, A. (2001). Differential Gene Expression in Pre-implantation Embryos from Mouse Oocytes Injected with Round Spermatids or Spermatozoa. *Hum. Reprod.* 16, 1449–1456. doi:10.1093/humrep/16.7.1449

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Sahoo, Choudhary, Sharma, Choudhary and Gupta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Comprehensive Analysis of mRNA, lncRNA, circRNA, and miRNA Expression Profiles and Their ceRNA Networks in the *Longissimus Dorsi* Muscle of Cattle-Yak and Yak

Chun Huang¹, Fei Ge¹, Xiaoming Ma¹, Rongfeng Dai¹, Renqing Dingkao², Zhuoma Zhaxi³, Getu Burechao³, Pengjia Bao¹, Xiaoyun Wu¹, Xian Guo¹, Min Chu¹, Ping Yan^{1*} and Chunnian Liang^{1*}

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Ran Di,
Institute of Animal Sciences (CAAS),
China
Ran Li,
Northwest A and F University, China

*Correspondence:

Ping Yan
pingyanlz@163.com
Chunnian Liang
chunnian2006@163.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 08 September 2021

Accepted: 15 November 2021

Published: 13 December 2021

Citation:

Huang C, Ge F, Ma X, Dai R,
Dingkao R, Zhaxi Z, Burechao G,
Bao P, Wu X, Guo X, Chu M, Yan P and
Liang C (2021) Comprehensive
Analysis of mRNA, lncRNA, circRNA,
and miRNA Expression Profiles and
Their ceRNA Networks in the
Longissimus Dorsi Muscle of Cattle-
Yak and Yak.
Front. Genet. 12:772557.
doi: 10.3389/fgene.2021.772557

¹Key Laboratory of Yak Breeding Engineering Gansu Province, Lanzhou Institute of Husbandry and Pharmaceutical Science, Chinese Academy of Agricultural Sciences, Lanzhou, China, ²Livestock Institute of Gannan Tibetan Autonomous Prefecture, Hezuo, China, ³Haixi Agricultural and Animal Husbandry Technology Extension Service Center, Qinghai, China

Cattle-yak, as the hybrid offspring of cattle (*Bos taurus*) and yak (*Bos grunniens*), demonstrates obvious heterosis in production performance. Male hybrid sterility has been focused on for a long time; however, the mRNAs and non-coding RNAs related to muscle development as well as their regulatory networks remain unclear. The phenotypic data showed that the production performance (i.e., body weight, withers height, body length, and chest girth) of cattle-yak was significantly better than that of the yak, and the economic benefits of the cattle-yak were higher under the same feeding conditions. Then, we detected the expression profiles of the *longissimus dorsi* muscle of cattle-yak and yak to systematically reveal the molecular basis using the high-throughput sequencing technology. Here, 7,126 mRNAs, 791 lncRNAs, and 1,057 circRNAs were identified to be differentially expressed between cattle-yaks and yaks in the *longissimus dorsi* muscle. These mRNAs, lncRNA targeted genes, and circRNA host genes were significantly enriched in myoblast differentiation and some signaling pathways related to muscle development (such as HIF-1 signaling pathway and PI3K-Akt signaling pathway). We constructed a competing endogenous RNA (ceRNA) network and found that some non-coding RNAs differentially expressed may be involved in the regulation of muscle traits. Taken together, this study may be used as a reference tool to provide the molecular basis for studying muscle development.

Keywords: cattle-yak, *Bos grunniens*, transcriptome, ceRNA, lncRNA, circRNA, skeletal muscle

INTRODUCTION

Yak, a special germplasm resource mainly inhabiting the Qinghai-Tibet Plateau, has been optimized for living and a source of living for the local herdsmen. The cattle-yak constitutes the hybrid of cattle (*Bos taurus*) and yak (*Bos grunniens*) exhibiting outstanding hybrid vigor in growth rate, meat performance, plateau adaptability, etc. The meat of cattle-yak is highly enriched in protein but has lower fat than yak, meeting the requirements of a popular healthy and

high-quality diet (Song et al., 2019; Wang et al., 2021b). Inhabiting the high-altitude environment, cattle-yak provides a natural green food favored by the consumers. The crossbreeding technology has been widely applied in animal breeding, as well as these hybrid individuals have been farmed for a long time forming gradually new indigenous breeds. Therefore, it is highly significant to understand the mechanisms regulating muscle growth and development of the generated crossbreed. The *longissimus dorsi* (LD) muscle is one of the important representatives of muscle tissues in animals, which is closely related to the individual skeletal muscle growth and development, as well as with the intramuscular fat content, and tenderness. A series of reported myogenic regulatory molecules regulate the myoblast's proliferation and myophagism, further influencing the growth process, including the identified genes *MYOG* (Rudnicki and Jaenisch, 1995) and *MYF5* (Xu et al., 2019), which are involved in regulating the myoblast differentiation, muscle growth, and meat quality traits in livestock. Previous studies have indicated that *MYOD* as a key regulator of myotube formation promotes the myotube's differentiation (Tapscott, 2005; Wang et al., 2017). The *CAPN* family are important candidate genes in the growth and degradation of the muscle fibers, and are specifically expressed in the skeletal muscle (Gandolfi et al., 2011).

In fact, the studies indicated that the biological processes were not only regulated by protein-coding RNA—mRNA, as the sequencing technology developed. Non-coding RNAs (ncRNAs), including long non-coding RNA (lncRNA), circular RNA (circRNA), and microRNA (miRNA), are profoundly involved in diverse biological processes and are regulating them by various mechanisms. The lncRNAs universally acknowledged participating in chromatin transcriptional/epigenetic regulation by interacting with the chromatin regulators as “molecular scaffold” or decoys to activate or repress transcription (Caretti et al., 2006; Korostowski et al., 2012). Many lncRNAs have been proven to play a vital role in skeletal muscle development; for example, the lncRNA *MAR1* positively correlates with muscle differentiation and growth *in vitro* and *in vivo* (Zhang et al., 2018). The lnc-*smad7*/miR-125b/*Smad7* (*SMAD* family member 7) and *IGF2* axes are instrumental in myoblast differentiation and regeneration of muscle in two different pig breeds (Song et al., 2018). In addition, continuously growing discoveries have reported the role of novel circRNA in skeletal muscle. CircZfp609 derived from Zinc Finger Protein 609, can inhibit the myogenic differentiation *via* the sponge miR-194-5p in the mouse myoblast cell line C2C12 (Wang et al., 2019b). The circFGFR4 was generated from the fibroblast growth factor receptor 4 (*FGFR4*) and could simulate the bovine primary myoblast's differentiation through the circFGFR4-miR-107-*WNT3A* axis in cattle (Li et al., 2018). MicroRNA response elements are considered to be “talking mediators” of mRNAs, lncRNAs, and transcribed pseudogenes (Ji et al., 2019), and these response elements have important roles in various biological processes by forming a large number of complex regulatory networks. Numerous reports on the conjoint effect of miRNA

and mRNA conjointly on developing skeletal muscle have been proved. *FGFR1*, which could prevent muscle fibrogenesis, is a functional target of miR-214-3p (Arrighi et al., 2021). The miR-183 and miR-96 were found to negatively regulate fat usage in the skeletal muscle *via* targeting *FoxO1* and *PDK4* (Wang et al., 2021a). Zhang et al. (2021) experimentally confirmed that miR-22-3p regulated the *WFIKK2* gene in adipocyte differentiation in muscle fat metabolism of Yanbian cattle. A targeted relationship between the oar-miR-655-3p and oar-miR-381-5p with *ACSM3* and *ABAT* has been found to have crucial roles in sheep muscle organogenesis, myoblast migration (Sun et al., 2019).

In recent years, with the deepening of the research on the function of miRNAs, a new theory named competing endogenous RNA (ceRNA) has emerged. At the same time, some studies reported that mRNAs, lncRNAs, and circRNAs might regulate the gene function *via* miRNA and act as ceRNAs in various biological processes (Salmena et al., 2011; Yu et al., 2019). In the whole transcriptome, a comprehensive post-transcriptional regulatory network formed by ceRNA activity has greatly widened the cognition of functional genetic information in the genome. There are effective interactions among the lncRNA, circRNA, and mRNA with miRNA, and they can take significant effect in various processes of regulation in animals. lncRNAs act as molecular sponges for the miRNAs that specifically inhibit the target mRNAs so that they can give play to the protection of mRNAs (Li et al., 2019). For instance, a previous study reported that *MAML1* and *MEF2C*, as transcription factors, activate the late-differentiation muscle genes, and linc-MD1 can regulate their expression as a ceRNA by sponging miR-133 and miR-135 (Cesana et al., 2011). lncRNA H19 can regulate muscle differentiation as a molecular sponge for the *LET7* family in the developing embryo and adult muscles (Kallen et al., 2013).

Until now, most studies have been based on focusing on the cattle-yak for exploring the male sterility mechanism and barely referred to the superiority in the growth mechanism. Here, we have measured the growth traits of cattle-yaks and domestic Ashidan yaks under the same feeding and management, and systematically explored the differences of the *longissimus dorsi* muscles for the first time using the whole-transcriptome sequencing. Furthermore, the ceRNA network was constructed to identify the key factors involved in muscle growth and development. This study will thus help in improving yak breeding and provide new ideas for studying the genetic mechanism of muscle growth.

MATERIALS AND METHODS

Ethics Approval

All the animal experiments were approved by Lanzhou Institute of Husbandry and Pharmaceutical Sciences of the Chinese Academy of Agricultural Sciences (CAAS) with the grant number: No. 2019-002. All the slaughter as well as sampling procedures strictly complied with the Guidelines on the Ethical Treatment of Experimental Animals of China.

Phenotypic Data Collection and Samples Preparation

Thirty cattle-yaks (Aberdeen Angus ♂ × Yak ♀) and 30 yaks were tracked to measure the production performance indices (withers height, body weight, chest girth, and body length) at three stages of growth (i.e., birth, 3 months, and 6 months). Cattle-yaks ($n = 3$, 6 months old) and yaks ($n = 3$, 6 months old) were selected randomly to be slaughtered for *longissimus dorsi* muscle. These samples were collected for transcriptome sequencing and Real-time quantitative polymerase chain reaction analysis (RT-qPCR). All the samples were stored in liquid nitrogen (-80°C) for the subsequent tests.

RNA Isolation and Illumina Sequencing

The total RNA was isolated with TRIzol (Invitrogen, Carlsbad, CA, United States) following the manufacturer's instructions, and the concentration and quality of RNA were assessed by 1.5% agarose gel electrophoresis and Thermo Scientific NanoDrop 2000c (ThermoFisher Scientific Inc., Waltham, MA, United States).

Equal quantities of RNA were pooled from each sample. Then, the TruSeq Stranded Total RNA with Ribo-Zero Gold Kit (Illumina, San Diego, CA, United States) was used for digesting the ribosomal RNA (rRNA) in the DNA-free RNA. According to the manufacturer's instructions, we performed the construction of library preparation with NEB Next Ultra Directional RNA LibraryPrep Kit for Illumina (NEB, Ipswich, MA, United States). The size and purity of libraries were validated by Agilent Technologies 2100 Bioanalyzer (Agilent, Santa Clara, CA). Finally, the samples were sequenced using Illumina HiSeq 2500 Technology (LC Sciences, Houston, TX, United States) with a 150-bp paired-end run.

Data Preprocessing, Read Mapping, and Transcript Assembly

Raw reads generated during high-throughput sequencing were in fastq format. The raw sequencing dataset supporting the results of cattle-yaks in this study was deposited at NCBI's Gene Expression Omnibus database (<http://www.ncbi.nlm.nih.gov/geo/>, accession number: PRJNA753699). The raw sequencing dataset of yaks has been uploaded to the NCBI Database in the previous study (Ma et al., 2020). To get high-quality clean reads that could be used for later analysis, the Trimmomatic software (Bolger et al., 2014) was applied to remove adaptors, low-quality bases, and N-bases.

The remaining high-quality cleaned reads of each sample were aligned to the yak reference genome (BosGru_v2.0) using the HiSAT2 software (Kim et al., 2015). All the samples were assessed by genomic and gene alignment. After obtaining the comparison result bam file, the reads on the comparison gene were assembled using the StringTie software (Pertea et al., 2015), and every single transcript assembled by each sample was fused and spliced into a merged transcript.

LncRNAs Identification and LncRNAs Target Gene Prediction

We performed the following steps to obtain the potential lncRNA candidates for subsequent analysis: (1) The Cuffcompare software

(Ghosh and Chan, 2016) was used to compare the merged transcripts with the reference transcripts one by one, and the transcripts marked with “i,” “u,” “x,” and “o” were retained after clarifying the position class of the remaining transcripts. (2) Transcripts with length >200 nt and exon number ≥ 2 were obtained. (3) The above transcripts were analyzed about the coding ability by CPC2 (Kang et al., 2017), CNCI (Sun et al., 2013), PLEK (Li et al., 2014), and Pfam (Sonnhammer et al., 1998) software to remove the transcripts with coding potential and obtain potential lncRNA candidates.

Cis-acting and *trans*-acting modes are two main ways to predict the targets of lncRNAs (Mercer et al., 2009). We calculated the locations of the paired lncRNAs and mRNAs for the *cis*-acting prediction. The lncRNA with no nearest protein-coding gene within 100 kb upstream or downstream was excluded in subsequent analysis. For *trans*-acting regulatory mode, the LncTar was used to calculate the free energy between them to predict the regulatory targets, since the expression of lncRNA is determined to be independent of the location of mRNA.

CircRNA Identification

We used the CIRI software and predicted the circRNA based on the BWA software (Li and Durbin, 2009; Gao et al., 2015). Since it is an authoritative software, it has the characteristics of high sensitivity and multiple screening for reducing false positives. Firstly, we aligned the clean reads to the reference genome to obtain the SAM file using the BWA software. Then, the CIRI software was used to scan for PCC signals (paired chiasmic clipping signals), and circRNA sequences were predicted based on junction reads and GT-AG cleavage signals. The expressional levels of circRNAs were quantified by the RPM algorithm.

Differential Expression Genes and Pathway Analysis

The expression levels of the mRNAs and lncRNAs were calculated through the fragments per kilobase of transcript per million reads mapped (FPKM) value (Liu et al., 2018) using the Cuffdiff program. The DESeq software (Anders and Huber, 2010) was used to standardize the counts of each sample and calculate the fold change (FC). A negative binomial distribution test (NB) was used to test the difference significance of counts. The differentially expressed mRNAs (DEMs), lncRNAs (DELs), and circRNAs (DECs) were screened according to the results of $|\log_2 \text{FC}| > 1$ and $p < 0.05$ eventually.

To further understand gene function, Gene Ontology (GO, <http://www.geneontology.org>) terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG, <https://www.genome.jp/kegg/>) were used for enrichment analysis of functional pathways. Each GO and KEGG enrichment term was confirmed by Hypergeometric Distribution Test. Then, the p -value was corrected by Benjamini and Hochberg multiple tests. The enrichment with p -value lower than 0.05 was considered significant.

Construction of ceRNA Network

To acquire a better understanding of the interactions of the mRNAs, lncRNAs, circRNAs, and miRNAs, a

TABLE 1 | The production performance between cattle-yaks and yaks.

Items	Birth (Mean ± SE)			3 months (Mean ± SE)			6 months (Mean ± SE)		
	Yaks	Cattle-yaks	p-Value	Yaks	Cattle-yaks	p-Value	Yaks	Cattle-yaks	p-Value
Body weight (kg)	16.37 ± 0.23	20.70 ± 0.41	<0.001	37.56 ± 0.51	48.21 ± 0.62	<0.001	81.40 ± 1.31	90.76 ± 0.81	<0.001
Body height (cm)	57.23 ± 0.87	63.56 ± 0.56	<0.001	73.73 ± 0.98	84.80 ± 0.95	<0.001	85.43 ± 1.12	91.73 ± 0.86	<0.001
Body length (cm)	50.50 ± 1.10	59.40 ± 0.55	<0.001	75.63 ± 0.72	87.27 ± 0.78	<0.001	89.90 ± 0.77	98.07 ± 1.04	<0.001
Chest girth (cm)	58.63 ± 0.79	63.86 ± 0.54	<0.001	84.46 ± 0.60	97.23 ± 0.68	<0.001	112.83 ± 1.23	120.77 ± 1.20	<0.001

Note: SE: standard error.

lncRNA–circRNA–miRNA–mRNA regulatory network was constructed based on the ceRNA hypothesis (Salmena et al., 2011). MiRanda (John et al., 2004) was used to predict the pairs of miRNA–lncRNA, miRNA–mRNA, and miRNA–circRNA. The Spearman correlation coefficient (SCC) was used to evaluate the pairwise correlations of miRNA–lncRNA, miRNA–mRNA, and miRNA–circRNA; the value greater than 0.8 was considered relevant for constructing the network and $p < 0.05$ was regarded as being statistically significant. The Cytoscape software (version 3.5.1) was used to display the results visually.

Quantitative Real-Time PCR for Validating Gene Expression

The RNA for verifying the gene expression was the same as RNA used in the above Illumina sequencing. The RNA was reverse transcribed into cDNA by HiScript II 1st Strand cDNA Synthesis Kit (Vazyme, Nanjing, Jiangsu, China). Glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), as endogenous control, was used to normalize target gene expression. The total 20- μ l reaction system of qPCR included 10 ng of cDNA, 10 μ l of SYBR Premix Ex Taq II (TaKaRa, Dalian, China), each 1 μ l of forward primer and reverse primer, and 7 μ l of ddH₂O. The RT-qPCR conditions included the preincubation at 95°C for the 30 s, 45 cycles of 10 s at 95°C and 60 s at 59°C, then ended at 72°C for 30 s. Each experiment was repeated three times, and the results of relative RNA expression were calculated according to the cycle threshold (Ct) value $2^{-\Delta\Delta Ct}$ (Livak and Schmittgen, 2001).

RESULTS

Comparison of Production Performance

Analysis of the measurement results showed that the cattle-yaks had a significant improvement in production performance compared to the yaks in each age stage ($p < 0.001$). As can be seen from **Table 1**, cattle-yaks were stronger and taller at birth than yaks. In terms of body weight, at birth, the cattle-yak increased by about 22.60% compared to the yaks, and the body length, height, and chest girth of cattle-yaks were also increased by 17.62%, 11.06%, and 8.92%, respectively. From birth to 6 months old, the cattle-yak showed obvious heterosis with varying degrees of improvement in various indicators.

The Prediction and Characteristic of lncRNAs and CircRNAs

After filtering out the low-quality and supernumerary reads, we obtained a total of 310,089,232 and 308,742,564 clean reads with greater than 93.77% of Q30 from the LD muscles of cattle-yaks and yaks, respectively. As shown in **Supplementary Table S1**, the successful alignment of reads to the reference genome was approximately over 94.55%.

After rigorous screening and filtration, we detected a large number of lncRNAs and circRNAs (**Figures 1A,E**). Among them, 1,817 lncRNAs with an average length of 1,449 bp were discovered as novel lncRNAs (**Figure 1B**). The largest proportion of lncRNAs was over 2,000 bp (19.37%) and the proportion of lncRNA containing two exons was about 70.61% (**Figure 1C**). The total lncRNAs of 135 exonic antisense, 219 intronic antisense, 105 intergenic downstream antisense, 202 intergenic upstream, 204 exonic sense, 380 intronic sense, 141 intergenic downstream sense and 140 intergenic up-stream sense were identified in our results (**Figure 1D**). The average length of detected circRNAs was 3,250 bp and the length of the most circRNAs was over 2,000 bp (16.92%) (**Figure 1F**). As evident in **Figure 1E**, the circRNAs of sense-overlapping accounts for 89% of the total. The number of circRNAs located in the exonic and intronic was 332 and 137, respectively, while the antisense-overlapping circRNAs occupied about 1%.

Differentially Expressed mRNAs, lncRNAs, and circRNAs Between CY and Y Groups

The study identified 7,126 mRNAs, 791 lncRNAs, and 1,057 circRNAs to have significant differential expressions (**Supplementary Tables S2–S4**). There were 6902 DE mRNAs, 742 DE lncRNAs, and 273 DE circRNAs, respectively, which were detected in the cattle-yaks and yaks (**Figures 2A–C**). However, 119 mRNAs, 41 lncRNAs, and 232 circRNAs were detected to express only in the CY group. Other 105 mRNAs, 8 lncRNAs, and 552 circRNAs were identified only in the Y group. To find the overall distribution of differential expression, the volcano plot was drawn based on the results of the differential expression. Compared with the yaks of the control group, we found 3,563 upregulated mRNAs, 455 upregulated lncRNAs, and 353 upregulated circRNAs in the cattle-yaks. Meanwhile, there were 3,563 downregulated mRNAs, 336 downregulated lncRNAs, and 704 downregulated circRNAs, respectively, in the cattle-yaks (**Figures 2D–F**).

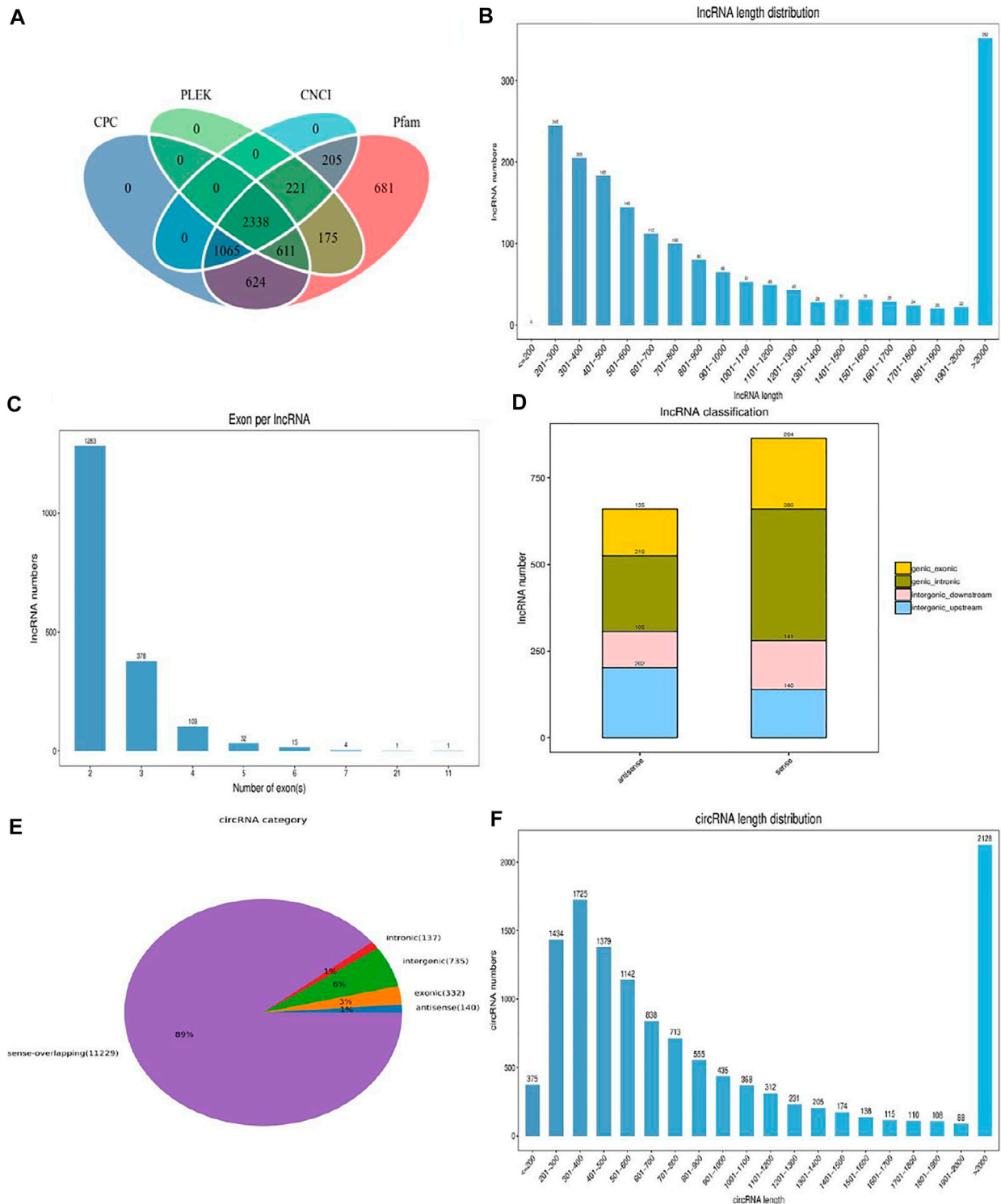


FIGURE 1 | The description of identified lncRNAs and circRNAs. **(A)** Screening of the candidate lncRNAs in *longissimus dorsi* muscle. **(B)** The length distribution of the novel lncRNAs. **(C)** Exon number distribution of novel lncRNAs. **(D)** The classification of novel lncRNAs. **(E)** The structure type pie chart of circRNAs. **(F)** The length distribution of circRNAs.

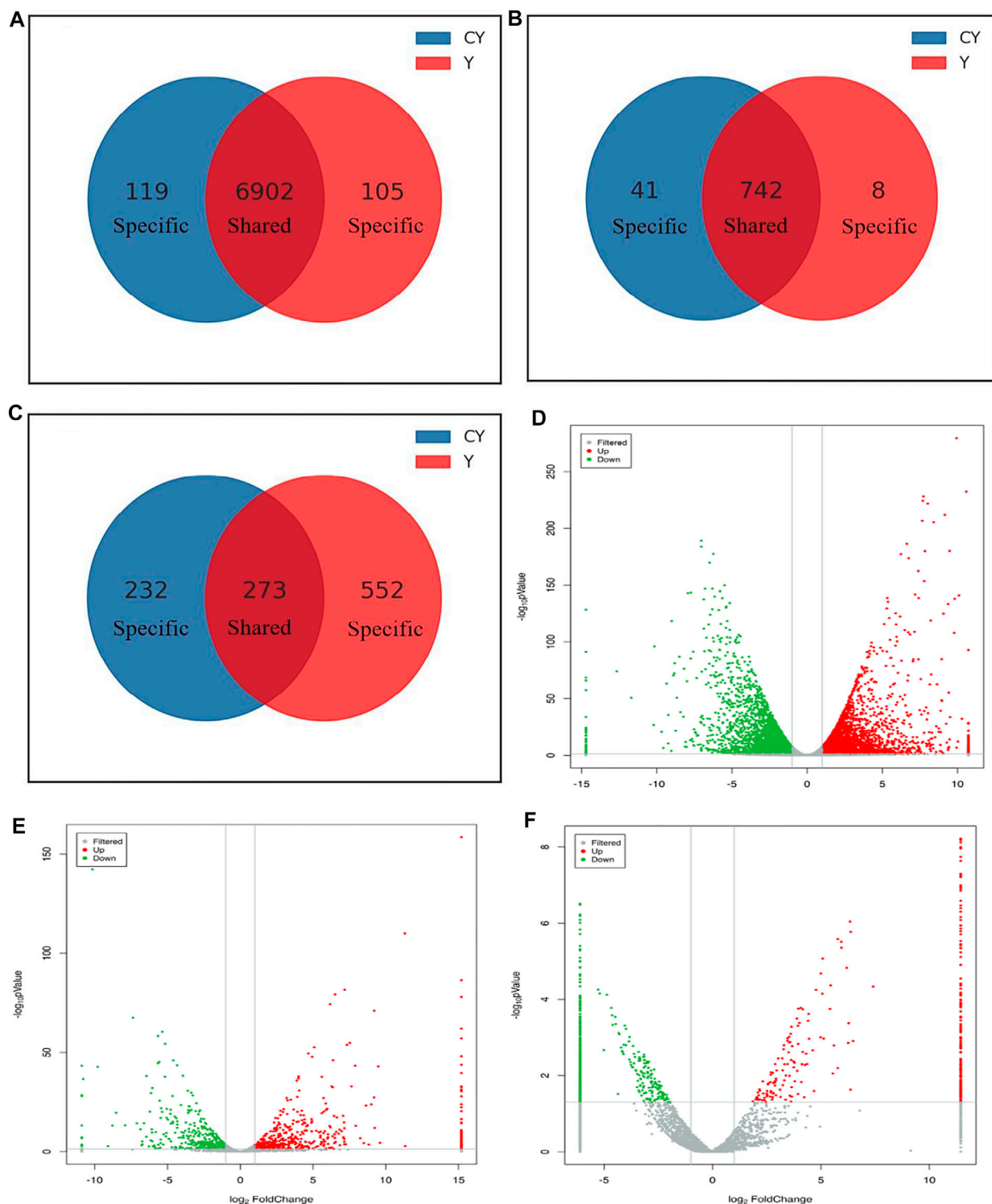


FIGURE 2 | Comparative analysis of the differentially expressed mRNAs and ncRNAs between cattle-yaks and yaks. The specific and shared mRNAs **(A)**, lncRNAs **(B)**, and circRNAs **(C)** between the two groups. The analysis of differentially expressed mRNAs **(D)**, lncRNAs **(E)**, and circRNAs **(F)** between two groups. Note: The red and green points represented upregulated and downregulated mRNAs, lncRNAs, and circRNAs, respectively. The gray points represented no significant differences. The gray vertical lines showed $|\log_2 \text{FC}| = 1$, and the gray horizontal lines showed $p = 0.05$.

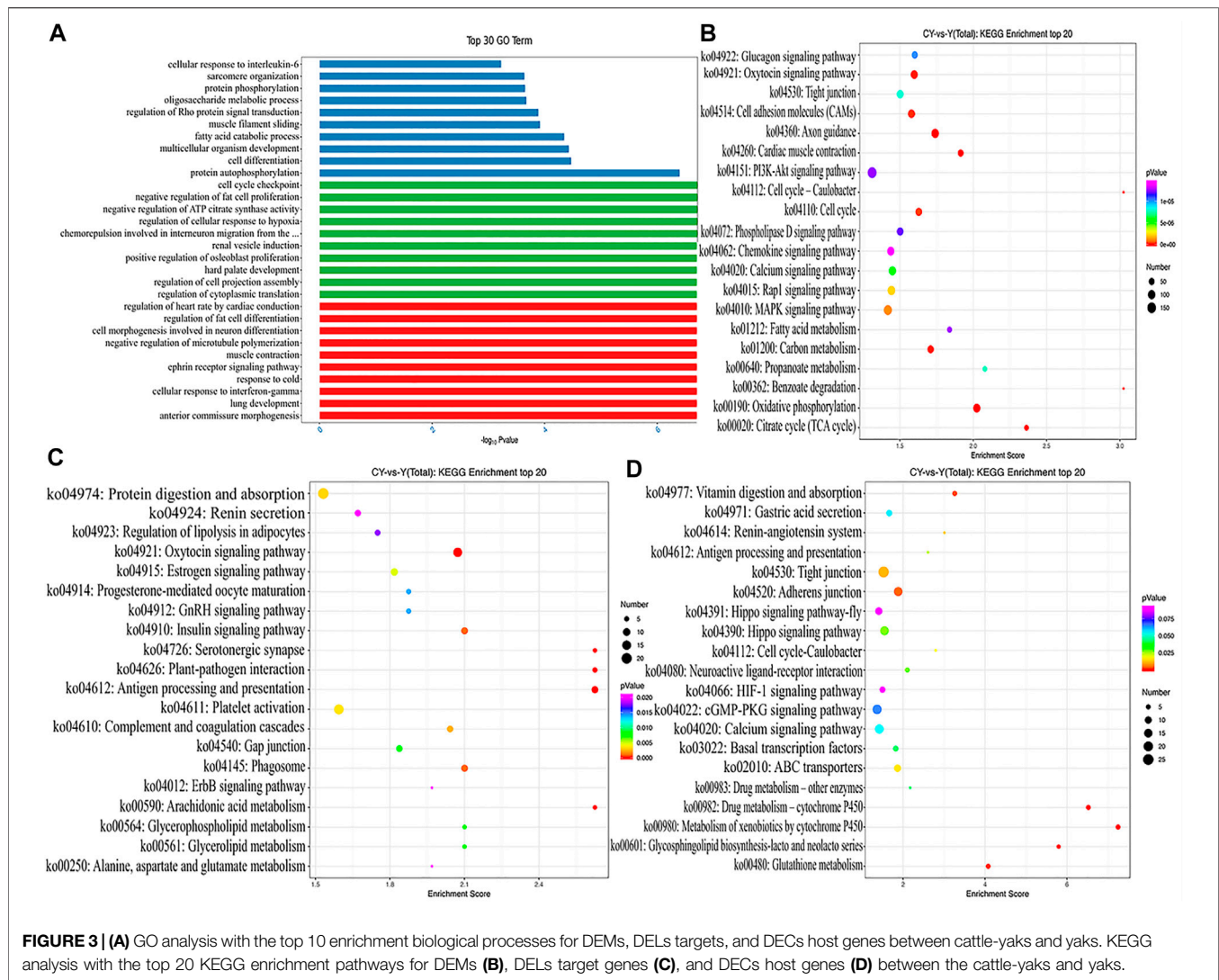


FIGURE 3 | (A) GO analysis with the top 10 enrichment biological processes for DEMs, DELs targets, and DEC host genes between cattle-yaks and yaks. KEGG analysis with the top 20 KEGG enrichment pathways for DEMs **(B)**, DELs target genes **(C)**, and DEC host genes **(D)** between the cattle-yaks and yaks.

Functional Analysis

GO analysis described the molecular functions performed by the DE ncRNAs, the cellular environment in which they were located, and the biological processes involved. We respectively selected the top 10 biological processes with the most significant enrichment of mRNAs, lncRNAs, and circRNAs (**Figure 3A**). Obviously, the DEMs were mainly enriched in regulation of osteoblast proliferation, negative regulation of fat cell proliferation, and regulation of cellular response to hypoxia. The targeted genes of DELs were most enriched in terms involving regulation of fat cell differentiation, ephrin receptor signaling pathway, and muscle contraction. Protein autophosphorylation, fatty acid catabolic process, and regulation of Rho protein signal transduction were the most significant enrichment terms for the sourced genes for DEC host genes. As evident from **Figures 3B–D**, we conducted the KEGG analysis using KEGG public pathway database and draw the augmented scatter diagram of the selected target genes. Between cattle-yaks and yaks, the DEMs were enriched in the PI3K–Akt signaling

pathway, MAPK signaling pathway, Fatty acid metabolism, Citrate cycle, etc. (**Figure 3B**). The DEL adjacent genes were significantly related to the protein digestion and absorption, GnRH signaling pathway, alanine, aspartate and glutamate metabolism, and so on (**Figure 3C**). The host genes of DEC host genes were significantly associated with some pathways, such as those related to vitamin digestion, ABC transporters, cGMP–PKG signaling pathway, etc. (**Figure 3D**). Interestingly, some DEMs and the host genes of DEC host genes were found to be enriched in the same pathways (i.e., hippo signaling pathway, cell cycle-caulobacter, and calcium signaling pathway), and some DEM and DEL adjacent genes were enriched in the same oxytocin signaling pathways.

In order to further explore the DEMs related to muscle growth and fatness, we screened out the related genes. These 117 DEMs related to muscle development and fat deposition are shown in **Supplementary Table S5**. The functional predictions of GO in DEMs (**Figure 4A**) mainly focused on some terms of skeletal muscle development, muscle cell differentiation, and regulation

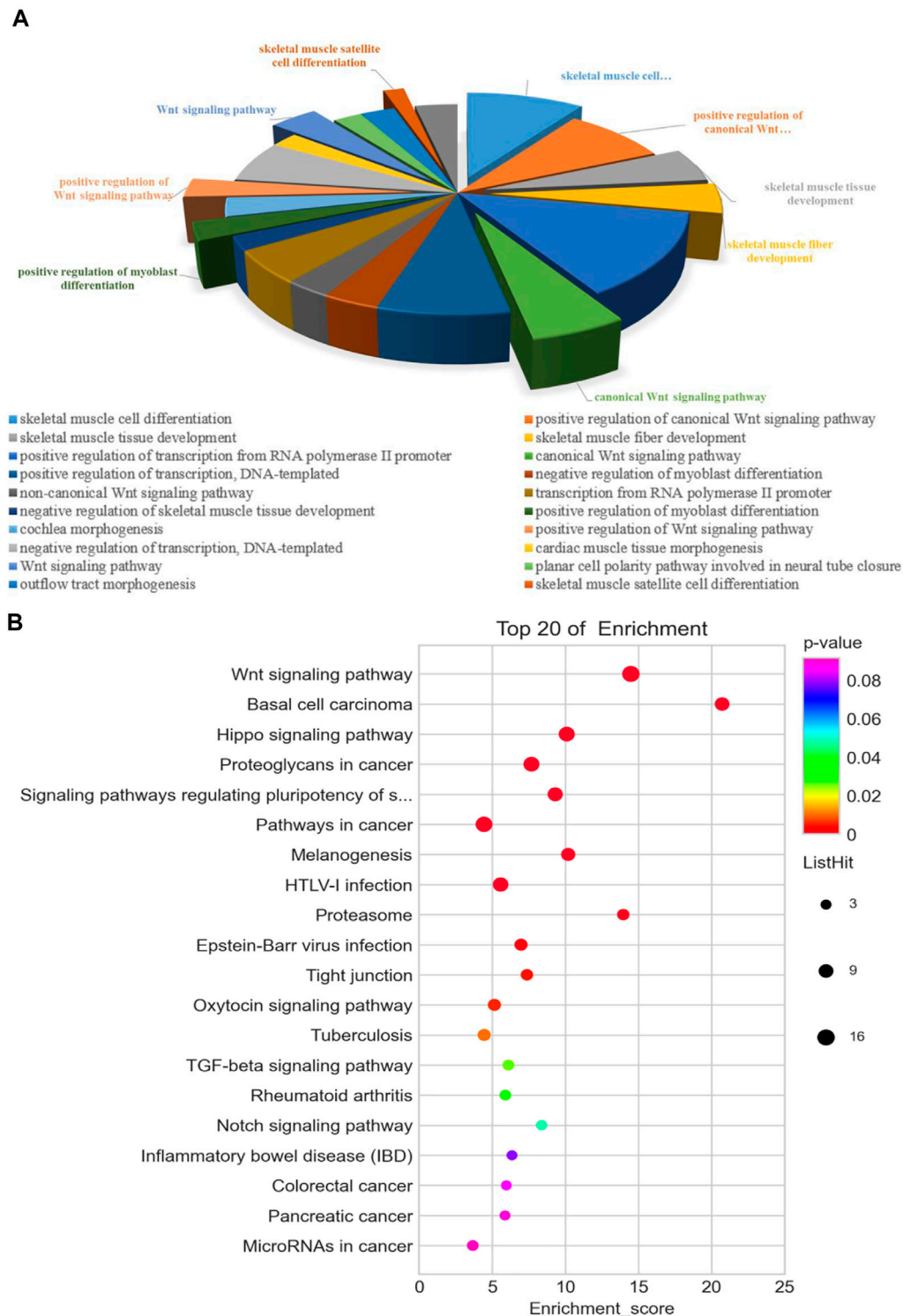


FIGURE 4 | Functional analysis of 117 differentially expressed genes associated with muscle development and fatness between the cattle-yak and yak. GO **(A)** and KEGG **(B)** analysis.

of canonical Wnt signaling pathway, including positive regulation of myoblast differentiation and skeletal muscle fiber development. According to the KEGG pathway analysis

(Figure 4B), some pathways were enriched significantly by these DEMs, such as Hippo signaling pathway, regulating many biological processes involved in proliferation, survival

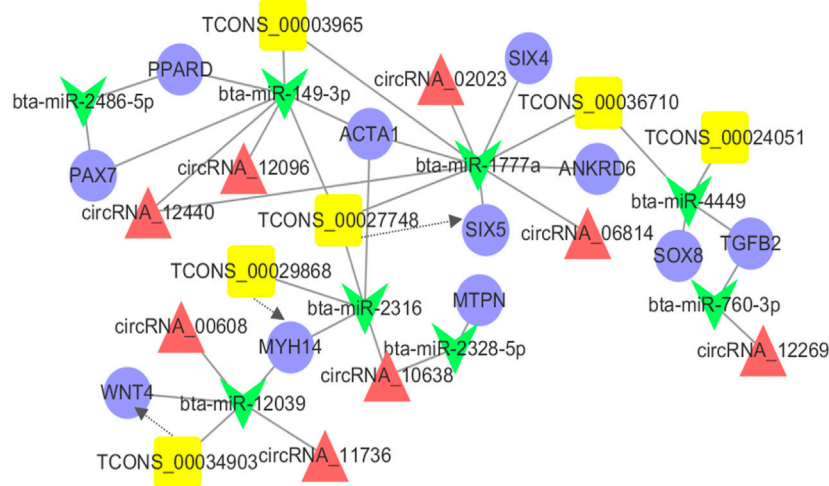


FIGURE 5 | The ceRNA co-regulation network. The blue circle, yellow box, red triangle and green V-type represented the differentially expressed mRNAs, lncRNAs, circRNAs, and miRNA, respectively. The solid line indicated the co-regulation between miRNAs and other transcripts. The dotted line indicated the co-regulation between the lncRNAs and mRNAs.

and differentiation of cell, organ size, and tissue homeostasis. In addition, some DEMs enriched significantly in Notch signaling pathway that can regulate the differentiation and development of cells, tissues, and organs through the interaction between the adjacent cells.

Construction of the ceRNA Coregulatory Network

Previous studies have shown that mRNAs, lncRNAs, and circRNAs may regulate gene function through miRNAs as ceRNAs in different processes (Salmena et al., 2011; Yu et al., 2019), indicating that ceRNAs and their miRNAs may work in concert with each other. Combined with the DEMs, DELs, and DECs related to muscle development and co-differentially expressed, we constructed the integrated ceRNA network. This ceRNA network contained 11 DEMs, 6 DELs, 8 DEC, and 33 relationships (Figure 5). Tcons-00034903 and bta-miR-2039 have a shared target gene *WNT4*. Similarly, we also found the same results in bta-miR-2316-Tcons-00029868-*MYH4* and bta-miR-1777a-Tcons-00027748-*SIX5*. This ceRNA network might provide valuable information for the development of the *longissimus dorsi* in cattle-yaks and yaks.

Real-Time Quantitative PCR Validation of Sequencing Data

Four mRNAs (*MYH14*, *LOC106700760*, *PIK3R2*, and *FGFR4*), 2 lncRNAs (Tcons-00034903 and Tcons-00004303), and 2 circRNAs (circ00012096 and circ00012564) were selected randomly for verifying the sequencing results through real-time quantitative PCR (Figure 6). The primers were designed using the Primer-BLAST web tool from the National Center for Biotechnology Information (NCBI) (Supplementary Table S6).

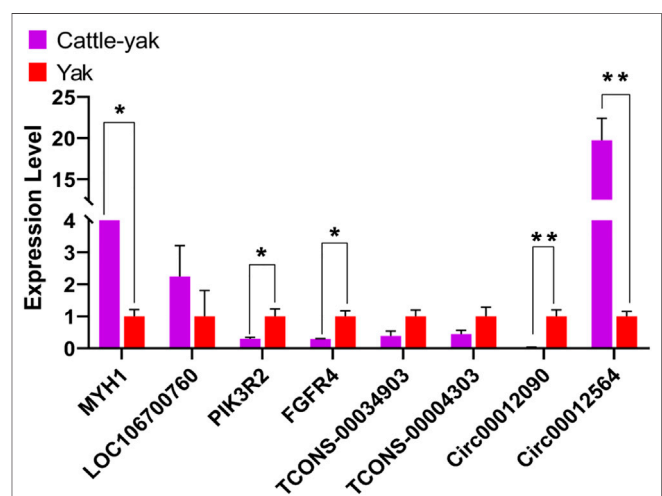


FIGURE 6 | The results of the real-time quantitative PCR validation of the expression level. ** indicates $p < 0.01$, * indicates $p < 0.05$. The data represented the mean \pm SEM from three biological replicates, and each measurement was repeated 3 times.

Their expression patterns were highly consistent with the sequencing results, indicating that the gene expression profiles obtained in this study had high repeatability and reliability.

DISCUSSION

Muscle growth is a complex economic trait owing to various physiological and biochemical indices. It concomitantly involves many gene expressions and regulations. To date, people have a great demand for high-quality meat and nutrition with improving living standards; it is an urgent problem for improving the

performance of yaks. Cattle-yak, as hybrid offspring of yak (♀) and cattle (♂), has been found to have significant improvement in the production performance (Guo et al., 2019; Dingkao et al., 2020; Luo et al., 2020; Jiang et al., 2021). However, the economic benefit of cattle-yaks has been ignored for a long time. The variations in the genes and proteins in the muscle structure are presumed generally to be affected by production performance (Joo et al., 1999); this study is the first time to systematically explore the differences of growth and development between cattle-yaks and yaks based on transcriptomics. To study the regulatory mechanism of growth and muscle development of cattle-yaks and yaks in a better way, we performed transcriptome analysis of the *longissimus dorsi* muscle. It was also the first time to compare the differences in the expression profiles of mRNA, lncRNA, and circRNA between the cattle-yaks and yaks, so as to determine the key factors involved in muscle growth and development. We measured the phenotypic data strictly of 30 cattle-yaks and yaks at three age groups of growth using the standard method of measurement, and the results obtained were similar to the previous studies; in each period of cattle-yaks with the same feeding conditions, production performance indexes of cattle-yaks were significantly higher than those of yaks ($p < 0.001$). We speculated that these differences may be due to genetic factors rather than the effects of feeding management on these traits. Therefore, the study on the mechanism regulating muscle growth and development in hybrid yaks and yak breeds can better help in enhancing the production performance of the yaks, providing help for exploring the regulatory mechanism of muscle development in mammals.

Typically, the growth and development of muscles are regulated by the core genes and signal transduction pathways (Myers et al., 2006; Ayuso et al., 2015). Compared to the yak group, a total of 7,126 DEMs, 791 DELs, and 1,057 DECs were identified in the cattle-yak group. Subsequently, the GO and KEGG pathway enrichment analysis revealed some important DEMs related to muscle growth and fat deposition, which was consistent with the results obtained after measuring the production performance, indicating that the production performance of the cattle-yak was better than that of the yak (Tumennasan et al., 1997). In addition, some DEMs, DELs, and DECs related to the immune system have reflected the adaptation of the yak to the high-altitude environment, and the adaptation of the yaks to the cold and high-altitude environment was well-preserved by the cattle-yaks. In succession, we focused on some DEMs related to the production performance of cattle-yaks and yaks. A total of 117 DEMs were identified that were related to the differentiation and proliferation of myoblasts, AMPK signaling pathway, MAPK signaling pathway, PI3K–Akt signaling pathway, etc. (Neri et al., 2002; Anderson, 2006; Gehart et al., 2010; Thomson, 2018). Among these DEMs, some have been found to have known functions in muscle growth and development. For example, myostatin (*MSTN*), as a member of the TGF- β superfamily, has a proven role as a growth differentiation factor (Kollias and McDermott, 2008; Li et al., 2008) playing an important role in mice (McPherron et al., 1997), cattle (McPherron and Lee, 1997), and humans (Schuelke et al., 2004) by negatively regulating the growth and differentiation of

myoblasts, as well as other mammals *via* controlling both the activation and proliferation of the satellite cells (skeletal muscle stem cells) (McCroskery et al., 2003). The expression level of the *MSTN* gene in our study was found to be downregulated in the cattle-yaks; hence, the expression of this gene was considered to be consistent with its function. This also provided new evidence for the function of the *MSTN* gene in *longissimus dorsi* of cattle-yaks and the high conservation across species. Myogenin (*MYOG*), including myogenic factor 6 (*MYF6*), is a regulatory factor in the family of Myogenic regulatory factors (*MRFs*), which is mainly involved in the fusion and differentiation of myoblasts (CHARGÉ and RUDNICKI, 2004; Buckingham and Vincent, 2009). Interestingly, the paired box transcription factor 7 (*PAX7*) gene and the *MYOG* gene were found to be downregulated, but the *MYF6* expression was upregulated in the cattle-yaks. Previous studies have shown that overexpression of gene *PAX7* can induce *MYOG* expression to inhibit myogenesis and prevent the differentiation of the muscle cell (Chen et al., 2010; Dey et al., 2011). Therefore, this also justified the downregulation of both *PAX7* and *MYOG* gene expression in cattle-yaks with better production performance than yaks. Additionally, some genes, such as actin alpha 1 skeletal muscle (*ACTA1*) and actin alpha cardiac muscle 1 (*ACTC1*), also play a key role in the differentiation and fusion of muscle cells, and have a positive impact on the myogenesis of the skeletal muscles (Ciecierska et al., 2020). Although these studies have helped in predicting the functions and accuracy of the key genes, the other functional DEMs related to muscle growth and development still need to be further studied on their expression regulation in *longissimus dorsi* muscle.

The main non-coding RNAs, the lncRNAs and circRNAs, are receiving more and more attention, as they can participate in the regulation of various biological processes in different ways (Liu et al., 2017; Marchese et al., 2017; Wang et al., 2019a). The identified DELs and DECs in this study were involved in regulating the promotion of muscle growth and development. lncRNA can exert its important action through various biological and pathological processes by demonstrating trans, cis, and antisense effects. Our results showed abundant differentially expressed lncRNAs in the skeletal muscle of cattle-yak and yak, suggesting that the lncRNAs may not be the exclusive by-products of mRNA in the cattle-yak but have specific roles. The long non-coding RNA acts differently in the nucleus and cytoplasm due to their different locations, where they are involved in the muscle development regulation on both embryonic and growing stages. According to the KEGG analysis, the differentially expressed lncRNAs and circRNAs were identified to functionally relate to some hormone regulation, myoblast proliferation, and metabolic pathways. CircRNAs being another type of non-coding RNA also act as an important regulatory role in skeletal muscle growth and development (Greco et al., 2018). Multiple studies have confirmed that abundant circRNAs exist in the skeletal muscle and their expression levels change dynamically during the process of myoblast differentiation (Legnini et al., 2017). Notably, we found that many DECs host genes are enriched in the HIF-1 signaling pathway in the KEGG analysis. HIF-1 signaling

pathway is the core signaling pathway induced by hypoxia involved in regulating the proliferation and differentiation of myoblasts upon hypoxic conditions (Ogilvie et al., 2000). HIF-1 pathway can upregulate its target genes, some of which involve the regulation of proliferation and differentiation of myoblasts. These DEC host genes were found to be significantly enriched in the HIF-1 signaling pathway, indicating a close relationship with the regulation of muscle growth and development under high-altitude and hypoxic environments. The regulatory mechanism of these circRNAs and their host genes on muscle development hence deserves further studies. Furthermore, our studies also indicated some lncRNAs and circRNAs to be specifically or mainly expressed in the *longissimus dorsi* muscle of cattle-yaks (e.g., Tcons-00005361, Tcons-00037918, and circRNA-02529), indicating that these ncRNAs were generated on purpose to have specific effects in the muscle development of the cattle-yaks.

In recent years, extensive studies on the function of miRNAs have provided a new theory named competing for endogenous RNA (ceRNA). To understand the process of development of the skeletal muscle, a ceRNA regulatory network was constructed based on the combination with the DEMs, DELs, and DECes related to muscle development and were expressed co-differentially. The results showed that 11 DEMs, 6 DELs, and 8 DECes cross-talked with another through 8 differential expressions of the microRNAs. This also indicated that the development of the *longissimus dorsi* muscle in cattle-yak was a complex regulation process of a balanced level of gene expression under a high-altitude and hypoxic environment. As reported, peroxisome proliferator-activated receptor delta (*PPARD*) acted as a vital regulator in adipogenesis and lipid metabolism (Kim et al., 2006; Chen and Yang, 2014). Ankyrin repeat domain 6 (*ANKRD6*) also played a role in regulating crucial events in developing vertebrates and invertebrates (Schwarz-Romond et al., 2002; Moeller et al., 2006). Therefore, we speculated that these ncRNAs might also contribute to muscle development by indirectly regulating the gene expression of *PPARD* and *ANKRD6*. Not only that, we observed three important ceRNA subnetworks from the ceRNA network, showing that TCONS-00024051 and its target *SOX8* “talked” to each other through the same bta-miR-1777a response element, while TCONS-00034903 and its target *WNT4*, TCONS-00029868 and its target *ACTA1* “talked” to each other through bta-miR-12039 and bta-miR-4449 response elements, respectively. Therefore, we speculated that these three subnetworks may be crucially associated and function in regulating muscle development. Notably, as a member of the “unconventional” non-muscle myosin II family of molecular motors, myosin heavy chain 14 (*MYH14*) gene has been identified as a key regulator of muscle fiber type (van Rooij et al., 2009; Bell et al., 2010). In our results, *MYH14* and Tcons-00029868 were both upregulated in *longissimus dorsi* muscle, which indicates that this lncRNA may have a *cis*-regulatory relationship with *MYH14* gene. From our results, these DEMs, DELs, and DECes were not only involved in the process of muscle development, but might also be involved in lipogenesis as ceRNAs.

CONCLUSION

In conclusion, we compared the production performance of cattle-yaks and yaks, it was also the first time to compare the expressional features of mRNAs, lncRNAs, and circRNAs in the *longissimus dorsi* tissue of cattle-yaks and yaks. In our results, the abundant mRNAs and ncRNAs were identified, some of which were specifically expressed in cattle-yaks. According to the bioinformatics analyses, the ncRNAs were found to be not only connected with the myoblast differentiation and proliferation, skeletal development, and signaling pathway of muscle growth, but be also useful as ceRNAs of important transcription factors (such as *SOX8* and *PPARD*). In addition, the ceRNA network (11 DEMs, 6 DELs, and 8 DECes) that we constructed may have the considerable effects on regulation of muscle growth. This study provided new insights into the genetic basis of muscle growth and laid the foundation for further study of the role of these ncRNAs in regulating muscle growth.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/>, PRJNA753699.

ETHICS STATEMENT

The animal study was reviewed and approved by the Lanzhou Institute of Husbandry and Pharmaceutical Sciences Chinese Academy of Agricultural Sciences. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

CL, PY, and CH contributed to conception and design of the study. CH and FG organized the database. FG and XM performed the statistical analysis. CH, FG, and XM wrote the first draft of the manuscript. RFD, RQD, ZZ, and GB collected the production data. XG, PB, XM, XW, and MC wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

Work presented in this paper was supported by the Project of Innovative Research on Yak Molecular Breeding Technology under the Innovation Population of Basic Research in Gansu Province, grant number 20JR5RA580, the Agricultural Science and Technology Innovation Program, grant

number CAAS-ASTIP-2014-LIHPS-01, the National Beef Cattle Industry Technology and System, grant number CARS-37, and the Herbivorous Livestock Industry and Technical System of Gansu Province, grant number GARS-08.

REFERENCES

- Anders, S., and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome Biol.* 11 (10), R106. doi:10.1186/gb-2010-11-10-r106
- Anderson, D. H. (2006). Role of Lipids in the MAPK Signaling Pathway. *Prog. Lipid Res.* 45 (2), 102–119. doi:10.1016/j.plipres.2005.12.003
- Arrighi, N., Moratal, C., Savary, G., Fassy, J., Nottet, N., Pons, N., et al. (2021). The FibromiR miR-214-3p Is Upregulated in Duchenne Muscular Dystrophy and Promotes Differentiation of Human Fibro-Adipogenic Muscle Progenitors. *Cells* 10 (7), 1832. doi:10.3390/cells10071832
- Ayuso, M., Fernández, A., Núñez, Y., Benítez, R., Isabel, B., Barragán, C., et al. (2015). Comparative Analysis of Muscle Transcriptome between Pig Genotypes Identifies Genes and Regulatory Mechanisms Associated to Growth, Fatness and Metabolism. *PLoS One* 10 (12), e0145162. doi:10.1371/journal.pone.0145162
- Bell, M. L., Buvoli, M., and Leinwand, L. A. (2010). Uncoupling of Expression of an Intronic microRNA and its Myosin Host Gene by Exon Skipping. *Mol. Cell Biol.* 30 (8), 1937–1945. doi:10.1128/mcb.01370-09
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170
- Buckingham, M., and Vincent, S. D. (2009). Distinct and Dynamic Myogenic Populations in the Vertebrate Embryo. *Curr. Opin. Genet. Dev.* 19 (5), 444–453. doi:10.1016/j.gde.2009.08.001
- Caretti, G., Schiltz, R. L., Dilworth, F. J., Di Padova, M., Zhao, P., Ogryzko, V., et al. (2006). The RNA Helicases P68/p72 and the Noncoding RNA SRA Are Coregulators of MyoD and Skeletal Muscle Differentiation. *Dev. Cell* 11 (4), 547–560. doi:10.1016/j.devcel.2006.08.003
- Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., et al. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. *Cell* 147 (2), 358–369. doi:10.1016/j.cell.2011.09.028
- Chargé, S. B. P., and Rudnicki, M. A. (2004). Cellular and Molecular Regulation of Muscle Regeneration. *Physiol. Rev.* 84 (1), 209–238. doi:10.1152/physrev.00019.2003
- Chen, J.-F., Tao, Y., Li, J., Deng, Z., Yan, Z., Xiao, X., et al. (2010). microRNA-1 and microRNA-206 Regulate Skeletal Muscle Satellite Cell Proliferation and Differentiation by Repressing Pax7. *J. Cell Biol.* 190 (5), 867–879. doi:10.1083/jcb.200911036
- Chen, L., and Yang, G. (2014). PPARs Integrate the Mammalian Clock and Energy Metabolism. *PPAR Res.* 2014, 1–6. doi:10.1155/2014/653017
- Ciecierska, A., Motyl, T., and Sadkowski, T. (2020). Transcriptomic Profile of Primary Culture of Skeletal Muscle Cells Isolated from Semitendinosus Muscle of Beef and Dairy Bulls. *Ijms* 21 (13), 4794. doi:10.3390/ijms21134794
- Dey, B. K., Gagan, J., and Dutta, A. (2011). miR-206 and -486 Induce Myoblast Differentiation by Downregulating Pax7. *Mol. Cell Biol.* 31(1), 203–214. doi:10.1128/MCB.01009-10
- Dingkao, R. Q., Shi, H. M., Li, P. X., Cairang, N. R., Ma, D. L., Niu, X. L., et al. (2020). Application of Early Pregnancy Diagnostic Technique in Xianan Cattle Production. *China Cattle Sci.* 46 (5), 11–14.
- Gandolfi, G., Pomponio, L., Ertbjerg, P., Karlsson, A. H., Nanni Costa, L., Lametsch, R., et al. (2011). Investigation on CAST, CAPN1 and CAPN3 Porcine Gene Polymorphisms and Expression in Relation to post-mortem Calpain Activity in Muscle and Meat Quality. *Meat Sci.* 88 (4), 694–700. doi:10.1016/j.meatsci.2011.02.031
- Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an Efficient and Unbiased Algorithm for De Novo Circular RNA Identification. *Genome Biol.* 16 (1), 4. doi:10.1186/s13059-014-0571-3
- Gehart, H., Kumpf, S., Ittner, A., and Ricci, R. (2010). MAPK Signalling in Cellular Metabolism: Stress or Wellness? *EMBO Rep.* 11 (11), 834–840. doi:10.1038/embor.2010.160
- Ghosh, S., and Chan, C.-K. K. (2016). “Analysis of RNA-Seq Data Using TopHat and Cufflinks,” in *Plant Bioinformatics: Methods and Protocols*. Editor D. Edwards (New York, NY: Springer New York), 339–361. doi:10.1007/978-1-4939-3167-5_18
- Greco, S., Cardinali, B., Falcone, G., and Martelli, F. (2018). Circular RNAs in Muscle Function and Disease. *Ijms* 19 (11), 3454. doi:10.3390/ijms19113454
- Guo, S., Ma, D., Li, B., Bao, Z., Zhang, T., and Xu, G. (2019). Determination of Growth and Development Indexes of Cattle-Yak in Gannan alpine Pasture Area. *China Herbivore Sci.* 039 (002), 73–75. doi:10.3969/j.issn.2095-3887.2019.02.021
- Ji, Q., Cai, G., Liu, X., Zhang, Y., Wang, Y., Zhou, L., et al. (2019). MALAT1 Regulates the Transcriptional and Translational Levels of Proto-Oncogene RUNX2 in Colorectal Cancer Metastasis. *Cell Death Dis.* 10 (6), 378. doi:10.1038/s41419-019-1598-x
- Jiang, H., Zhang, C. F., Zhang, Q., Sun, G. M., Cao, H. W., Ciji, Y. D., et al. (2021). Study on Yak Economic Crossbreeding and Growth Performance of F1 Generation. *Tibet J. Agric. Sci.* 43 (02), 22–24.
- John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA Targets. *Plos Biol.* 2 (11), e363. doi:10.1371/journal.pbio.0020363
- Joo, S. T., Kauffman, R. G., Kim, B. C., and Park, G. B. (1999). The Relationship of Sarcoplasmic and Myofibrillar Protein Solubility to Colour and Water-Holding Capacity in Porcine Longissimus Muscle. *Meat Sci.* 52 (3), 291–297. doi:10.1016/S0309-1740(99)00005-4
- Kallen, A. N., Zhou, X.-B., Xu, J., Qiao, C., Ma, J., Yan, L., et al. (2013). The Imprinted H19 lncRNA Antagonizes Let-7 microRNAs. *Mol. Cell* 52 (1), 101–112. doi:10.1016/j.molcel.2013.08.027
- Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. *Nucleic Acids Res.* 45 (W1), W12–W16. doi:10.1093/nar/gkx428
- Kim, D. J., Bility, M. T., Billin, A. N., Willson, T. M., Gonzalez, F. J., and Peters, J. M. (2006). Ppar β/δ Selectively Induces Differentiation and Inhibits Cell Proliferation. *Cell Death Differ* 13 (1), 53–60. doi:10.1038/sj.cdd.4401713
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a Fast Spliced Aligner with Low Memory Requirements. *Nat. Methods* 12 (4), 357–360. doi:10.1038/nmeth.3317
- Kollias, H. D., and McDermott, J. C. (2008). Transforming Growth Factor- β and Myostatin Signaling in Skeletal Muscle. *J. Appl. Physiol.* 104 (3), 579–587. doi:10.1152/japplphysiol.01091.2007
- Korostowski, L., Sedlak, N., and Engel, N. (2012). The Kcnq1ot1 Long Non-coding RNA Affects Chromatin Conformation and Expression of Kcnq1, but Does Not Regulate its Imprinting in the Developing Heart. *Plos Genet.* 8 (9), e1002956. doi:10.1371/journal.pgen.1002956
- Legnini, I., Di Timoteo, G., Rossi, F., Morlando, M., Briganti, F., Sthandier, O., et al. (2017). Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol. Cell* 66 (1), 22–37. e29. doi:10.1016/j.molcel.2017.02.017
- Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a Tool for Predicting Long Non-coding RNAs and Messenger RNAs Based on an Improved K-Mer Scheme. *BMC Bioinformatics* 15 (1), 311. doi:10.1186/1471-2105-15-311
- Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324
- Li, H., Wei, X., Yang, J., Dong, D., Hao, D., Huang, Y., et al. (2018). circFGFR4 Promotes Differentiation of Myoblasts via Binding miR-107 to Relieve its Inhibition of Wnt3a. *Mol. Ther. - Nucleic Acids* 11, 272–283. doi:10.1016/j.omtn.2018.02.012

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.772557/full#supplementary-material>

- Li, Z. B., Kollias, H. D., and Wagner, K. R. (2008). Myostatin Directly Regulates Skeletal Muscle Fibrosis. *J. Biol. Chem.* 283 (28), 19371–19378. doi:10.1074/jbc.M802585200
- Li, Z., Cai, B., Abdalla, B. A., Zhu, X., Zheng, M., Han, P., et al. (2019). lncIRS1 Controls Muscle Atrophy via Sponging miR-15 Family to Activate IGF1-PI3K/AKT Pathway. *J. Cachexia, Sarcopenia Muscle* 10 (2), 391–410. doi:10.1002/jcsm.12374
- Liu, J., Liu, T., Wang, X., and He, A. (2017). Circles Reshaping the RNA World: from Waste to Treasure. *Mol. Cancer* 16 (1), 58. doi:10.1186/s12943-017-0630-y
- Liu, X., Liu, K., Shan, B., Wei, S., Li, D., Han, H., et al. (2018). A Genome-wide Landscape of mRNAs, lncRNAs, and circRNAs during Subcutaneous Adipogenesis in Pigs. *J. Anim. Sci. Biotechnol.* 9 (1), 76. doi:10.1186/s40104-018-0292-7
- Livak, K. J., and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2- $\Delta\Delta$ CT Method. *Methods* 25 (4), 402–408. doi:10.1006/meth.2001.1262
- Luo, Q. B., Buren, C. G. T., Si, Q., Zhaxi, Z. M., Zhang, S., Chang, L., et al. (2020). Study on Crossbreeding Effect between Angus Beef Cattle and Yak. *China Cattle Sci.* 46 (5), 8–10.
- Ma, X., Fu, D., Chu, M., Ding, X., Wu, X., Guo, X., et al. (2020). Genome-Wide Analysis Reveals Changes in Polled Yak Long Non-coding RNAs in Skeletal Muscle Development. *Front. Genet.* 11 (365). doi:10.3389/fgene.2020.00365
- Marchese, F. P., Raimondi, I., and Huarte, M. (2017). The Multidimensional Mechanisms of Long Noncoding RNA Function. *Genome Biol.* 18 (1), 206. doi:10.1186/s13059-017-1348-2
- McCroskey, S., Thomas, M., Maxwell, L., Sharma, M., and Kambadur, R. (2003). Myostatin Negatively Regulates Satellite Cell Activation and Self-Renewal. *J. Cell Biol.* 162 (6), 1135–1147. doi:10.1083/jcb.200207056
- McPherron, A. C., Lawler, A. M., and Lee, S.-J. (1997). Regulation of Skeletal Muscle Mass in Mice by a New TGF- β Superfamily Member. *Nature* 387 (6628), 83–90. doi:10.1038/387083a0
- McPherron, A. C., and Lee, S.-J. (1997). Double Muscling in Cattle Due to Mutations in the Myostatin Gene. *Proc. Natl. Acad. Sci.* 94 (23), 12457–12461. doi:10.1073/pnas.94.23.12457
- Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521
- Moeller, H., Jenny, A., Schaeffer, H.-J., Schwarz-Romond, T., Mlodzik, M., Hammerschmidt, M., et al. (2006). Diversin Regulates Heart Formation and Gastrulation Movements in Development. *Proc. Natl. Acad. Sci.* 103 (43), 15900–15905. doi:10.1073/pnas.0603808103
- Myers, S. A., Wang, S.-C. M., and Muscat, G. E. O. (2006). The Chicken Ovalbumin Upstream Promoter-Transcription Factors Modulate Genes and Pathways Involved in Skeletal Muscle Cell Metabolism. *J. Biol. Chem.* 281 (34), 24149–24160. doi:10.1074/jbc.M601941200
- Neri, L. M., Borgatti, P., Capitani, S., and Martelli, A. M. (2002). The Nuclear Phosphoinositide 3-kinase/AKT Pathway: a New Second Messenger System. *Biochim. Biophys. Acta* 1584 (2), 73–80. doi:10.1016/S1388-1981(02)00300-1
- Ogilvie, M., Yu, X., Nicolas-Metral, V., Pulido, S. M., Liu, C., Ruegg, U. T., et al. (2000). Erythropoietin Stimulates Proliferation and Interferes with Differentiation of Myoblasts. *J. Biol. Chem.* 275 (50), 39754–39761. doi:10.1074/jbc.M004999200
- Perte, M., Perte, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122
- Rudnicki, M. A., and Jaenisch, R. (1995). The MyoD Family of Transcription Factors and Skeletal Myogenesis. *Bioessays* 17 (3), 203–209. doi:10.1002/bies.950170306
- Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* 146 (3), 353–358. doi:10.1016/j.cell.2011.07.014
- Schuelke, M., Wagner, K. R., Stolz, L. E., Hübner, C., Riebel, T., Kömen, W., et al. (2004). Myostatin Mutation Associated with Gross Muscle Hypertrophy in a Child. *N. Engl. J. Med.* 350 (26), 2682–2688. doi:10.1056/NEJMoa040933
- Schwarz-Romond, T., Asbrand, C., Bakkers, J., Kühl, M., Schaeffer, H.-J., Huelsken, J., et al. (2002). The Ankyrin Repeat Protein Diversin Recruits Casein Kinase I ϵ to the β -catenin Degradation Complex and Acts in Both Canonical Wnt and Wnt/JNK Signaling. *Genes Dev.* 16 (16), 2073–2084. doi:10.1101/gad.230402
- Song, C., Huang, Y., Yang, Z., Ma, Y., Chaogetu, B., Zhuoma, Z., et al. (2019). RNA-seq Analysis Identifies Differentially Expressed Genes in Subcutaneous Adipose Tissue in Qaidaford Cattle, Cattle-Yak, and Angus Cattle. *Animals* 9 (12), 1077. doi:10.3390/ani9121077
- Song, C., Wang, J., Ma, Y., Yang, Z., Dong, D., Li, H., et al. (2018). linc-smad7 Promotes Myoblast Differentiation and Muscle Regeneration via Sponging miR-125b. *Epigenetics* 13 (6), 591–604. doi:10.1080/15592294.2018.1481705
- Sonnhammer, E., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains. *Nucleic Acids Res.* 26 (1), 320–322. doi:10.1093/nar/26.1.320
- Sun, L., Luo, S., Bai, M., Xiang, L., Li, J., Jia, C., et al. (2019). Integrative microRNA-mRNA Analysis of Muscle Tissues in Qianhua Mutton Merino and Small Tail Han Sheep Reveals Key Roles for Oar-miR-655-3p and Oar-miR-381-5p. *DNA Cel Biol.* 38 (5), 423–435. doi:10.1089/dna.2018.4408
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing Sequence Intrinsic Composition to Classify Protein-Coding and Long Non-coding Transcripts. *Nucleic Acids Res.* 41 (17), e166. doi:10.1093/nar/gkt646
- Tapscott, S. J. (2005). The Circuitry of a Master Switch: MyoD and the Regulation of Skeletal Muscle Gene Transcription. *Development* 132 (12), 2685–2695. doi:10.1242/dev.01874
- Thomson, D. (2018). The Role of AMPK in the Regulation of Skeletal Muscle Size, Hypertrophy, and Regeneration. *Ijms* 19 (10), 3125. doi:10.3390/ijms19103125
- Tumennasan, K., Tuya, T., Hotta, Y., Takase, H., Speed, R. M., and Chandley, A. C. (1997). Fertility Investigations in the F1 Hybrid and Backcross Progeny of Cattle (*Bos taurus*) Andyak (B. Grunniens) in Mongolia. *Cytogenet. Cell Genet* 78 (1), 69–73. doi:10.1159/000134633
- van Rooij, E., Quat, D., Johnson, B. A., Sutherland, L. B., Qi, X., Richardson, J. A., et al. (2009). A Family of microRNAs Encoded by Myosin Genes Governs Myosin Expression and Muscle Performance. *Dev. Cell* 17 (5), 662–673. doi:10.1016/j.devcel.2009.10.013
- Wang, C., Liu, W., Nie, Y., Qaher, M., Horton, H. E., Yue, F., et al. (2017). Loss of MyoD Promotes Fate Transdifferentiation of Myoblasts into Brown Adipocytes. *EBioMedicine* 16, 212–223. doi:10.1016/j.ebiom.2017.01.015
- Wang, H., Ma, M., Li, Y., Liu, J., Sun, C., Liu, S., et al. (2021a). miR-183 and miR-96 Orchestrate Both Glucose and Fat Utilization in Skeletal Muscle. *EMBO Rep.* 22, e52247. doi:10.15252/embr.202052247
- Wang, J., Ren, Q., Hua, L., Chen, J., Zhang, J., Bai, H., et al. (2019a). Comprehensive Analysis of Differentially Expressed mRNA, lncRNA and circRNA and Their ceRNA Networks in the Longissimus Dorsi Muscle of Two Different Pig Breeds. *Ijms* 20 (5), 1107. doi:10.3390/ijms20051107
- Wang, Y., Li, M., Wang, Y., Liu, J., Zhang, M., Fang, X., et al. (2019b). A Zfp609 Circular RNA Regulates Myoblast Differentiation by Sponging miR-194-5p. *Int. J. Biol. Macromolecules* 121, 1308–1313. doi:10.1016/j.ijbiomac.2018.09.039
- Wang, Y., Wang, Z., Hu, R., Peng, Q., Xue, B., and Wang, L. (2021b). Comparison of Carcass Characteristics and Meat Quality between Simmental Crossbred Cattle, Cattle-yaks and Xuanhan Yellow Cattle. *J. Sci. Food Agric.* 101 (9), 3927–3932. doi:10.1002/jsfa.11032
- Xu, X., Mishra, B., Qin, N., Sun, X., Zhang, S., Yang, J., et al. (2019). Differential Transcriptome Analysis of Early Postnatal Developing Longissimus Dorsi Muscle from Two Pig Breeds Characterized in Divergent Myofiber Traits and Fatness. *Anim. Biotechnol.* 30 (1), 63–74. doi:10.1080/10495398.2018.1437045
- Yu, C., Longfei, L., Long, W., Feng, Z., Chen, J., Chao, L., et al. (2019). lncRNA PVT1 Regulates VEGFC through Inhibiting miR-128 in Bladder Cancer Cells. *J. Cel Physiol.* 234 (2), 1346–1353. doi:10.1002/jcp.26929

- Zhang, J. S., Xu, H. Y., Fang, J. C., Yin, B. Z., Wang, B. B., Pang, Z., et al. (2021). Integrated microRNA-mRNA Analysis Reveals the Roles of microRNAs in the Muscle Fat Metabolism of Yanbian Cattle. *Anim. Genet.* 52, 598–607. doi:10.1111/age.13126
- Zhang, Z.-K., Li, J., Guan, D., Liang, C., Zhuo, Z., Liu, J., et al. (2018). A Newly Identified lncRNA MAR1 Acts as a miR-487b Sponge to Promote Skeletal Muscle Differentiation and Regeneration. *J. Cachexia, Sarcopenia Muscle* 9 (3), 613–626. doi:10.1002/jcsm.12281

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Huang, Ge, Ma, Dai, Dingkao, Zhaxi, Burechao, Bao, Wu, Guo, Chu, Yan and Liang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Integrated Analysis of mRNA and MicroRNA Co-expressed Network for the Differentiation of Bovine Skeletal Muscle Cells After Polyphenol Resveratrol Treatment

Dan Hao^{1,2†}, Xiao Wang^{3†}, Yu Yang¹, Bo Thomsen², Lars-Erik Holm², Kaixing Qu⁴, Bizhi Huang^{5*} and Hong Chen^{1,6*}

¹ College of Animal Science and Technology, Northwest A&F University, Shaanxi Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, Yangling, China, ² Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark, ³ Konge Larsen ApS, Kongens Lyngby, Denmark, ⁴ Academy of Science and Technology, Chuxiong Normal University, Chuxiong, China, ⁵ Yunnan Academy of Grassland and Animal Science, Kunming, China, ⁶ College of Animal Science, Xinjiang Agricultural University, Urumqi, China

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

Mao Yongjiang,
Yangzhou University, China
Yulin Jin,
Emory University, United States

*Correspondence:

Bizhi Huang
hbz@ynbp.cn
Hong Chen
chenhong1212@263.net

†These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Veterinary Science

Received: 15 September 2021

Accepted: 15 November 2021

Published: 24 December 2021

Citation:

Hao D, Wang X, Yang Y, Thomsen B, Holm L-E, Qu K, Huang B and Chen H (2021) Integrated Analysis of mRNA and MicroRNA Co-expressed Network for the Differentiation of Bovine Skeletal Muscle Cells After Polyphenol Resveratrol Treatment. *Front. Vet. Sci.* 8:777477. doi: 10.3389/fvets.2021.777477

Resveratrol (RSV) has been confirmed to benefit human health. Resveratrol supplemented in the feeds of animals improved pork, chicken, and duck meat qualities. In this study, we identified differentially expressed (DE) messenger RNAs (mRNAs) ($n = 3,856$) and microRNAs (miRNAs) ($n = 93$) for the weighted gene co-expression network analysis (WGCNA) to investigate the co-expressed DE mRNAs and DE miRNAs in the primary bovine myoblasts after RSV treatment. The mRNA results indicated that RSV treatments had high correlations with turquoise module (0.91, P -value = 0.01) and blue module (0.93, P -value < 0.01), while only the turquoise module (0.96, P -value < 0.01) was highly correlated with the treatment status using miRNA data. After biological enrichment analysis, the 2,579 DE genes in the turquoise module were significantly enriched in the Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The top two GO terms were actin filament-based process (GO:0030029) and actin cytoskeleton organization (GO:0030036). The top two KEGG pathways were regulation of actin cytoskeleton (bta04810) and tight junction (bta04530). Then, we constructed the DE mRNA co-expression and DE miRNA co-expression networks in the turquoise module and the mRNA-miRNA targeting networks based on their co-expressions in the key module. In summary, the RSV-induced miRNAs participated in the co-expression networks that could affect mRNA expressions to regulate the primary myoblast differentiation. Our study provided a better understanding of the roles of RSV in inducing miRNA and of the characteristics of DE miRNAs in the key co-expressed module in regulation of mRNAs and revealed new candidate regulatory miRNAs and genes for the beef quality traits.

Keywords: primary bovine myoblast, resveratrol, differentially expressed analysis, WGCNA, mRNA co-expression, miRNA co-expression, mRNA-miRNA network

INTRODUCTION

Resveratrol (RSV) is a natural polyphenol compound found in grapes, nuts, and some blackberries. Researchers have studied its health-promoting effects of neuroprotection (1) and cardioprotection (2) as well as its inhibiting actions to tumor cell proliferation (3) and microbial activity (4) and its diminishing effects on inflammation in humans and animals (5, 6). Its pro-differentiation properties to human lung fibroblasts (7), embryonic cardiomyoblasts (8), and skeletal myoblast have also been studied (9). For example, Dirks Naylor (10) demonstrated the RSV effects on skeletal muscle metabolism, protein catabolism, and muscle-related ischemia and reperfusion injury disease in a review study (10). Resveratrol could also help to improve muscle fatigue resistance (11), reduce aging-induced muscle loss (12), improve muscle atrophy (13), and enhance exercise performance (14). Resveratrol is contained in the wine grape pomace, and adding it to feed will benefit the feed efficiency and meat tenderness in lamb (15). In addition, the antioxidant effects of RSV improved the heat-stressed and the transport-stressed meat quality of broilers (16, 17). Resveratrol alleviated the skeletal muscle mitochondrial dysfunction and oxidative damage, when 80 mg/kg/day RSV was supplied in the intrauterine growth retardation piglets (18). The same dose of RSV also improved the meat quality by increasing the content of oxidative muscle fiber and decreasing the lipid accumulation in pigs (19). Dietary RSV supplements of 300–450 mg/kg in Peking ducks improved meat quality through decreasing abdominal fat rate and shear force, as well-increasing the flavor amino acid and intramuscular fat deposition (20). In cattle, the beneficial effects of RSV have been concluded in several studies on bovine oocyte maturation and subsequent embryonic development (21), inhibition of apoptosis and lipid peroxidation for the fertilization capacity of bovine sex-sorted semen (22), rumen fermentation, methane production, and prokaryotic community composition (23). However, the effect of RSV on beef production and quality still needs further investigation.

The carcass composition of beef cattle is influenced by intrinsic factors (e.g., genetic, age, and sex) and extrinsic factors (e.g., nutrition, environment, and management) (24). Bassel et al. (25) suggested the establishment of global co-expression network connections between genes by considering all samples in *Arabidopsis*. Gene co-expression networks are constructed by genes with significant co-expression relationships, where the co-expressed genes show similar expression patterns across samples that are controlled by the same transcriptional regulatory programs (26, 27). The weighted gene co-expression network analysis (WGCNA) has been used in analyzing the feed efficiency, residual feed intake, carcass traits, and lactation in cattle (28–30).

Abbreviations: *ACTG1*, actin gamma 1 gene; *DGAT1*, diacylglycerol acyltransferase-1 gene; DE, differentially expressed; FPKM, fragments per kilobase of transcript sequence per million base pairs sequenced; FDR, false discovery rate; FCs, fold changes; GO, Gene Ontology; IACUC, Institutional Animal Care and Use Committee; MAD, median absolute deviation; mRNAs, messenger RNAs; MRFs, myogenic regulatory factors; *MEF2*, myocyte enhancer factor 2; RSV, resveratrol; TPM, transcript per million; KEGG, Kyoto Encyclopedia of Genes and Genomes; WGCNA, weighted gene co-expression network analysis.

Our previous study focused on transcriptomic changes in bovine skeletal muscle cells after RSV treatment and was conducted to identify the differentially expressed (DE) genes and microRNAs (miRNAs) (31, 32); therefore, this study mainly focuses on the combined co-expressed transcriptomes, i.e., messenger RNA (mRNA) and miRNA studies for bovine muscle in response to treatment with RSV, which aims to investigate the roles of RSV in inducing miRNA for the better understanding, to identify the characteristics of DE miRNAs in the key co-expressed module in regulation of mRNAs, and to reveal new candidate regulatory miRNAs and genes underlying the beef quality traits.

MATERIALS AND METHODS

Primary Bovine Myoblast, Transcriptome Sequencing, and Differential Expression Analysis

The cultured primary myoblasts from the fetal beef *longissimus dorsi* muscle, the transcriptome sequencing datasets after quality control and alignment, and the differential expression analysis results were achieved from our previous studies (31, 32).

All the animal procedures were carried out according to the protocols approved by the Institutional Animal Care and Use Committee (IACUC) of the College of Animal Science and Technology, Northwest A&F University, China. Ninety-day-old fetal cattle were collected from Tumen slaughterhouse in Xi'an, Shaanxi Province. First, we used 75% alcohol and 1% double-antibody sterile phosphate buffer solution (phosphate buffered saline, PBS) to gently wash the epidermis of the fetal cow to eliminate blood stains and bacterial contamination in an extra-large plate. Second, we cut the muscle tissue pieces and placed them in a 50-ml centrifuge tube, with collagenase I digestion in Dulbecco's modified Eagle's medium (DMEM) at 37°C for 1.5 h. Third, the suspension was filtered and centrifuged, and the supernatant was removed. Fourth, we added four times the volume of 0.25% trypsin in the sediment at 37°C for 30 min. Fifth, the digested sample was filtered through 1-mm stainless steel mesh and 100-μm mesh. Sixth, we added 500 μl of a medium containing 15% fetal bovine serum (FBS) to the pellet to terminate the digestion, and the cells were inoculated in a 6-cm Petri dish. After 20-min culturing, we drew the upper culture medium and continued the culturing process in a new 6-cm Petri dish. When the cell density reached 80–90%, we used them for the subsequent experiments.

RNA for each cell sample was isolated for mRNA sequencing to generate 125 or 150-bp paired-end reads and for miRNA sequencing to generate 50-bp single-end reads on an Illumina HiSeq platform (Illumina, USA). We removed the unqualified reads by quality control to achieve clean reads by in-house *perl* scripts that were used in our previous studies (31, 32). Then, the miRNA tags were mapped to the reference genome of *Bos taurus* (UMD_3.1.1/bosTau8) by Bowtie software (version 0.12.9) (33). The mapped miRNA tags were used to seek the known miRNAs using miRBase20.0 as the reference, so the potential miRNA and the secondary structures were obtained by miRDeep2 software (version 2.0.0.5) (34).

The expected number of fragments per kilobase of transcript sequence per million base pairs sequenced (FPKM) of each gene was calculated to achieve the average FPKM values for the replicates. The threshold of gene expression was set for mRNA, when the FPKM value is larger than 1. Following the normalization formula (35), miRNA expression levels were also estimated by transcript per million (TPM). A differential expression analysis of two groups (case and control) for both mRNA and miRNA was performed using the R package DESeq (version 1.18.0) (36). The *P*-values were adjusted using the Benjamini and Hochberg's method for controlling the false discovery rate (FDR). Differentially expressed mRNAs and DE miRNAs were defined when the adjusted *P*-values were <0.05 . In addition, we calculated fold changes (FCs) between the case and control groups based on the averaged FPKM values and TPM values to define the up-regulated ($\log_2\text{FC} > 0$) and down-regulated ($\log_2\text{FC} < 0$) mRNAs and miRNAs, respectively.

We used the skeletal muscle cells under the polyphenol RSV treatment as the case group, while the skeletal muscle cells without RSV treatment were considered as the control group. Meanwhile, both the case and control groups had three independent experiments separately for the skeletal muscle cell collections. The correlation coefficients among samples were visualized in the heatmaps using $\log_{10}(\text{FPKM} + 1)$ for mRNA and $\log_{10}(\text{TPM} + 1)$ for miRNA in **Figure 1**.

Gene Co-expression Network of mRNA and miRNA and Their Associations With RSV Treatment

The R package WGCNA (37) was used to construct the co-expression network. It constructs a similarity matrix by Pearson correlation coefficients to measure the similarity between the gene expression profiles and then transforms the similarity matrix into an adjacency matrix (*A*) raised to a β exponent (soft threshold) based on the free-scale topology model. In this study, a total of 18,329 genes were filtered from 26,332 genes in mRNA data based on the median absolute deviation (MAD) of each gene bigger than 0.01. The β power parameter (soft threshold) was equal to 12 when the R^2 of the free-scale topology was equal to 0.8 (**Figure 2A**). In the miRNA data, 650 miRNAs were filtered from 765 genes, and the β power parameter (soft threshold) was equal to 4 (**Figure 2B**).

We chose the soft threshold power ($\beta = 12$ for mRNA and $\beta = 4$ for miRNA) based on the criterion of approximate scale-free topology to construct a weighted gene network and detect the consensus modules with the topological overlap matrix (TOM). The minimum module size was set at 30 for both miRNA data and mRNA data, and the maximum module sizes were set at 18,329 and 650 for mRNA data and miRNA data, respectively. Based on the dissimilarity between module eigengenes (MEs), the modules can be merged, where the first principal component of each module represents the gene expression profiles within the modules (38). Here, we set the cut height for module merging at 0.25, so the modules whose eigengenes are correlated above 0.75 will be merged.

A module association analysis was conducted between the ME and the RSV treatment status (i.e., 0, 0, 0, 1, 1, 1 for three control and three case groups, respectively) to calculate the correlations for the relevant module identifications. We calculated the module significance (MS) [i.e., the average absolute gene significance (GS) of all the genes involved in the module, where GS is measured as $\log P$ -value in the linear regression between gene expression and RSV treatment status] to evaluate the correlation strengths. Normally, the module with the highest MS score is the key module (37). Module significance genes in the association analysis (*P*-value < 0.1) were assigned for functional enrichment analysis. The hub genes were defined as the TOM values up to 0.8.

Gene Ontology and Pathway Enrichment Analysis

R package *clusterProfiler* (version 3.6) (39) was used to test the Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichments. Significant enrichment was defined with the adjusted *P*-value < 0.1 for both GO terms and pathways.

Predicted Target mRNAs of miRNA and mRNA–miRNA Networks

The bovine genomic sequence (release-99) and the gene annotation file were downloaded from the Ensembl FTP site (<http://www.ensembl.org/index.html>). We used TBtools to obtain the 3' UTR sequence of the bovine genomic transcripts (40). Then, the transcript stable IDs were converted to the Ensembl stable IDs using the BioMart website (<http://www.ensembl.org/biomart>). Accordingly, we found the 3' UTR sequence of the genes in the turquoise modules. The binding capability of miRNAs and their target genes in the turquoise modules was assessed by RNAhybrid (version 2.1.2) (41), with the minimal free energy hybridization under -20 and the helix constraint from 8 to 12.

RESULTS

Differentially Expressed mRNAs and miRNAs Between the RSV Treatment and Control Groups

From our previous study (31, 32), a total of 3,856 DE mRNAs were identified from 18,329 mRNAs based on the threshold of adjusted *P*-value < 0.05 ; meanwhile, 93 DE miRNAs were also identified from 650 miRNAs based on the same thresholds (**Table 1**). The details of $\log_2\text{FC}$, *P*-value, adjusted *P*-value, and DE mRNAs with FPKM and miRNAs with TPM of each sample are listed in the **Supplemental File 1**.

Module Identification of the Gene Co-expression Network for mRNA and miRNA and Their Associations With RSV Treatment

Using 18,329 mRNA and 650 miRNA data for the sample clustering, we found that the samples with RSV treatment were clustered together in mRNA analysis (**Figure 3A**), while

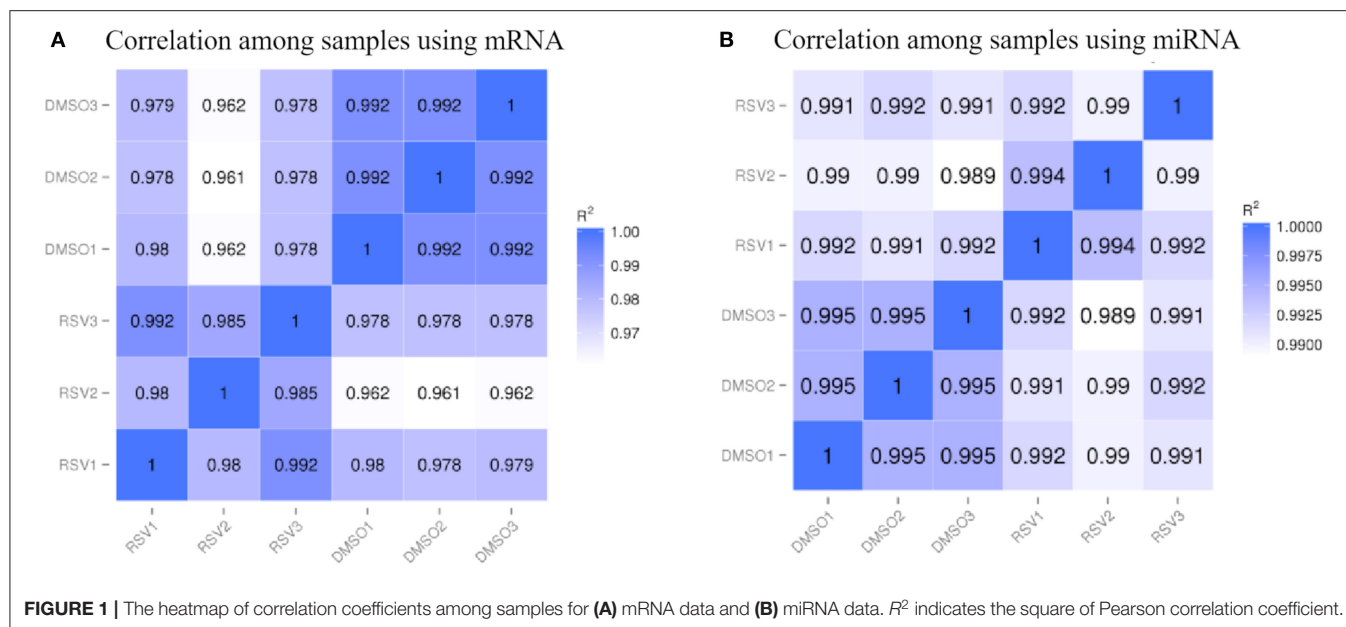


FIGURE 1 | The heatmap of correlation coefficients among samples for (A) mRNA data and (B) miRNA data. R^2 indicates the square of Pearson correlation coefficient.

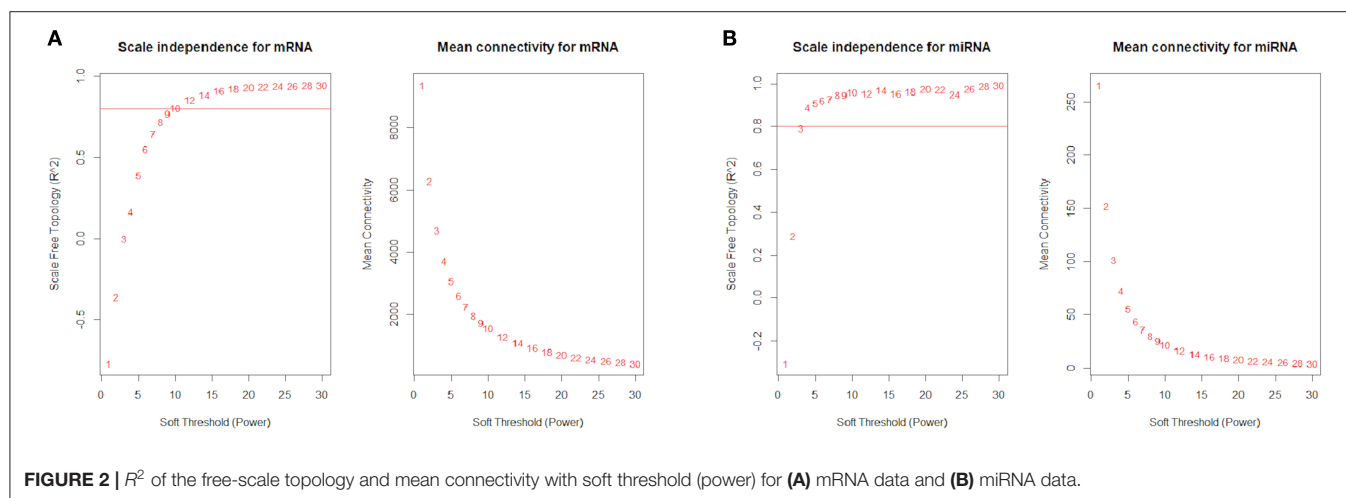


FIGURE 2 | R^2 of the free-scale topology and mean connectivity with soft threshold (power) for (A) mRNA data and (B) miRNA data.

TABLE 1 | Summary of differentially expressed mRNAs and miRNAs.

	DE mRNA (adjusted P-value <0.05)	DE mRNA (adjusted P-value <0.0001)	DE miRNA (adjusted P-value <0.05)
Up-regulated	1,805	450	44
Down-regulated	2,051	681	49
Total	3,856	1,131	93

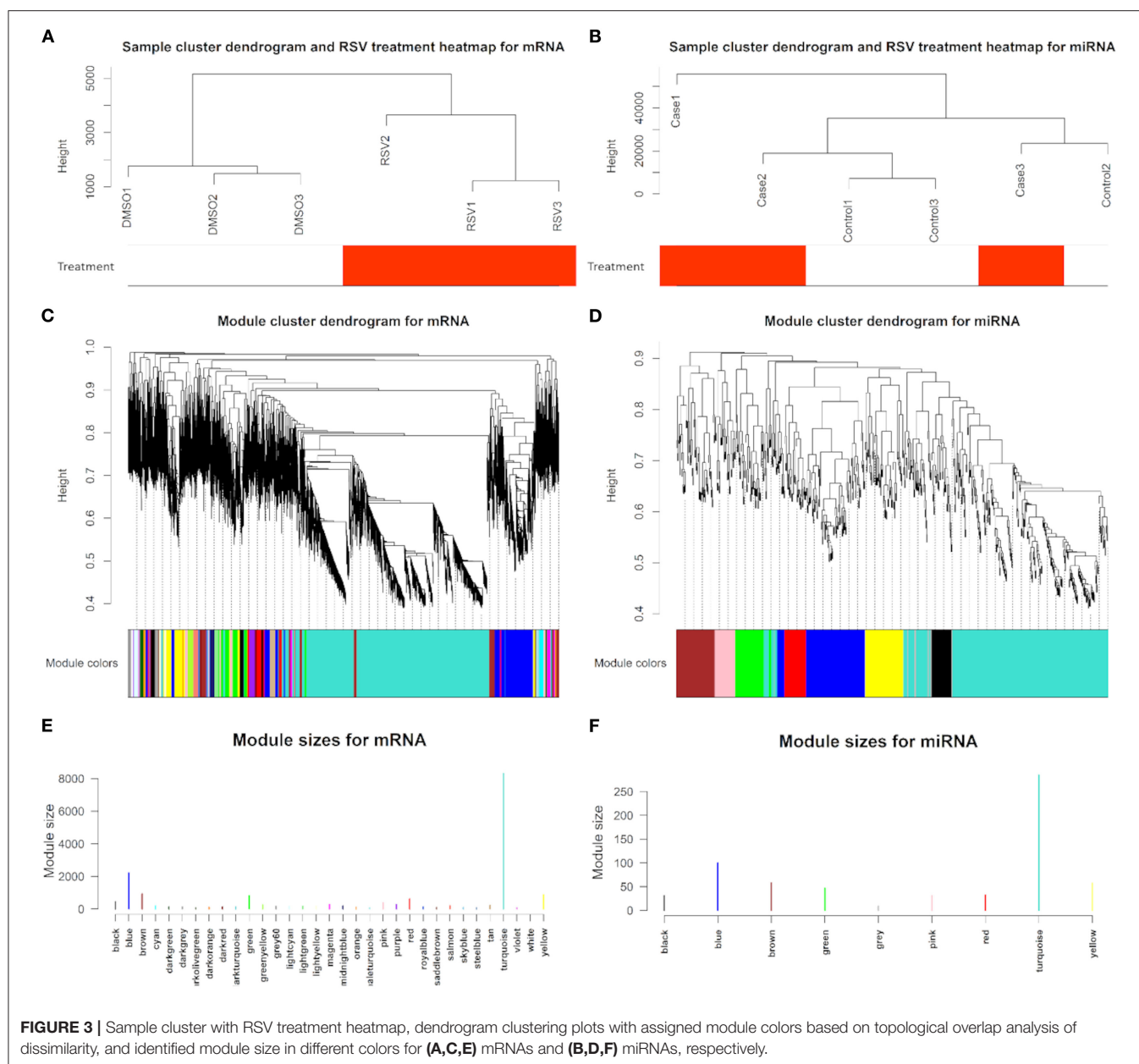
Differentially expressed (DE) mRNA and miRNA results were achieved from our previous study (31, 32).

RSV-treated samples were not clustered together by miRNA data (Figure 3B). Generally, 18,329 mRNAs were grouped into 32 modules that had similar co-expressions using the average linkage hierarchical clustering algorithm (Figures 3C,E), where

8,311 mRNAs were grouped into turquoise module as the key module, followed by 2,210 mRNAs into blue module, etc. (Figure 3E; Table 2). However, miRNAs were only grouped into eight modules (Figures 3D,F; Table 2), where 285 miRNAs were grouped into turquoise modules as the key module, followed by 100 miRNAs into blue modules. The mRNAs and miRNAs that were not assigned to any modules were grouped into gray modules (Figures 3C–F).

The eigengene adjacency heatmap indicated that these modules of mRNAs and miRNAs could be clustered further together into groups (Figure 4). After incorporating the RSV treatment trait, we found that the treatment status was clustered with blue and turquoise modules for mRNAs (Figure 4A) and with turquoise modules in a single cluster for miRNAs (Figure 4B).

The module–trait relationship results revealed that RSV treatment had high correlations with turquoise module (0.91,



P -value = 0.01), blue module (0.93, P -value < 0.01), and tan module (-0.81 , P -value = 0.05) using mRNA data (Figure 5A), whereas only turquoise module (0.96, P -value < 0.01) was highly correlated with treatment status using miRNA data (Figure 5B). Therefore, the turquoise module showed a strongly positive relationship with RSV treatment no matter whether mRNA data or miRNA data is used.

We also used the 3,856 DE mRNAs and 93 DE miRNAs for the weighted gene network construction based on the same soft threshold power ($\beta = 12$ for mRNA and $\beta = 4$ for miRNA) and then visualized them in the TOM clusters (Figure 6). Here, the minimum module size and maximum block size were set at 30 and 3,856, respectively, for mRNA and were set at 30 and

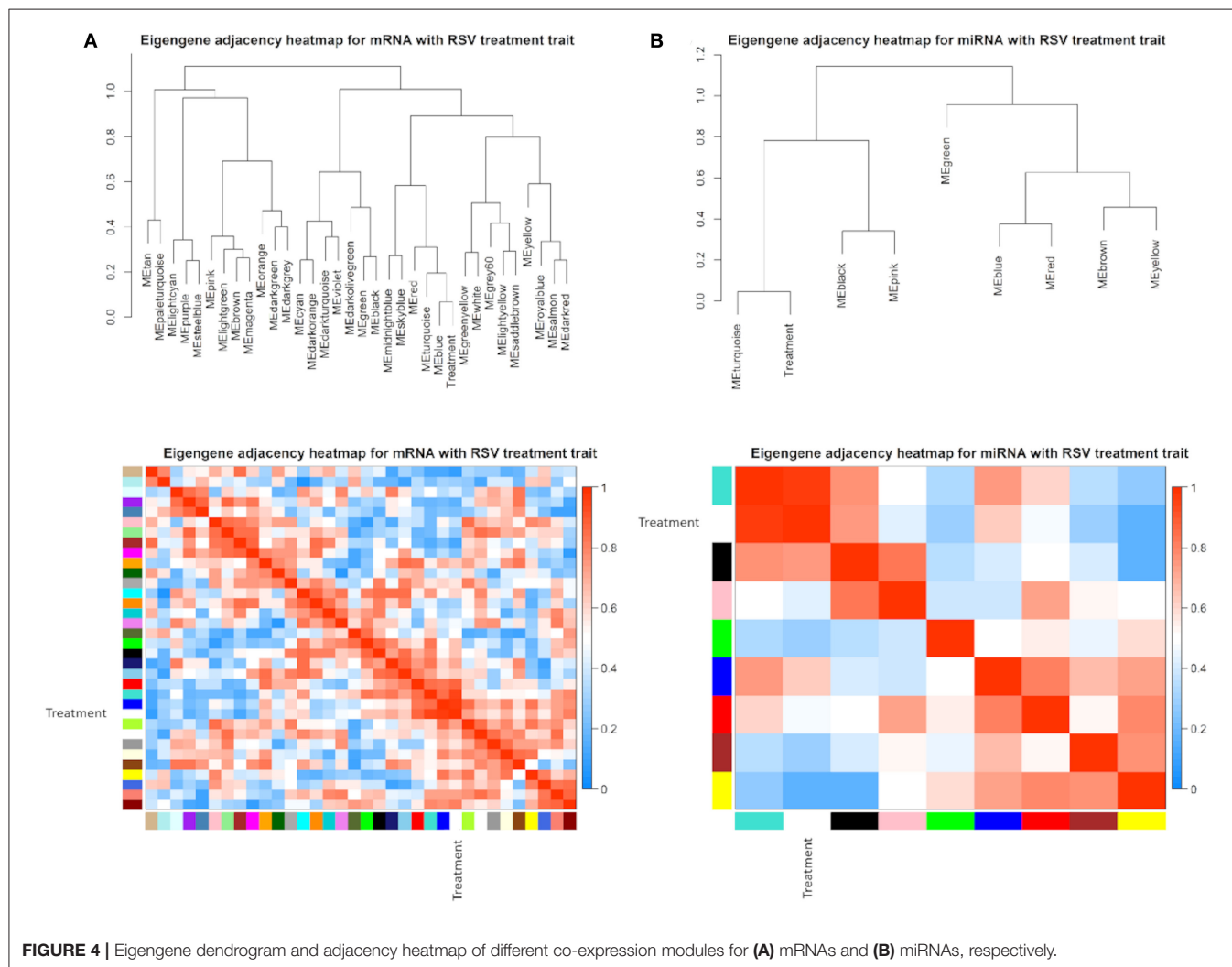
93, respectively, for miRNA. Only two cluster modules were displayed for both mRNAs and miRNAs, i.e., turquoise and blue modules. In the turquoise module, 2,579 DE mRNAs and 59 DE miRNAs were found, while 1,277 DE mRNAs and 34 DE miRNAs were found in the blue module. Furthermore, the network heatmap of DE mRNAs and DE miRNAs showed a high level of overlap densities among the two clusters (Figure 6).

Gene Ontology and Pathway Enrichment Analysis

Based on the 2,579 DE mRNAs in the turquoise module, we performed an enrichment analysis to reveal significant GO terms and KEGG pathways. We found that the three most significant

TABLE 2 | Summary of mRNA and miRNA module identifications.

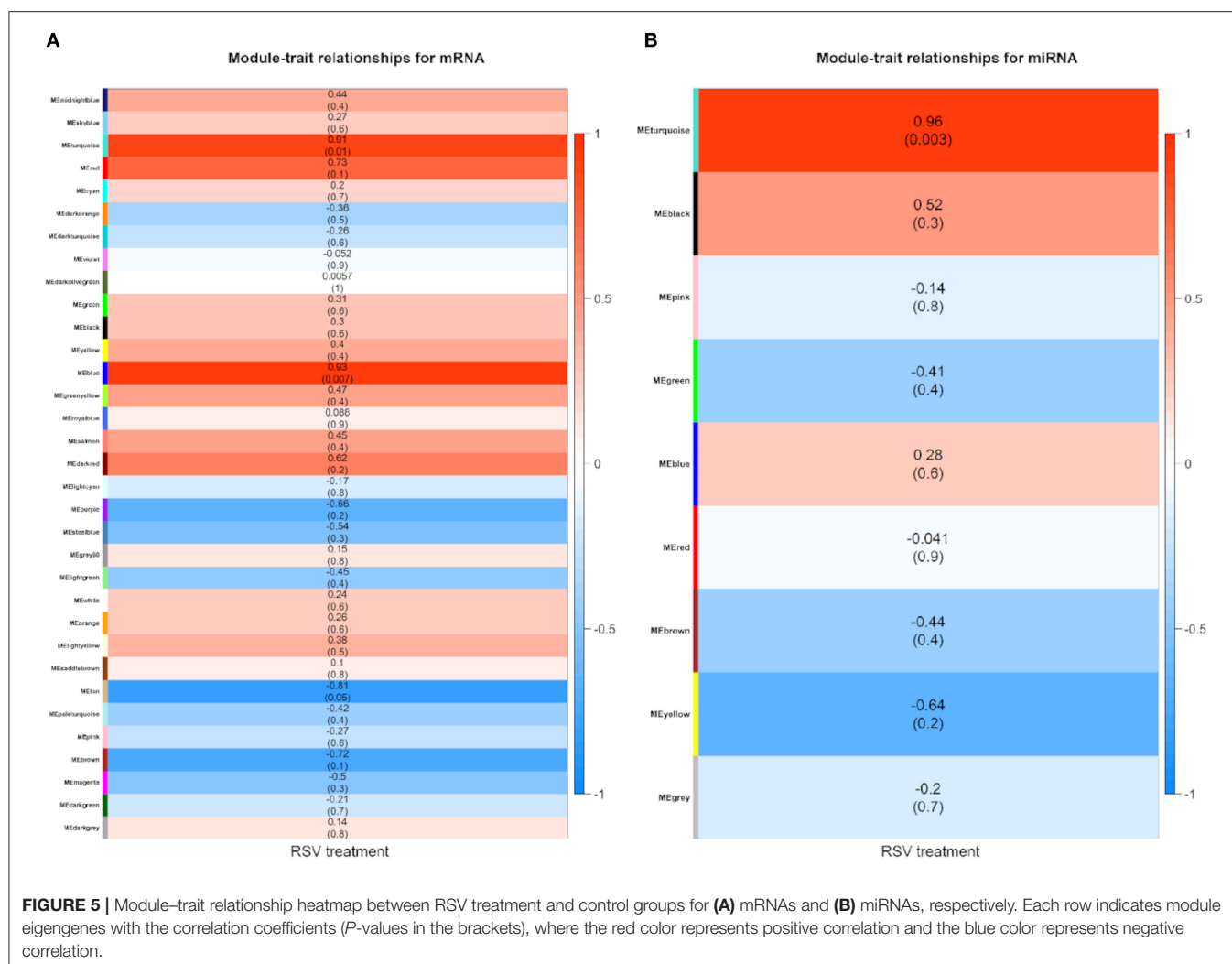
Transcriptome sequencing	Modules					
	Turquoise	Blue	Brown	Yellow	Green	Red
mRNAs ($n = 18,329$)	8,311 (45.34%)	2,210 (12.06%)	925 (5.05%)	886 (4.83%)	824 (4.50%)	613 (3.34%)
miRNAs ($n = 650$)	285 (43.85%)	100 (15.38%)	58 (8.92%)	57 (8.77%)	47 (7.23%)	32 (4.92%)

**FIGURE 4** | Eigengene dendrogram and adjacency heatmap of different co-expression modules for (A) mRNAs and (B) miRNAs, respectively.

GO terms were actin filament-based process (GO:0030029, adjusted P -value = 1.86×10^{-6}) with 37 DE genes that were enriched in, followed by actin cytoskeleton organization (GO:0030036, adjusted P -value = 4.33×10^{-6}) with 34 DE genes, and actin filament organization (GO:0007015, adjusted P -value = 6.86×10^{-5}) with 24 DE genes in the down-regulated category (Figure 7A). Similarly, the three most significant pathways were regulation of actin cytoskeleton (bta04810, adjusted P -value = 1.40×10^{-11}) with 47 genes, tight junction (bta04530, adjusted P -value = 2.32×10^{-6}) with 33 genes, and axon guidance (bta04360, adjusted P -value = 2.32×10^{-6}) with 33 genes (Figure 7B).

Networks Displaying the Relationships Among Genes Within Co-expressed Modules

We selected and constructed the network of four genes within the turquoise module including two significantly up-regulated genes, i.e., sushi domain containing 4 (*SUSD4*) gene and diacylglycerol O-acyltransferase 1 (*DGAT1*) gene, and two significantly down-regulated genes, i.e., fibroblast growth factor 18 (*FGF18*) gene and actin gamma 1 (*ACTG1*) gene (Figure 8). *SUSD4* has associations with 186 genes including 8 up-regulated genes and 176 down-regulated genes (Figure 8A). *DGAT1* was closely connected with



70 genes where 32 genes were up-regulated (e.g., proliferating cell nuclear antigen, *PCNA*) and 38 genes were down-regulated (e.g., phosphatase and tensin homolog, *PTEN*) (Figure 8B). *FGF18* and *ACTG1* were negatively connected with 36 and 51 up-regulated genes, respectively (Figures 8C,D). The selected hub gene *ACTG1* was negatively related with skeletal muscle myosin heavy chain 3 (*MYH3*) gene, fibroblast growth factor 2 (*FGF2*) gene, and uncoupling protein-2 (*UCP2*) (Figure 8D).

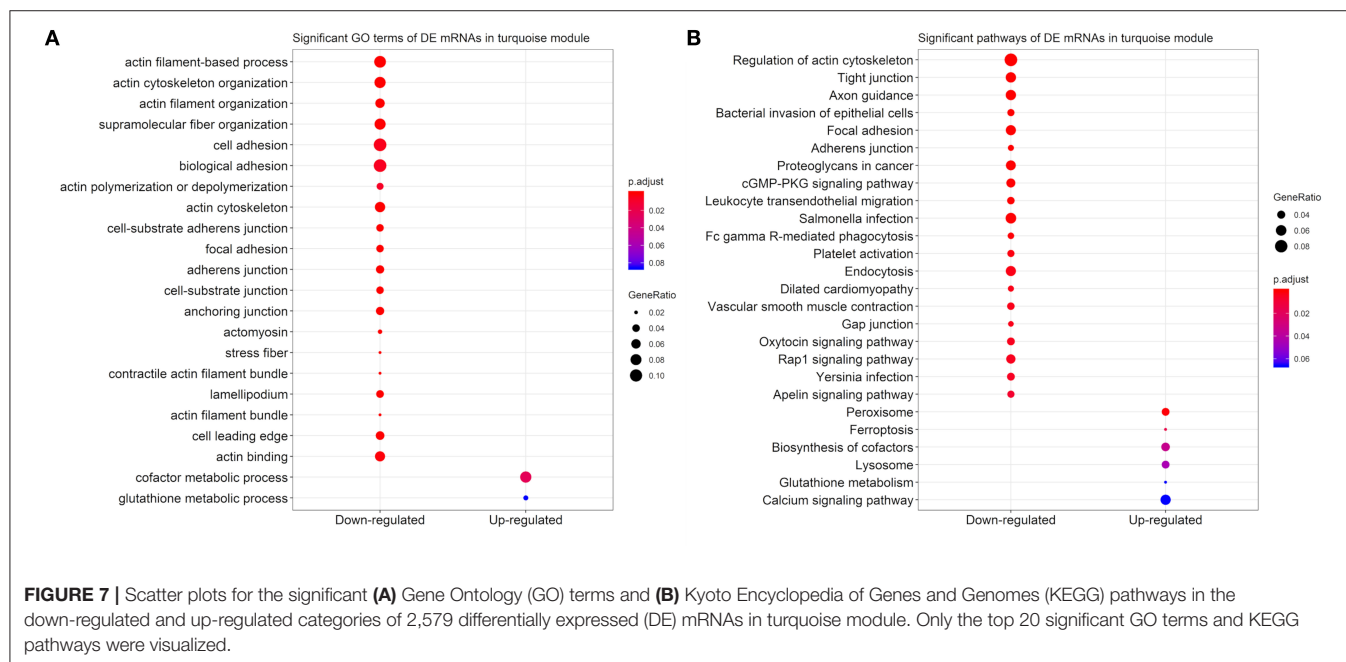
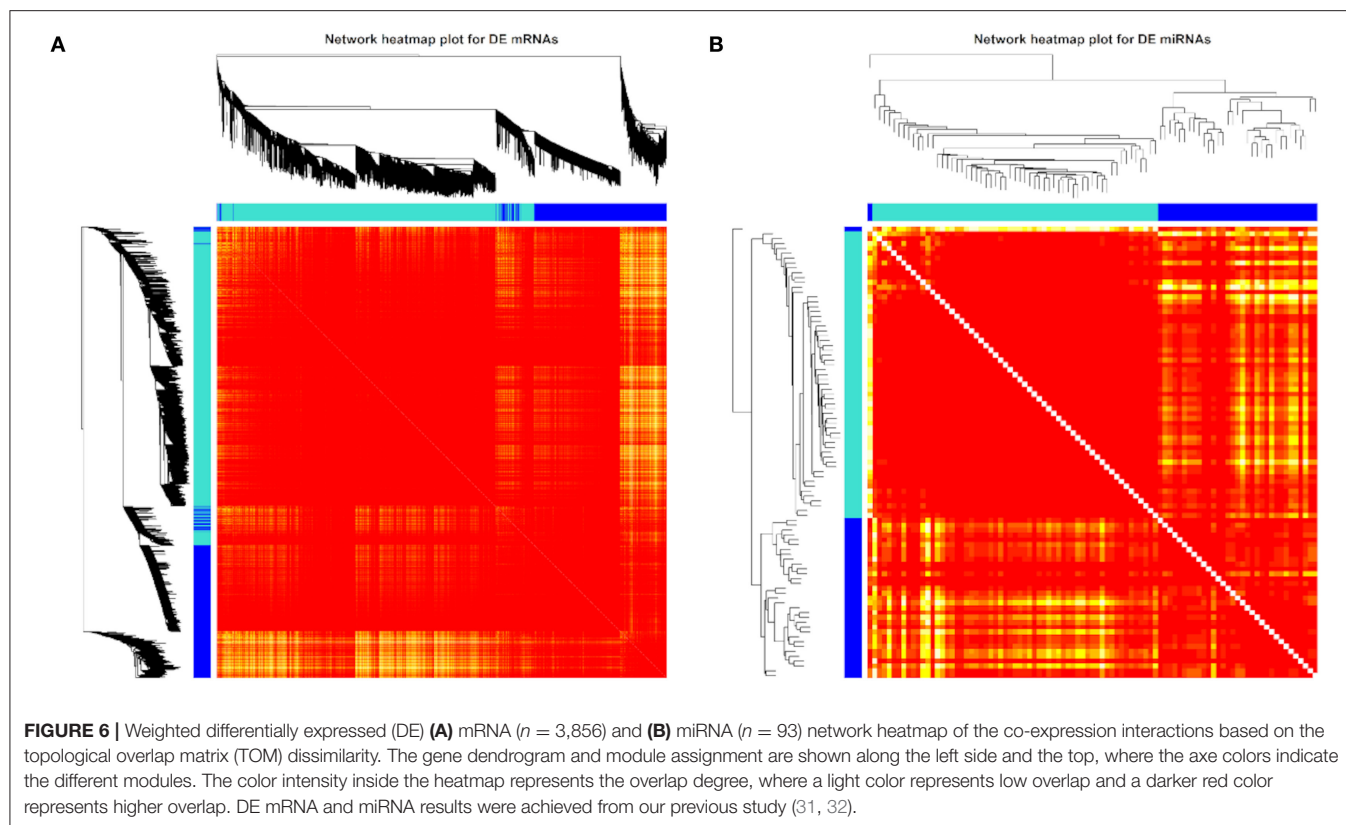
Identification of miRNAs in the Key Module and Target Gene Prediction of the miRNAs for mRNA–miRNA Network

Based on the top significant RSV-induced up- and down-regulated DE miRNAs in the turquoise module, we predicted their target genes. The top four up-regulated miRNAs were bta-miR-34c, bta-miR-432, bta-miR-2344, and bta-miR-154c that targeted 21, 78, 22, and 49 down-regulated genes in the turquoise module, respectively (Figure 9A). Likewise, the top four down-regulated miRNAs, i.e., bta-miR-2310, bta-miR-452, bta-miR-1814c, and bta-miR-199b targeted 59, 62, 58, and 15 up-regulated

genes in the turquoise module, respectively, where bta-miR-2310 and bta-miR-1814c targeted the same genes ($n = 57$) (Figure 9B). Three up-regulated and three down-regulated miRNAs targeted the top up-regulated *CDKN1A* [adjusted *P*-value = 2.50×10^{-104} and $\log_2(\text{FC}) = 1.97$], while five up-regulated and one down-regulated miRNAs targeted the top down-regulated *KCNK12* [adjusted *P*-value = 5.92×10^{-63} and $\log_2(\text{FC}) = -2.32$] (Figure 9C).

DISCUSSION

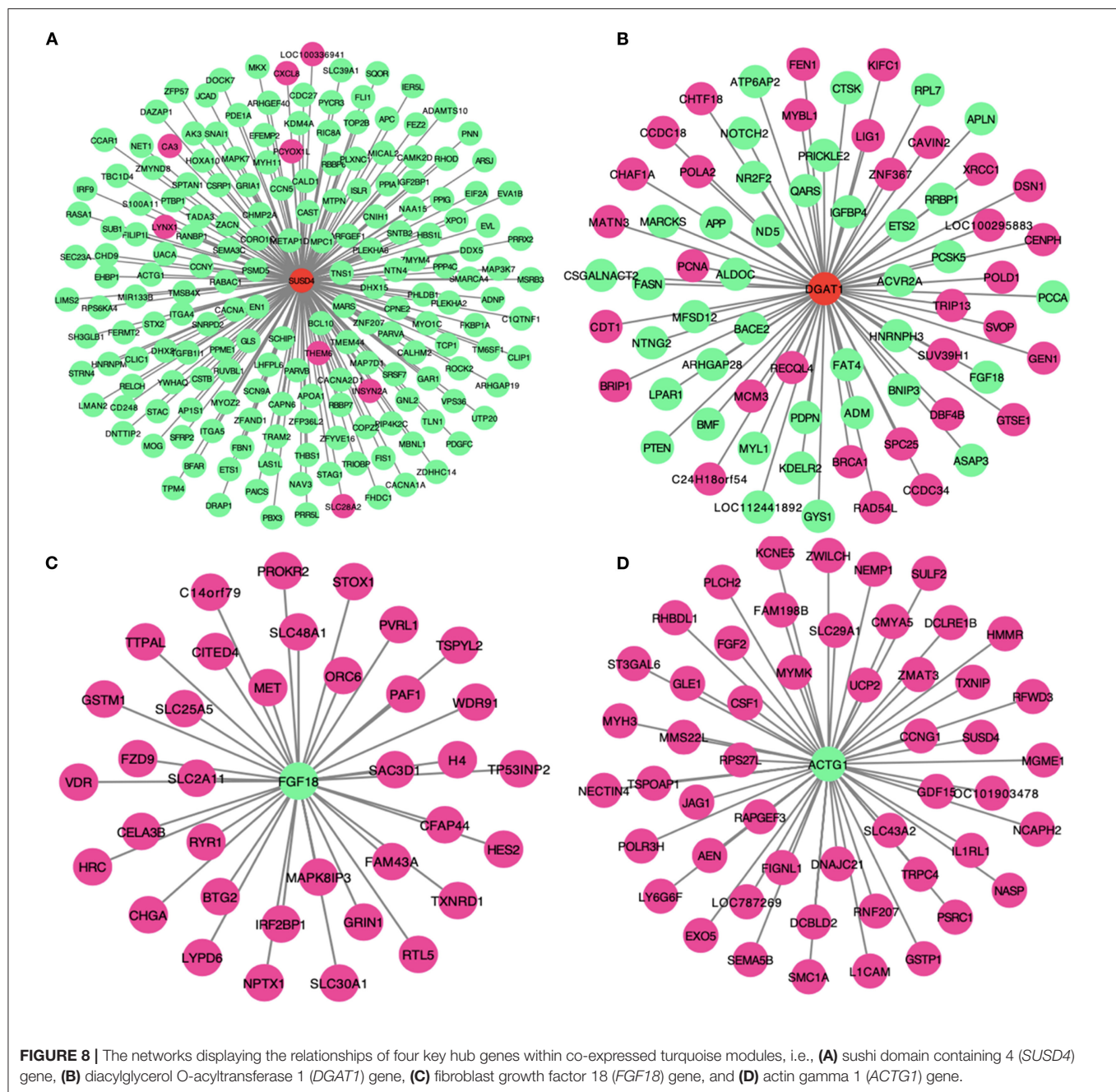
Carcass weight is mainly influenced by the number and the size of myoblasts that are generated from somite and through proliferation, differentiation, and fusion into myofibers in an embryonic stage (42). Skeletal muscle fiber characteristics can be divided into fast and slow types based on the contraction speed that can determine the meat quality traits such as marbling (43). Internal (heredity) and external factors (nutrition and environment) are combined to regulate the conversions among the fiber types, such as arginine (44) and linoleic



acid (45) that have been considered as the nutrients that influence the conversion of skeletal fiber type. Researchers also analyzed the omics data to reveal the effects of functional feed additives to improve carcass characteristics and beef quality, such as vitamin A, zinc propionate, etc. (46, 47). Our study

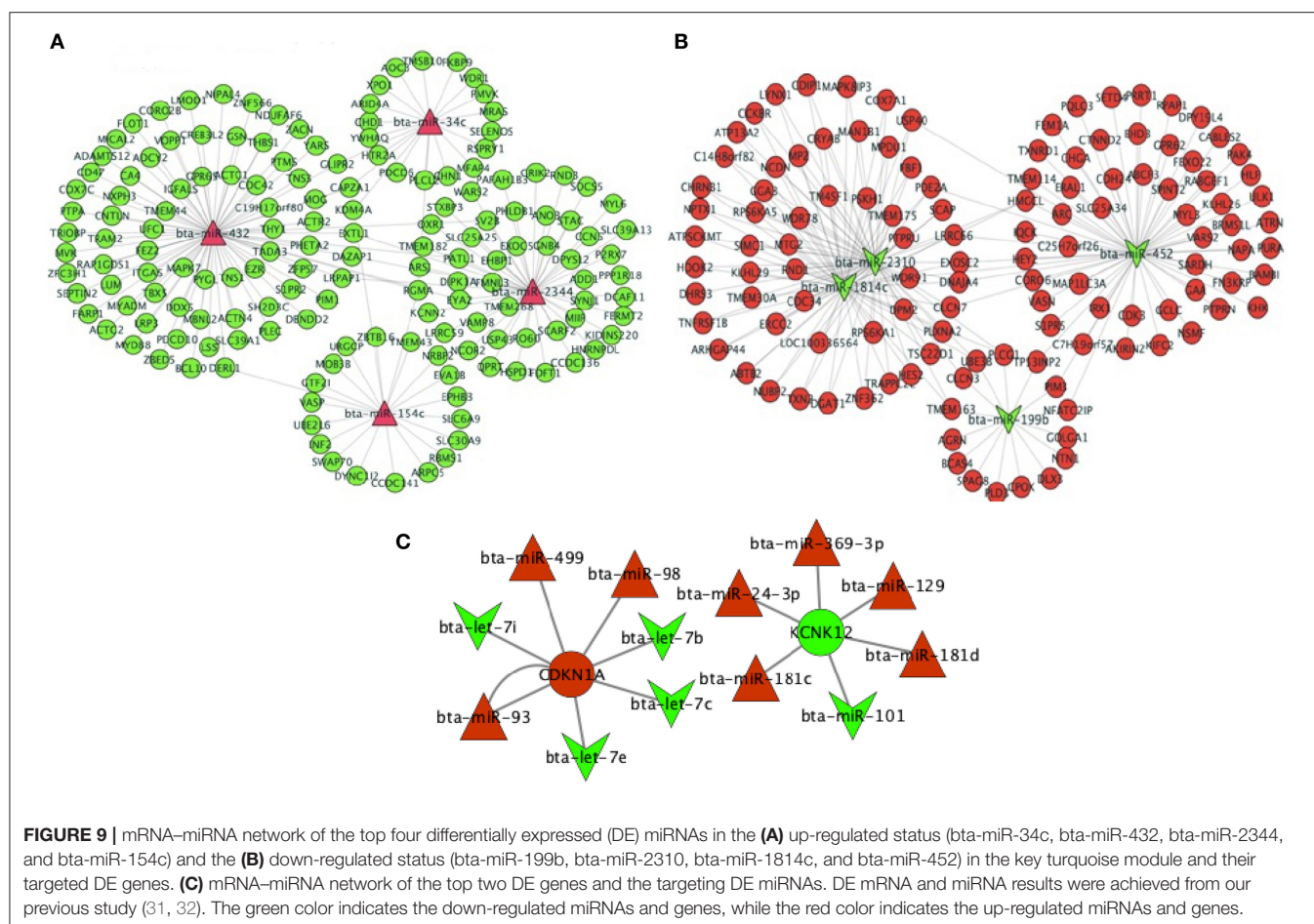
aims to illustrate the additive effects of natural polyphenol compound RSV on primary bovine myoblast differentiation through transcriptome sequencing.

Resveratrol effects have been extensively studied on various cell types including cardiomyoblasts (8), fibroblasts (7),



hepatocytes (48), smooth muscle cells (49), mammary epithelial cells (50), immune cells (6), and a multitude of various cancer cell lines (3). Moreover, researchers paid more attention to the RSV functions on human myoblast proliferation, injury, and death, as well as the resistance against oxidative stress of RSV to the skeletal muscle tissues in livestock (4, 11, 17, 20, 31). In this study, we identified biologically relevant key modules of the co-expressed mRNA and miRNA networks for the differentiation of bovine primary myoblast after RSV treatment. The results showed that the significant modules of DE mRNAs and DE miRNAs were strongly associated with RSV treatment status;

for example, the key mRNA module was turquoise module with 2,579 DE mRNAs induced by RSV, where 47 down-regulated DE genes were enriched in the regulation of actin cytoskeleton pathway (Figure 7B). The actin cytoskeleton is essential for cell proliferation, differentiation, migration, phagocytosis, and exocytosis, even when cells experience oxidative damage (51). The actin cytoskeleton is also essential to maintaining the stability of skeletal muscle functions. The absence of *ACTG1* resulted in muscle weakness and a progressive myopathy in mice (52); meanwhile, the reduced expression of *ACTG1* was closely associated with up-regulated *MYH3* induced by RSV



(Figure 8D). *MYH3* plays an important role in the skeletal muscle metabolism and the content of distinct types of skeletal muscle fibers (53). Thus, we suggested that RSV had functions on the type switch of the primary bovine myoblast fiber, which was consistent with the previous studies of RSV effects on C2C12 cells (54, 55).

Skeletal muscle development is elaborately regulated by myogenic regulatory factors (MRFs), growth factors (e.g., TGF- β and IGFs), signal pathways (e.g., IGF1-Akt-mTOR and Smad2/3 pathway), and non-coding RNAs (miRNAs) (56–58). MiRNAs play important roles in regulating myogenesis and regeneration, hypertrophy and atrophy, muscle disease, and aging (59, 60). In recent years, the identified functions of miRNAs in skeletal muscle development have been widely studied in cattle. Our study also identified 59 DE miRNAs in the turquoise module including bta-miR-432 and bta-miR-365-3p. They were highly expressed in skeletal muscle tissues and differently expressed in the fetal and adult stages of Qinchuan cattle; they may participate in the myoblast differentiation with vital roles (61, 62).

Network construction that integrates miRNA and mRNA data to identify the complex transcriptional regulating mechanism of RSV is more significant to the biological pathways for primary bovine myoblast than a separate analysis. This study

found that *ACTG1* could be targeted by the significantly RSV-induced up-regulated bta-miR-432 that potentially activated the IGF2/AKT signaling pathway to promote the proliferation and differentiation of myoblasts (61) (Figure 9A). Interestingly, the RSV-induced down-regulated bta-miR-2310 and bta-miR-1814c could co-target 57 RSV-induced up-regulated DE genes, such as *DGAT1* that is the functional candidate gene for the improvement of meat and carcass fatness quality in beef cattle (63) (Figure 9B). *DGAT1* was also positively connected with the myoblast proliferation gene (*PCNA*) and negatively related with *PTEN*, which regulated the skeletal satellite cell proliferation and differentiation (64) (Figure 9B). The RSV-induced miRNAs participated in the complex co-expression networks to regulate primary myoblast differentiation through affecting mRNA expressions, which subsequently participated in regulating skeletal development, metabolism, and skeletal muscle fiber type switch, thereby functionally regulating the cattle carcass weight and meat quality.

CONCLUSIONS

In summary, RSV treatments had high correlations with the turquoise module (0.91, P -value = 0.01) and blue module

(0.93, P -value <0.01) using mRNA data, but only had high correlations with the turquoise module (0.96, P -value <0.01) using miRNA data. The two top GO terms of actin filament-based process (GO:0030029) and actin cytoskeleton organization (GO:0030036) and the two top KEGG pathways of regulation of actin cytoskeleton (bta04810) and tight junction (bta04530) were revealed using 2,579 DE genes in the turquoise module. The mRNA–miRNA network was then constructed based on the co-expressions of DE mRNA and miRNA in the key module. Our study provided a better understanding of the roles of RSV in inducing miRNA and of the characteristics of DE miRNAs in the key co-expressed module in regulation of mRNAs and revealed new candidate regulatory miRNAs and genes for beef quality traits.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GEO, GSE186730.

ETHICS STATEMENT

All the animal procedures were carried out according to the protocols approved by the College of Animal Science and Technology, Northwest A&F University, China. All the experimental animals were approved by the Institutional Animal Care and Use Committee in the College of Animal Science and Technology, Northwest A&F University, China.

REFERENCES

- Albani D, Polito L, Signorini A, Forloni G. Neuroprotective properties of resveratrol in different neurodegenerative disorders. *BioFactors*. (2010) 36:370–6. doi: 10.1002/biof.118
- Hung LM, Chen JK, Huang SS, Lee RS, Su MJ. Cardioprotective effect of resveratrol, a natural antioxidant derived from grapes. *Cardiovasc Res*. (2000) 47:549–55. doi: 10.1016/S0008-6363(00)00102-4
- Lin JN, Lin VCH, Rau KM, Shieh PC, Kuo DH, Shieh JC, et al. Resveratrol modulates tumor cell proliferation and protein translation via SIRT1-dependent AMPK activation. *J Agric Food Chem*. (2010) 58:1584–92. doi: 10.1021/jf9035782
- Ahmed ST, Hossain ME, Kim GM, Hwang JA, Ji H, Yang CJ. Effects of resveratrol and essential oils on growth performance, immunity, digestibility and fecal microbial shedding in challenged piglets. *Asian Aust J Anim Sci*. (2013) 26:683–90. doi: 10.5713/ajas.2012.12683
- Das S, Das DK. Anti-inflammatory responses of resveratrol. *Inflamm Allergy Drug Targets*. (2007) 6:168–73. doi: 10.2174/187152807781696464
- Zhou ZX, Mou SF, Chen XQ, Gong LL, Ge WS. Anti-inflammatory activity of resveratrol prevents inflammation by inhibiting NF- κ B in animal models of acute pharyngitis. *Mol Med Rep*. (2018) 17:189–93. doi: 10.3892/mmr.2017.7933
- Fagone E, Conte E, Gili E, Fruciano M, Pistorio MP, Lo Furno D, et al. Resveratrol inhibits transforming growth factor- β -induced proliferation and differentiation of *ex vivo* human lung fibroblasts into myofibroblasts through ERK/Akt inhibition and PTEN restoration. *Exp Lung Res*. (2011) 37:162–74. doi: 10.3109/01902148.2010.524722
- Leong CW, Wong CH, Lao SC, Leong EC, Lao IF, Law PTW, et al. Effect of resveratrol on proliferation and differentiation of

AUTHOR CONTRIBUTIONS

DH performed all the experiments. DH and XW analyzed the data and wrote the manuscript. XW, YY, BT, and L-EH improved the manuscript. HC and BH conceived and designed the experiments. KQ collected experimental samples. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 31772574), the Program of National Beef Cattle and Yak Industrial Technology System (CARS-37), the Program of Yunling Scholar and the Young and Middle-aged Academic Technology Leader Backup Talent Cultivation Program in Yunnan Province, China (No. 2018HB045), the Yunnan Provincial Major S&T Project (2019ZG007 and 2019ZG011), and the Doctoral Startup Project of Chuxiong Normal University (No. BSQD2101).

ACKNOWLEDGMENTS

DH appreciates the scholarship from the China Scholarship Council, China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fvets.2021.777477/full#supplementary-material>

- embryonic cardiomyoblasts. *Biochem Biophys Res Commun*. (2007) 360:173–80. doi: 10.1016/j.bbrc.2007.06.025
- Kaminski J, Lançon A, Aires V, Limagne E, Tili E, Michaille JJ, et al. Resveratrol initiates differentiation of mouse skeletal muscle-derived C2C12 myoblasts. *Biochem Pharmacol*. (2012) 84:1251–9. doi: 10.1016/j.bcp.2012.08.023
- Dirks Naylor AJ. Cellular effects of resveratrol in skeletal muscle. *Life Sci*. (2009) 84:637–40. doi: 10.1016/j.lfs.2009.02.011
- Alway SE, McCrory JL, Kearcher K, Vickers A, Frear B, Gilleland DL, et al. Resveratrol enhances exercise-induced cellular and functional adaptations of skeletal muscle in older men and women. *J Gerontol A Biol Sci Med Sci*. (2017) 72:1595–606. doi: 10.1093/gerona/glx089
- Joseph AM, Malamo AG, Silvestre J, Wawrzyniak N, Carey-Love S, Nguyen LMD, et al. Short-term caloric restriction, resveratrol, or combined treatment regimens initiated in late-life alter mitochondrial protein expression profiles in a fiber-type specific manner in aged animals. *Exp Gerontol*. (2013) 48:858–68. doi: 10.1016/j.exger.2013.05.061
- Wang D, Sun H, Song G, Yang Y, Zou X, Han P, et al. Resveratrol improves muscle atrophy by modulating mitochondrial quality control in STZ-induced diabetic mice. *Mol Nutr Food Res*. (2018) 62:1–11. doi: 10.1002/mnfr.201700941
- Dolinsky VW, Jones KE, Sidhu RS, Haykowsky M, Czubyrt MP, Gordon T, et al. Improvements in skeletal muscle strength and cardiac function induced by resveratrol during exercise training contribute to enhanced exercise performance in rats. *J Physiol*. (2012) 590:2783–99. doi: 10.1113/jphysiol.2012.230490
- Zhao JX, Li Q, Zhang RX, Liu WZ, Ren YS, Zhang CX, et al. Effect of dietary grape pomace on growth performance, meat quality and antioxidant activity in ram lambs. *Anim Feed Sci Technol*. (2018) 236:76–85. doi: 10.1016/j.anifeedsci.2017.12.004

16. Zhang C, Zhao X, Wang L, Yang L, Chen X, Geng Z. Resveratrol beneficially affects meat quality of heat-stressed broilers which is associated with changes in muscle antioxidant status. *Anim Sci J.* (2017) 88:1569–74. doi: 10.1111/asj.12812
17. Zhang C, Wang L, Zhao XH, Chen XY, Yang L, Geng ZY. Dietary resveratrol supplementation prevents transport-stress-impaired meat quality of broilers through maintaining muscle energy metabolism and antioxidant status. *Poult Sci.* (2017) 96:2219–25. doi: 10.3382/ps/pex004
18. Cheng K, Wang T, Li S, Song Z, Zhang H, Zhang L, et al. Effects of early resveratrol intervention on skeletal muscle mitochondrial function and redox status in neonatal piglets with or without intrauterine growth retardation. *Oxid Med Cell Longev.* (2020) 2020:1–12. doi: 10.1155/2020/4858975
19. Cheng K, Yu C, Li Z, Li S, Yan E, Song Z, et al. Resveratrol improves meat quality, muscular antioxidant capacity, lipid metabolism and fiber type composition of intrauterine growth retarded pigs. *Meat Sci.* (2020) 170:108237. doi: 10.1016/j.meatsci.2020.108237
20. Yu Q, Fang C, Ma Y, He S, Ajuwon KM, He J. Dietary resveratrol supplement improves carcass traits and meat quality of Pekin ducks. *Poult Sci.* (2021) 100802. doi: 10.1016/j.psj.2020.10.056
21. Wang F, Tian X, Zhang L, He C, Ji P, Li Y, et al. Beneficial effect of resveratrol on bovine oocyte maturation and subsequent embryonic development after in vitro fertilization. *Fertil Steril.* (2014) 101:577–86. doi: 10.1016/j.fertnstert.2013.10.041
22. Li CY, Zhao YH, Hao HS, Wang HY, Huang JM, Yan CL, et al. Resveratrol significantly improves the fertilisation capacity of bovine sex-sorted semen by inhibiting apoptosis and lipid peroxidation. *Sci Rep.* (2018) 8:7603. doi: 10.1038/s41598-018-25687-z
23. Ma T, Wu W, Tu Y, Zhang N, Diao Q. Resveratrol affects in vitro rumen fermentation, methane production and prokaryotic community composition in a time- and diet-specific manner. *Microb Biotechnol.* (2020) 13:1118–31. doi: 10.1111/1751-7915.13566
24. Hocquette JF, Botreau R, Picard B, Jacquet A, Pethick DW, Scollan ND. Opportunities for predicting and manipulating beef quality. *Meat Sci.* (2012) 92:197–209. doi: 10.1016/j.meatsci.2012.04.007
25. Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J. Functional network construction in arabidopsis using rule-based machine learning on large-scale data sets. *Plant Cell.* (2011) 23:101–16. doi: 10.1105/tpc.111.088153
26. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science.* (2003) 302:249–55. doi: 10.1126/science.1087447
27. Weirauch MT. Gene coexpression networks for the analysis of DNA microarray data. In: Dehmer M, Emmert-Streib F, Graber A, Salvador A, editors. *Applied Statistics for Network Biology: Methods in Systems Biology.* New York, NY: John Wiley & Sons (2011). p. 215–50. doi: 10.1002/9783527638079.ch11
28. De Oliveira PSN, Coutinho LL, Tizioto PC, Cesar ASM, de Oliveira GB, Diniz WJ, et al. An integrative transcriptome analysis indicates regulatory mRNA-miRNA networks for residual feed intake in Nelore cattle. *Sci Rep.* (2018) 8:17072. doi: 10.1038/s41598-018-35315-5
29. Novais FJ, Pires PRL, Alexandre PA, Dromms RA, Iglesias AH, Ferraz JBS, et al. Identification of a metabolomic signature associated with feed efficiency in beef cattle. *BMC Genomics.* (2019) 20:1–10. doi: 10.1186/s12864-018-5406-2
30. Fan Y, Arbab AAI, Zhang H, Yang Y, Nazar M, Han Z, et al. Lactation associated genes revealed in holstein dairy cows by weighted gene co-expression network analysis (WGCNA). *Animals.* (2021) 11:1–16. doi: 10.3390/ani11020314
31. Hao D, Wang X, Wang X, Thomsen B, Kadarmideen HN, Lan X, et al. Transcriptomic changes in bovine skeletal muscle cells after resveratrol treatment. *Gene.* (2020) 754:144849. doi: 10.1016/j.gene.2020.144849
32. Hao D, Wang X, Wang X, Thomsen B, Qu K, Lan X, et al. Resveratrol stimulates microRNA expression during differentiation of bovine primary myoblasts. *Food Nutr Res.* (2021) 65:1–9. doi: 10.29219/fnr.v65.5453
33. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* (2009) 10:R25. doi: 10.1186/gb-2009-10-3-r25
34. Friedländer MR, MacKowiak SD, Li N, Chen W, Rajewsky N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* (2012) 40:37–52. doi: 10.1093/nar/gkr688
35. Zhou L, Chen J, Li Z, Li X, Hu X, Huang Y, et al. Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq273 associate with clear cell renal cell carcinoma. *PLoS ONE.* (2010). 5:e15224. doi: 10.1371/journal.pone.0015224
36. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* (2010) 11:R106. doi: 10.1186/gb-2010-11-10-r106
37. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics.* (2008) 9:559. doi: 10.1186/1471-2105-9-559
38. Silva-Vignato B, Coutinho LL, Poleti MD, Cesar ASM, Moncau CT, Regitano LCA, et al. Gene co-expression networks associated with carcass traits reveal new pathways for muscle and fat deposition in Nelore cattle 06 Biological Sciences 0604 Genetics. *BMC Genomics.* (2019) 20:32. doi: 10.1186/s12864-018-5345-y
39. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J Integr Biol.* (2012) 16:284–7. doi: 10.1089/omi.2011.0118
40. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant.* (2020) 13:1194–202. doi: 10.1016/j.molp.2020.06.009
41. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA.* (2004) 10:1507–17. doi: 10.1261/rna.5248604
42. Mueck T, Berger F, Buechler I, Valchanova RS, Landuzzi L, Lollini PL, et al. TRAF6 regulates proliferation and differentiation of skeletal myoblasts. *Differentiation.* (2011) 81:99–106. doi: 10.1016/j.diff.2010.11.002
43. Hwang YH, Kim GD, Jeong JY, Hur SJ, Joo ST. The relationship between muscle fiber characteristics and meat quality traits of highly marbled Hanwoo (Korean native cattle) steers. *Meat Sci.* (2010) 86:456–61. doi: 10.1016/j.meatsci.2010.05.034
44. Chen X, Guo Y, Jia G, Liu G, Zhao H, Huang Z. Arginine promotes skeletal muscle fiber type transformation from fast-twitch to slow-twitch via Sirt1/AMPK pathway. *J Nutr Biochem.* (2018) 61:155–62. doi: 10.1016/j.jnutbio.2018.08.007
45. Lu ZQ, Ren Y, Zhou XH, Yu XF, Huang J, Yu DY, et al. Maternal dietary linoleic acid supplementation promotes muscle fibre type transformation in suckling piglets. *J Anim Physiol Anim Nutr.* (2017) 101:1130–6. doi: 10.1111/jpn.12626
46. Diniz WJS, Mazzoni G, Coutinho LL, Banerjee P, Geistlinger L, Cesar ASM, et al. Detection of co-expressed pathway modules associated with mineral concentration and meat quality in nelore cattle. *Front Genet.* (2019) 10:210. doi: 10.3389/fgene.2019.00210
47. Knutson EE, Menezes ACB, Sun X, Fontoura ABP, Liu JH, Bauer ML, et al. Effect of feeding a low-vitamin A diet on carcass and production characteristics of steers with a high or low propensity for marbling. *Animal.* (2020) 14:2308–14. doi: 10.1017/S1751731120001135
48. Gnoni GV, Paglialonga G. Resveratrol inhibits fatty acid and triacylglycerol synthesis in rat hepatocytes. *Eur J Clin Invest.* (2009) 39:211–8. doi: 10.1111/j.1365-2362.2008.02077.x
49. Li H, Xia N, Hasselwander S, Daiber A. Resveratrol and vascular function. *Int J Mol Sci.* (2019) 20:1–16. doi: 10.3390/ijms20092155
50. Licznarska B, Szafer H, Wierchowski M, Sobierajska H, Baer-Dubowska W. Resveratrol and its methoxy derivatives modulate the expression of estrogen metabolism enzymes in breast epithelial cells by AhR down-regulation. *Mol Cell Biochem.* (2017) 19:1907–14. doi: 10.1007/s11010-016-2871-2
51. Davidson AJ, Wood W. Unravelling the actin cytoskeleton: a new competitive edge? *Trends Cell Biol.* (2016) 26:569–76. doi: 10.1016/j.tcb.2016.04.001
52. Rubenstein PA. The functional importance of multiple actin isoforms. *BioEssays.* (1990) 12:309–15. doi: 10.1002/bies.950120702
53. Cho IC, Park HB, Ahn JS, Han SH, Lee JB, Lim HT, et al. A functional regulatory variant of MYH3 influences muscle fiber-type composition and intramuscular fat content in pigs. *PLoS Genet.* (2019) 15:e1008279. doi: 10.1371/journal.pgen.1008279
54. Jiang Q, Cheng X, Cui Y, Xia Q, Yan X, Zhang M, et al. Resveratrol regulates skeletal muscle fibers switching through the AdipoR1-AMPK-PGC-1 α pathway. *Food Funct.* (2019) 10:3334–43. doi: 10.1039/c8fo02518e

55. Wen W, Chen X, Huang Z, Chen D, Chen H, Luo Y, et al. Resveratrol regulates muscle fiber type conversion via miR-22-3p and AMPK/SIRT1/PGC-1 α pathway. *J Nutr Biochem.* (2020) 77:108297. doi: 10.1016/j.jnutbio.2019.108297
56. Perry RL, Rudnick MA. Molecular mechanisms regulating myogenic determination and differentiation. *Front Biosci.* (2000) 5:d750–67. doi: 10.2741/a548
57. Goljanek-Whysall K, Sweetman D, Münsterberg AM. microRNAs in skeletal muscle differentiation and disease. *Clin Sci.* (2012) 123:611–25. doi: 10.1042/CS20110634
58. Schiaffino S, Dyar KA, Ciciliot S, Blaauw B, Sandri M. Mechanisms regulating skeletal muscle growth and atrophy. *FEBS J.* (2013) 280:4294–314. doi: 10.1111/febs.12253
59. Güller I, Russell AP. MicroRNAs in skeletal muscle: their role and regulation in development, disease and function. *J Physiol.* (2010) 588:4075–87. doi: 10.1113/jphysiol.2010.194175
60. Zheng Y, Kong J, Li Q, Wang Y, Li J. Role of miRNAs in skeletal muscle aging. *Clin Interv Aging.* (2018) 13:2407–19. doi: 10.2147/CIA.S169202
61. Wang X, Cao X, Dong D, Shen X, Cheng J, Jiang R, et al. Circular RNA TTN acts as a miR-432 sponge to facilitate proliferation and differentiation of myoblasts via the IGF2/PI3K/AKT signaling pathway. *Mol Ther Nucl Acids.* (2019) 18:966–80. doi: 10.1016/j.omtn.2019.10.019
62. Hao D, Wang X, Wang X, Thomsen B, Yang Y, Lan X, et al. MicroRNA bta-miR-365-3p inhibits proliferation but promotes differentiation of primary bovine myoblasts by targeting the activin A receptor type I. *J Anim Sci Biotechnol.* (2021) 12:16. doi: 10.1186/s40104-020-00528-0
63. Li X, Ekerljung M, Lundström K, Lundén A. Association of polymorphisms at DGAT1, leptin, SCD1, CAPN1 and CAST genes with color, marbling and water holding capacity in meat from beef cattle populations in Sweden. *Meat Sci.* (2013) 94:153–8. doi: 10.1016/j.meatsci.2013.01.010
64. Yue F, Bi P, Wang C, Shan T, Nie Y, Ratliff TL, et al. Pten is necessary for the quiescence and maintenance of adult muscle stem cells. *Nat Commun.* (2017) 8:14328. doi: 10.1038/ncomms14328

Conflict of Interest: XW was employed by company Konge Larsen ApS.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Hao, Wang, Yang, Thomsen, Holm, Qu, Huang and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Characterization of XR_311113.2 as a MicroRNA Sponge for Pre-ovulatory Ovarian Follicles of Goats *via* Long Noncoding RNA Profile and Bioinformatics Analysis

Hu Tao¹, Juan Yang¹, Pengpeng Zhang², Nian Zhang¹, Xiaojun Suo¹, Xiaofeng Li¹, Yang Liu¹ and Mingxin Chen^{1*}

¹Hubei Key Laboratory of Animal Embryo Engineering and Molecular Breeding, Institute of Animal Husbandry and Veterinary, Hubei Academy of Agricultural Sciences, Wuhan, China, ²Department of Biotechnology, College of Life Sciences, Xinyang Normal University, Xinyang, China

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Xin Qi,
Ocean University of China, China
Shahin Eghbalsaeed,
Islamic Azad University, Isfahan, Iran

*Correspondence:

Mingxin Chen
cmx1963@163.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 18 August 2021

Accepted: 10 December 2021

Published: 03 January 2022

Citation:

Tao H, Yang J, Zhang P, Zhang N,
Suo X, Li X, Liu Y and Chen M (2022)
Characterization of XR_311113.2 as a
MicroRNA Sponge for Pre-ovulatory
Ovarian Follicles of Goats *via* Long
Noncoding RNA Profile and
Bioinformatics Analysis.
Front. Genet. 12:760416.
doi: 10.3389/fgene.2021.760416

Long noncoding RNAs (lncRNAs) were identified recently as a large class of noncoding RNAs (ncRNAs) with a length ≥ 200 base pairs (bp). The function and mechanism of lncRNAs have been reported in a growing number of species and tissues. In contrast, the regulatory mechanism of lncRNAs in the goat reproductive system has rarely been reported. In the present study, we sequenced and analyzed the lncRNAs using bioinformatics to identify their expression profiles. As a result, 895 lncRNAs were predicted in the pre-ovulatory ovarian follicles of goats. Eighty-eight lncRNAs were differentially expressed in the Macheng black goat when compared with Boer goat. In addition, the lncRNA XR_311113.2 acted as a sponge of chi-miR-424-5p, as assessed via a luciferase activity assay. Taken together, our findings demonstrate that lncRNAs have potential effects in the ovarian follicles of goats and may represent a promising new research field to understand follicular development.

Keywords: goat, ovarian follicle, lncRNA, transcriptome, miRNA sponge

INTRODUCTION

Ovulation rate is an important goat reproductive trait that determines the upper limit of the female goat litter size; in turn, the growth and development of follicles in the ovary determine the ovulation rate (Cui et al., 2009). Follicle development is a very complex biological process and its regulatory mechanism warrants further study. During mammalian ovarian folliculogenesis, which involves multiple transcription factors, primordial follicles develop into pre-ovulatory follicles, followed by ovulation, which releases mature oocytes (Choi and Rajkovic, 2006; Pan et al., 2012). The Macheng black goat is an excellent goat breed that is unique to China and is characterized by pure black hair color and its breed function is higher than Boer goat (Tao et al., 2018). In the breeding production process, PMSG-hCG is usually used to treat the ewe, and then artificial fertilization. We found that after hormone treatment, the litter size of Macheng black goat remains above Boer goat. To investigate the phenotypic differences, we analyze the pre-ovulatory follicles of ewes after hormone treatment via High-throughput sequencing and analyze the development of follicles in a highly fertile local goat.

Noncoding RNAs (ncRNAs) have always been regarded as “DNA junk” without the capability to encode proteins. More recently ncRNAs, which include mainly microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), and circular RNAs (circRNAs), were demonstrated to have important biological functions. They are basically derived from the transcription activity of the genome (Jin et al., 2018). Long noncoding RNAs (lncRNAs), which were recently identified as a large class of ncRNAs, are defined as ncRNAs of more than 200 bp in length with no protein-coding capacity (Mercer et al., 2009; Hung and Chang, 2010). A recent study showed that lncRNAs play multifunctional roles in a wide variety of important biological processes, including cell proliferation (Wang et al., 2017), cell apoptosis (Lu et al., 2016), genomic splicing (Parikshak et al., 2016), chromatin remodeling (Han and Chang, 2015), and transcription (McHugh et al., 2015). In addition, lncRNAs regulate the expression of miRNA by acting as molecular sponges and reducing the inhibitory effect of miRNAs on their target genes (Paraskevopoulou and Hatzigeorgiou, 2016).

During ovarian folliculogenesis, functional lncRNAs, such as growth arrest specific 5 (GAS5) (Wang et al., 2018), lncRNA HCG26 (Liu et al., 2017), lncRNA Neat1 (Nakagawa et al., 2014; Li et al., 2017), and lncRNA MALAT1 (Li et al., 2018a), have been found to regulate follicle development and regeneration through diverse mechanisms. For instance, GAS5, which is expressed in female germ-line stem cells (FGSCs) and oocytes, promotes FGSC reproduction and survival *in vitro* and is highly expressed in neonatal mouse ovaries (Wang et al., 2018). High levels of lncRNA SRA stimulate the growth of mouse follicle granulosa cells, increase the levels of estrogen and progesterone, and upregulate the expression of key enzymes (i.e., cytochrome P450 family 19, subfamily A, member 1 (CYP19A1) and cytochrome P450 family 11, subfamily A, member 1 (CYP11A1)) (Li et al., 2018b). Although miRNAs and piRNAs have been extensively studied in ovarian folliculogenesis, the role of other valuable noncoding transcripts remains to be studied. Moreover, researchers are paying increasing attention to the role of lncRNAs in follicular development, and have used high-throughput sequencing to analyze the expression pattern of lncRNAs during follicular development. Finally, the identification of the key lncRNAs for follicular development is expected.

In the present study, differentially expressed lncRNAs were detected through deep RNA sequencing (RNA-seq) in samples of pre-ovulatory follicles from Macheng black goats and Boer goats. Eight hundred and ninety-five new lncRNAs were identified in the ovarian follicles of goats. Eighty-eight lncRNAs were differentially expressed between Macheng black and Boer goats. The lncRNA expression patterns were validated using quantitative real-time polymerase chain reaction (qRT-PCR). The greatly differentially expressed lncRNA, XR_311113.2 was demonstrated to sponge chi-miR-424-5p *via* a luciferase activity assay. Therefore, in this study we examined the expression of lncRNAs by RNA-seq during goat follicle development and explored their expression profiles and functions in goat follicles.

MATERIALS AND METHODS

Ethics Statement

All studies involving animals were conducted according to the regulation (No. 5 proclaim of the Standing Committee of Hubei People's Congress) approved by the Standing Committee of Hubei People's Congress, P. R. China. Sample collection was approved by the ethics committee of Hubei Academy of Agricultural Sciences. Animals were humanely sacrificed as necessary to ameliorate suffering.

Animals and Tissues

Aged adult ewes weighing 40 kg were obtained from the goat stud farm of the Institute of Animal Husbandry and Veterinary Medicine, Hubei Academy of Agricultural Sciences, feeding with the same pattern and environment. Macheng black goats (Kidding rate and fecundity rate: 219 and 346%) have higher reproductive performance than Boer goats (Kidding and fecundity rates: 189 and 210%). Three Macheng black goats and three Boer goats exhibiting normal estrous cycles were treated with 1000 IU PMSG (SanSheng, Zhejiang, China) and 500 IU hCG (SanSheng, Zhejiang, China) as previously described (Saharrea et al., 1998; Torner et al., 1998). PMSG-hCG was used to initiate and synchronize the follicular phase. After 36 h, the PMSG-hCG stimulated ewes were slaughtered, and the ovaries of each goat in the pre-ovulatory phase were immediately removed, stored in liquid nitrogen until RNA extraction. The total RNAs were extracted from fourteen tissues (i.e., liver, spleen, heart, kidneys, small intestine, fat, ovarian follicle, lungs, uterus, abomasum, muscle, reticulum, rumen and omasum) of the Boer goats.

RNA Extraction and Qualification

A total of 5 µg of RNA was isolated from each individual sample using the TRIzol reagent (Invitrogen, MD, United States) according to the manufacturer's protocol. The purity and quantity of the total RNA were measured using a NanoDrop instrument and agarose gel electrophoresis. RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, United States), and only the samples showing RNA integrity number (RIN) scores higher than 6 were used in this study.

Library Preparation for lncRNA and mRNA Sequencing

Three micrograms (µg) of RNA was used as input material. Ribosomal RNA was completely removed using the Epicentre Ribo-zero™ rRNA Removal Kit (Epicentre, WI, United States), and then sequencing libraries were generated with the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (NEB, Ipswich, United States) following manufacturer's recommendations. First-strand cDNA was synthesized using M-MuLV Reverse transcriptase and random hexamer primers, and second-strand cDNA was synthesized using RNase H and DNA polymerase I. After adenylation of the 3' ends of DNA

fragments, NEB Next Adaptors were ligated to prepare for hybridization. The library fragments were purified for selecting cDNA fragments (150–200 bp in length). Finally, the products were purified with an AMPure XP system (Beckman Coulter, Beverly, United States). The libraries were sequenced on an Illumina HiSeq 4000 platform, and 150 bp paired-end reads were generated. Sequencing data were deposited with the NCBI Sequence Read Archive (SRA) under accession number PRJNA648013.

Quality Control and Transcriptome Assembly

Firstly, Raw reads were processed through in-house Perl scripts developed by the Novogene Bioinformatics Institute (Beijing, China). The clean data were obtained by removing reads that contain adapter or ploy-N and low-quality reads from raw data. Meanwhile, The Phred score (Q20, Q30) and GC content of the clean data were calculated. Reference genome and gene model annotation files were downloaded from National Center for Biotechnology Information (ARS1, <https://www.ncbi.nlm.nih.gov/genome/?term=goat>). Index of the reference genome was built by Bowtie v2.4.2 and paired-end clean reads were aligned to the reference genome using TopHat v2.1.1. The mapped reads of each sample were assembled by both Scripture (beta2) and Cufflinks (v2.2.1).

Coding Potential and Conservative Analysis

We used a Coding Potential Calculator (CPC, 0.9-r2) (Kong et al., 2007), Coding-Non-Coding-Index (CNCI, v2) (Sun et al., 2013) and PfamScan (v1.3) (Punta et al., 2012) to assess the protein-coding potential of each novel transcript. CPC examined sequences in NCBI eukaryotes' protein database and set the e-value "1e-10." Default parameters were used in CNCI profiles. Pfam searches use default parameters of -E 0.001 -domE 0.001. Candidate sets of lncRNAs that possessed no coding potential were chosen and filtered out from the predicted transcript results from the three tools listed.

Analysis of Differential Expression

Cuffdiff (v2.2.1) was used to estimate the fragment per kilobase of exon per million fragments mapped (FPKM) of the transcripts in each sample (Trapnell et al., 2012). The unit of measurement is FPKM. Cuffdiff calculated for the level of expression in each transcript. Gene and transcript expressions based on FPKM values were calculated by Cufflinks transcript quantification engine. Transcripts with a $p < 0.05$ were assigned as differentially expressed.

Target Gene and miRNA Binding Site Prediction of lncRNAs

To explore the function of candidate lncRNAs, the target genes of candidate lncRNAs were predicted in the cis way. Coding genes 100 k upstream and downstream of lncRNAs were examined as the target genes in cis (co-location genes). The co-expression of coding genes with lncRNAs in different chromosomes were determined with the Pearson correlation coefficient (PCC);

with $PCC > 0.95$ or < -0.95 , the lncRNA-mRNA pair was considered to represent the target genes in trans (co-expressed genes) (Schober et al., 2018). MiRNA binding sites were predicted for the goats with miRanda software, whose principles are based on the miRanda prediction algorithm (Enright et al., 2003). The miRNAs and target lncRNAs were considered when the miRanda score was 140 or higher, and the energy threshold was set to -1.

Analysis of lncRNA-miRNA Network Interactions

The lncRNA-miRNA interaction network was built according to the prediction of miRNA binding sites. The lncRNA-miRNA interaction analysis was conducted with Cytoscape software (Version 3.6.1). In the network diagram, the connections indicate possible regulatory relationships. The square represents lncRNAs, the circle represents miRNAs, the red represents up-regulated expression and the green represents down-regulated expression.

Quantitative real-time PCR Analysis

Total cDNA was synthesized using reverse transcriptase Kit (TaKaRa, Dalian, China). QRT-PCR were performed using CFX96 Touch™ Real-Time PCR Detection System and SYBR® Green PCR Supermix (Bio-Rad, CA, United States). Each PCR reaction (in 20 μ L) involved 10 μ L SYBR® Green PCR Supermix (Bio-Rad, CA, United States), 0.25 μ L of each primer, 1 μ L cDNA and 8.5 μ L H₂O. The cycling conditions included an initial single cycle (94°C for 3 min), and followed by 40 cycles (94°C for 30 s; 60°C for 30 s; 72°C for 20 s). The primers were designed using Primer 5 (shown in Table 1). The expression level was defined based on the threshold cycle (Ct), and relative expression levels were calculated via the $2^{-\Delta\Delta C_t}$ method. The correlation between the RNA-seq and qRT-PCR results was calculated using the SPSS (Version 18.0.0). β -actin was served as internal standard control, and all reactions were performed in triplicate.

Plasmid Construction

Two fragments of XR_311113.2 sequence were amplified using I-5™ 2 × High-Fidelity Master Mix (Tsingke, Wuhan, China). All the PCR products were inserted into pmirGLO vector (Promega, WI, United States), respectively. The primers are listed in Table 1. Then two recombinant plasmids were digested with *PmeI* and *NheI* (Thermo Scientific, WLM, United States). All constructs were sequenced by Sangon Biotech Co., Ltd. (Shanghai, China). The mutants of binding sites were generated using a MutanBEST Kit (TaKaRa, Dalian, China) and mutagenic primers (shown in Table 1).

Cell Culture, Cell Transfection, and Dual Luciferase Reporter Assays

The goat kidney epithelial cells (GK cells) were obtained from the Kunming cell bank of the Chinese academy of sciences. Cells were seeded at a density of 1.5×10^5 /ml using Dulbecco minimum essential medium (DMEM) (Hyclone, UT, United States) supplied with 10% fetal bovine serum (FBS) medium (Hyclone, UT, United States). The miRNA mimics and negative control (NC) were synthesized by Ribobio (Ribobio,

TABLE 1 | The nucleotide sequence used in this study.

ID	Forward sequence (5'–3')	Reverse sequence (5'–3')	Length (bp)
XR_001297559.1	TTGTA ^{CTCGTGGCCCTAAT}	CCAGGCTAATCCTCCAACC	150
XR_311113.2	TTGAGAAAACAGCCAGTGC	TACCGCCAGTGACAAGGAT	125
XR_001297560.1	TCTCATGCTAACCAGGACCC	AAAGCCACTGTAACCGCACC	150
LNC_000026	GCTGGAGTCTTA ^{CTTATTGGAT}	ATCAGAAAGGATGGGTGTG	148
XR_310768.2	AGGCTTCCTCCTGCTTGTG	ATCCGCATCATTTGTCCATT	279
LNC_000155	AGCCACAGTGAGCAGCATC	AAAGGGAGTCATAGAGTGGG	451
XR_001295597.1	ATGTTCTTCATCGGCTTCACC	CTCGTTCTTGTCTAGTCCCAC	177
β-actin	GTCACCAACTGGGACGACA	AGGCGTACAGGGACAGCA	208
chi-miR-424-5p mimics	CAGCAGCAAUUC ^{AUGUUUUGA}		
chi-miR-3955-5p mimics	UUUGAUGGCUGAUCCUCUCACU		
NC	UUCUCCGAACGUGUACGUTT	ACGUGACACGUUCGGAGAATT	
pmiRGLO-1	GGGTTTAAACCTTGGAGCATAGTCTACAGCA	CCGCTCGAGTCATACTGACCGACTTGTGC	1005
pmiRGLO-2	GGGTTTAAACAGAGAGGCATCCTTGTCACTG	CCGCTCGAGAATTC ^{CAACAGGCAATCGTT}	1525
chi-miR-135a mut	ACCCTTTGAGGGGGATTGCTCGGCTGAATCC	GGATTAGCCGAGCAATCCCCCTCAAAGGGT	1005
chi-miR-424-5p mut1	TGCCATATTGGGCATCATCATAGGGCCAAAG	CTTTGGCCCTATGATGATGCCCAATATGGCA	1005
chi-miR-424-5p mut2	AGGAGAGAAGAACATCATCTTGTCTATTGTAG	CTACAATGACAAGATGATGTTCTTCTCTCTCT	1525
chi-miR-544-5p mut	CTACCGGCTGGACGGTGGACCA ^{CCATCTCAG}	CTGAGATGGTGGTCCACCGTCCAGCCGGTACG	1005
chi-miR-3955-5p mut	CCCAGCAAGAGATTGCTGGTCA ^{CCCTCTGGG}	CCCAGAGGGTGACCAGCAATCTCTTGTCTGGG	1525

Note: The part highlighted with gray was enzyme site induced.

Guangzhou, China). The sequences used are shown in **Table 1**. Cells were seeded in 24-well plates at a density of 1.5×10^5 cells per well 24 h before transfection. The cells were co-transfected with a mixture of 250 ng pmirGLO recombinant vector plasmid and 3 μ L miRNA mimics or NC per wells using lipofectamine 3000 (Life Technologies, MD, United States). The luciferase activity in the transfected cells was then measured after 24 h according to the manufacturer's instructions with a VICTOR™ X3 PerkinElmer 2030 Multilabel Plate Reader (PerkinElmer, MA, United States).

Statistical Analyses

Descriptive results of the study are expressed as means \pm SD. All data analyses were performed by GraphPad Prism 5 (San Diego, CA). The means of the groups were compared by Student's t-test. $p < 0.05$ was considered to indicate statistically.

RESULTS

Identification of lncRNAs in Goat Pre-ovulatory Follicles

To identify the expression of lncRNAs in goat pre-ovulatory follicles, six goats average from two goat breeds, Boer and Macheng goats, were sequenced using an Illumina HiSeq 4000 platform. A total of 41.3 Gb of lncRNA clean data were obtained and mapped to the goat reference genome using TopHat2. We carried out a series of rigorous screening analyses using three analytical software programs (CPC, PFAM, and CNCI) (**Figures 1A,B**), and a total of 895 lncRNAs from 99,106 assembled transcripts were identified (**Figure 1C** and **Supplementary Table S1**). These lncRNAs consisted of 88.0% long intergenic noncoding RNAs (lincRNAs) and 12.0% antisense lncRNAs, with virtually no intronic lncRNAs being detected (**Figure 1D**). We analyzed the characteristics and expression levels of the lncRNA and mRNA transcripts. As shown in **Figure 1E**, we found that the

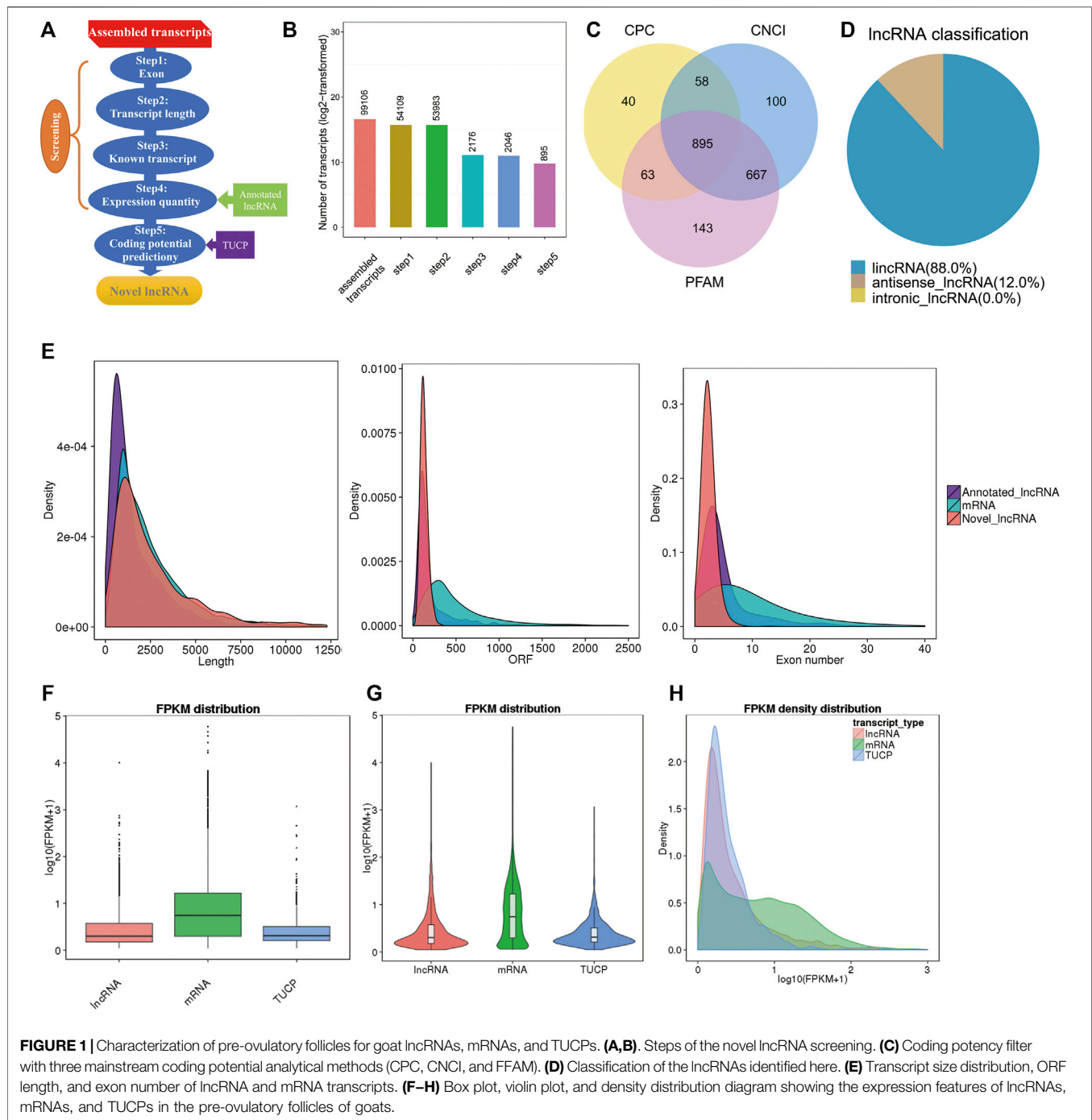
predicted lncRNAs were shorter in length than the mRNAs and had fewer exons than the average mRNA. Furthermore, most of the lncRNAs tended to exhibit a relatively shorter open reading frame (ORF) length than the mRNAs. The mRNA levels were significantly higher than those of lncRNAs or transcripts of uncertain coding potential (TUCP) in the pre-ovulatory follicles of goat samples (**Figures 1F–H**).

Expression Profiles of lncRNAs and mRNAs in Goat Follicles

Here, we identified the differential expression profiles of lncRNAs and mRNAs in the pre-ovulatory follicles of Boer and Macheng Black goats. Specifically, a total of 88 lncRNAs (67 upregulated and 21 downregulated) and 1216 mRNAs (743 upregulated and 473 downregulated) were differentially expressed (**Figures 2A,B** and **Supplementary Tables S2, S3**). A hierarchical clustering analysis was performed on the differentially expressed lncRNAs and mRNAs, respectively. To gain insight into the similarities of the pre-ovulatory follicles between the Boer and Macheng goats, data from all of the differentially expressed lncRNAs and mRNAs were used in a systematic cluster analysis. The heat map clearly indicated self-segregated clusters in the Boer (BO) and Macheng (MC) groups (**Figures 2C,D**).

Overview of Differential lncRNA Expression Profiles

Five upregulated (XR_001297559.1, XR_311113.2, XR_001297560.1, LNC_000026, and XR_310768.2) and two downregulated (LNC_000155 and XR_001295597.1) lncRNAs were randomly selected and amplified *via* qRT-PCR to confirm RNA-seq accuracy. As shown in **Figure 3A**, log2-fold changes (MC/BO) were calculated based on the RNA-seq and qRT-PCR results. The observed expression trends indicated that the two methods produced consistent results. Furthermore, the fold



change values measured by RNA-seq and qRT-PCR were significantly correlated (correlation coefficient, 0.944) (**Figure 3B**). Moreover, we investigated the relative expression levels of the greatly differentially expressed lncRNA XR_311113.2 in 14 goat tissues of Macheng black goat and Boer goat (i.e., liver, spleen, heart, kidneys, small intestine, fat, ovarian follicle, lungs, uterus, abomasum, muscle, reticulum, rumen, and omasum) (**Figure 3C**). XR_311113.2 tended to exhibit high expression levels in the liver, spleen, heart, and kidneys of Macheng black goat and Boer goat, followed by the small intestine,

fat, ovarian follicle, and lungs. By contrast, its low expression was noted in the uterus, abomasum, muscle, reticulum, rumen, and omasum. Notably, the expression level of XR_311113.2 in Macheng ovarian follicle was far greater than Boer, and the fold change was nearly 6. These results demonstrated that there was a general consistency between the qRT-PCR and RNA-seq results, although the fold changes were not exactly the same between the two different technologies. In addition, XR_311113.2 was only significantly dimorphically expressed in ovarian follicle.

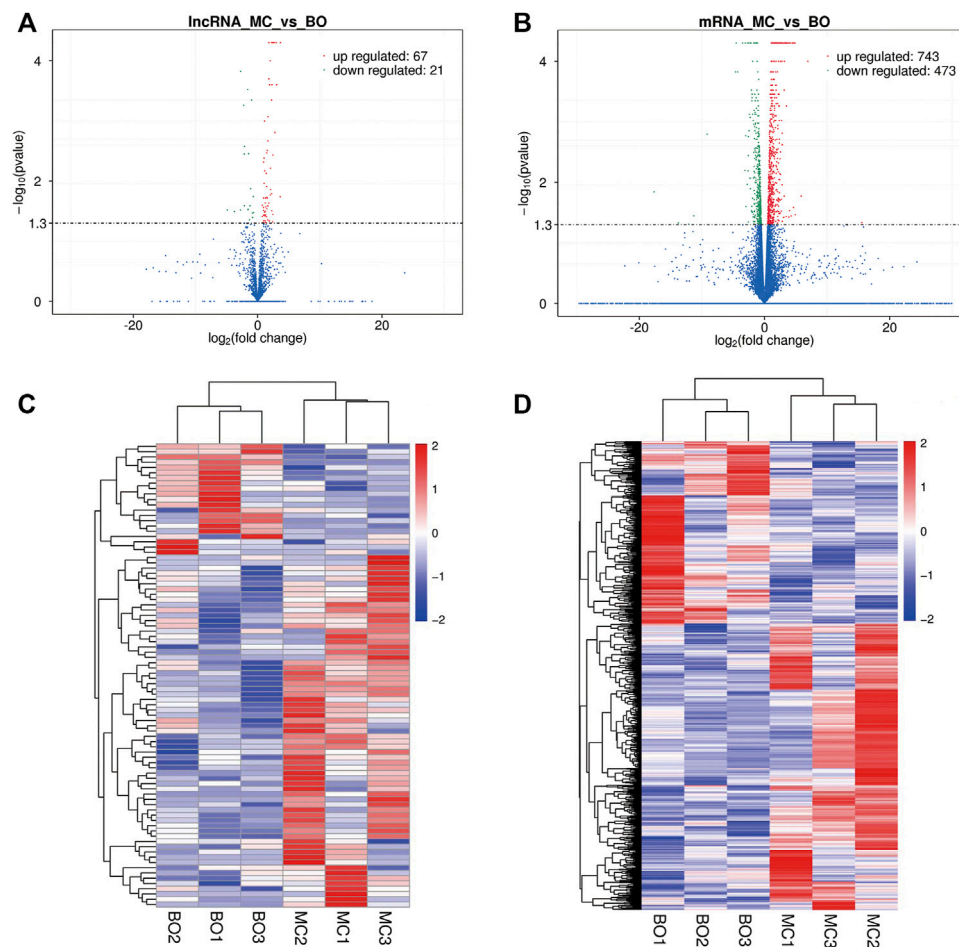


FIGURE 2 | Differential expression of lncRNAs and mRNAs in the pre-ovulatory follicles of goats. **(A,B)** Volcano plots of differentially expressed lncRNA and mRNA transcripts. **(C,D)** Hierarchical clustering of the expression profiles of differentially expressed lncRNAs and mRNAs.

lncRNA-miRNA Interaction Networks

The lncRNA-miRNA interaction networks were constructed using the RNA-seq data and potential lncRNA-miRNA connections were explored using the Cytoscape 3.6.1 software (<http://www.cytoscape.org/>). The lncRNA-miRNA interaction network was established based on the relationship between the differentially expressed (DE) lncRNAs and DE miRNAs. In the network depicted in **Figure 4A**, we identified the top 20 lncRNAs and top 20 miRNAs as differentially expressed profiles. The top differentially expressed lncRNA XR_311113.2 included binding sites for four miRNAs (chi-miR-135a, chi-miR-424-5p, chi-miR-544-5p, and chi-miR-3955-5p). The interaction network of XR_311113.2 and these four miRNAs was predicted by miRanda (**Figure 4B**). This result implies that These results indicate that the aforementioned miRNA and lncRNA might have a tight correlation and regulation relationship, which could be our main focus for further study.

The lncRNAs may affect the expression of miRNAs by sponging them.

XR_311113.2 Serves as a Sponge for Chi-miR-424-5p

The miRanda software was used to predict the putative miRNA binding sites of XR_311113.2. **Figure 5A** indicates that four miRNAs (chi-miR-135a, chi-miR-424-5p, chi-miR-544-5p, and chi-miR-3955-5p) bind to XR_311113.2 *via* putative binding sites. To verify that these four miRNAs bind to XR_311113.2, the miRNA binding site regions and the mutated regions of chi-miR-135a (719 bp/741 bp), chi-miR-424-5p (975 bp/995 and 3197 bp/3216 bp), chi-miR-544-5p (680 bp/700 bp), and chi-miR-3955-5p (3569 bp/3591 bp) were cloned into two pmirGLO vectors, named pmirGLO-1 (159 bp/1163 bp, 1005 bp) and pmirGLO-2 (2104 bp/3628 bp, 1525 bp), respectively. Three miRNA mutants (chi-miR-544-5p mut, chi-miR-135a mut, and chi-miR-424-5p mut) of pmirGLO-1 and two miRNA mutants (chi-miR-424-5p mut and chi-miR-3955-5p mut) of pmirGLO-2 were transfected into GK cells. We discovered that chi-miR-424-5p mut ($p < 0.001$) and chi-miR-3955-5p mut ($p < 0.01$) completely abolished the suppression of

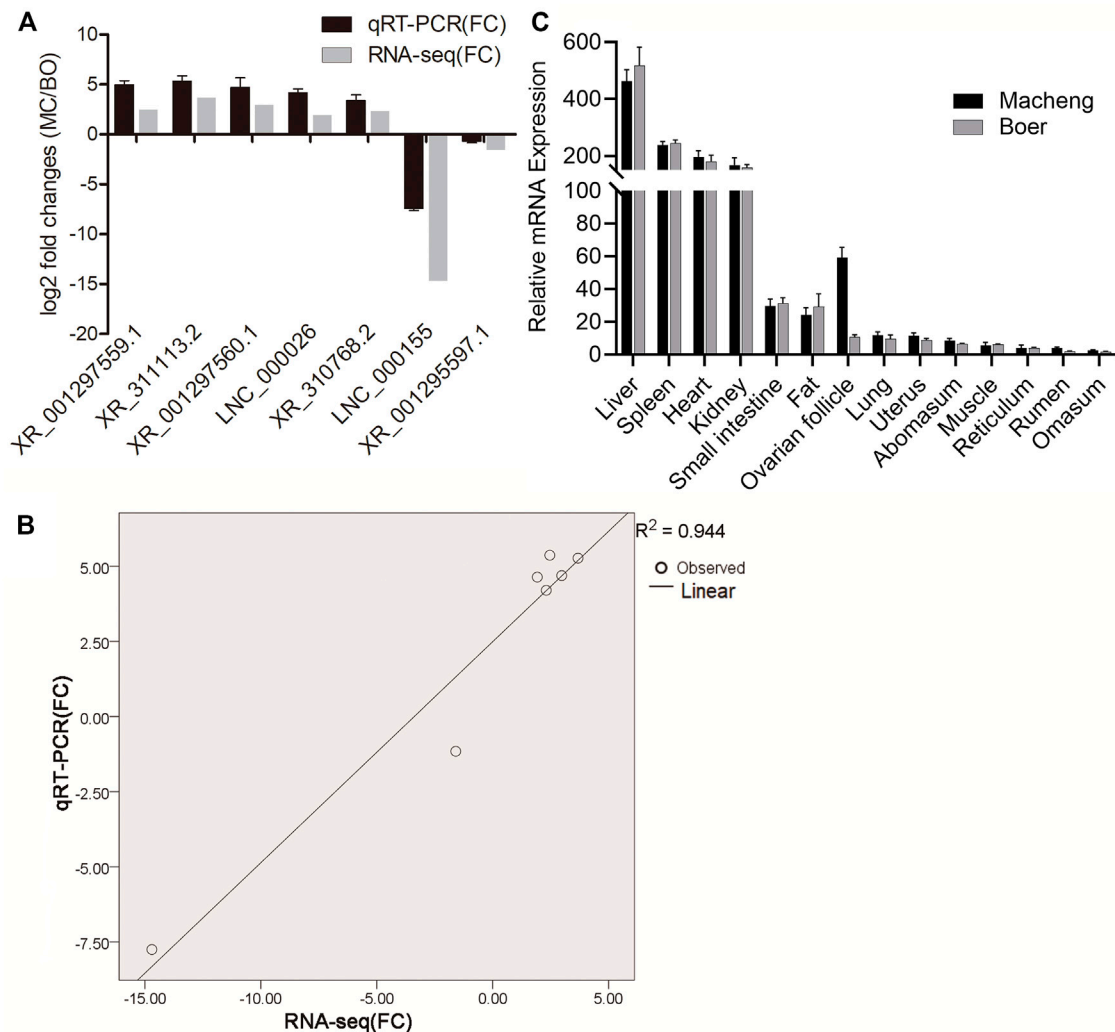


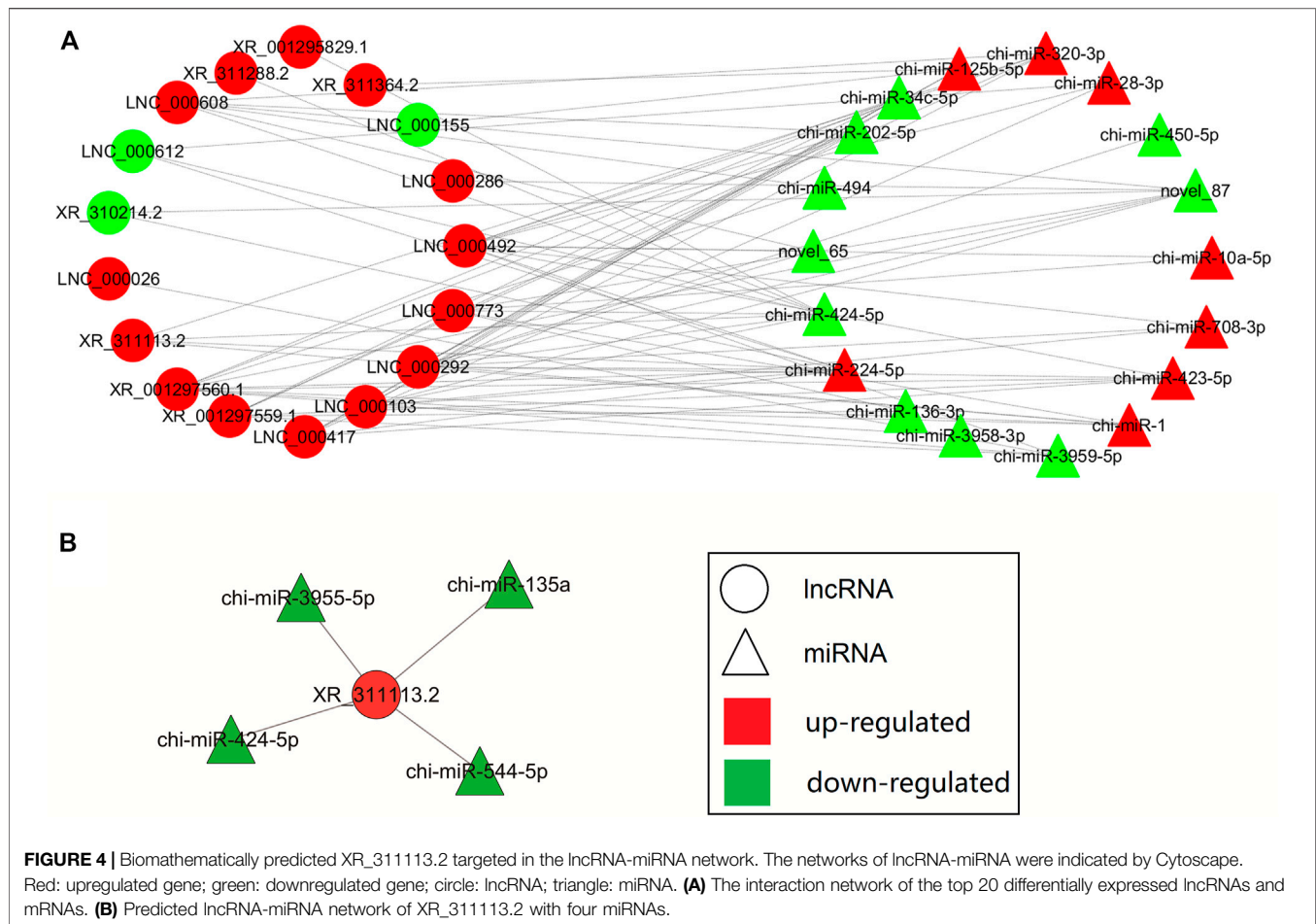
FIGURE 3 | Validation of selected lncRNAs and mRNAs using qRT-PCR. **(A)** Fold changes in the relative expression of lncRNAs, as assessed by qRT-PCR. Results of the comparison of the seven lncRNAs using qRT-PCR and RNA-seq. The vertical axis indicates the mean fold change (log2-fold change) of each lncRNA. **(B)** Correlation analysis of the fold changes between qRT-PCR and RNA-seq. **(C)** Relative expression of lncRNAs. Expression profile of lncRNA XR_311113.2 in 14 tissues of Macheng and Boer goat expressed as the mean \pm SD. Data were normalized to the reference gene (β -actin).

luciferase activity compared with the wild-type vector (**Figure 5B**). These results suggest that chi-miR-424-5p and chi-miR-3955-5p bind to the putative miRNA binding site of XR_311113.2. To clarify further the binding activity of miRNAs, the mimic of chi-miR-424-5p and chi-miR-3955-5p was cotransfected with the pmirGLO-2 luciferase reporter plasmid into GK cells. Compared with the NC group, chi-miR-424-5p significantly reduced luciferase activity ($p < 0.01$) (**Figure 5C**). These results suggest that XR_311113.2 serves as a sponge of chi-miR-424-5p.

DISCUSSION

Goat is a domestic animal with a long history. The Macheng black goat is a unique breed indigenous to mountainous areas of

Central China Region and characterized by systemic black and high prolificacy. Boer goat is one of the most popular breeds of meat goat in the world due to their excellent carcass qualities and high growth efficiency (Naing et al., 2010). The reproductive performance of ewes is an important part of goat productivity. The kidding and fecundity rates of Boer goats are approximately 189% and 210% (Malan, 2000). The Macheng black goat is highly fertile, the kidding rate and fecundity rate are approximately 219% and 346% (Tao et al., 2018). Recent studies indicate that a high ovulation rate can increase the probability of higher litter size (Wang et al., 2020). As a high propagation capability goat breed, how lncRNAs regulate its follicular development is still unclear. Therefore, we investigated the differential expression of lncRNAs in ewe follicular development between Macheng goat and Boer goat using RNA-seq.



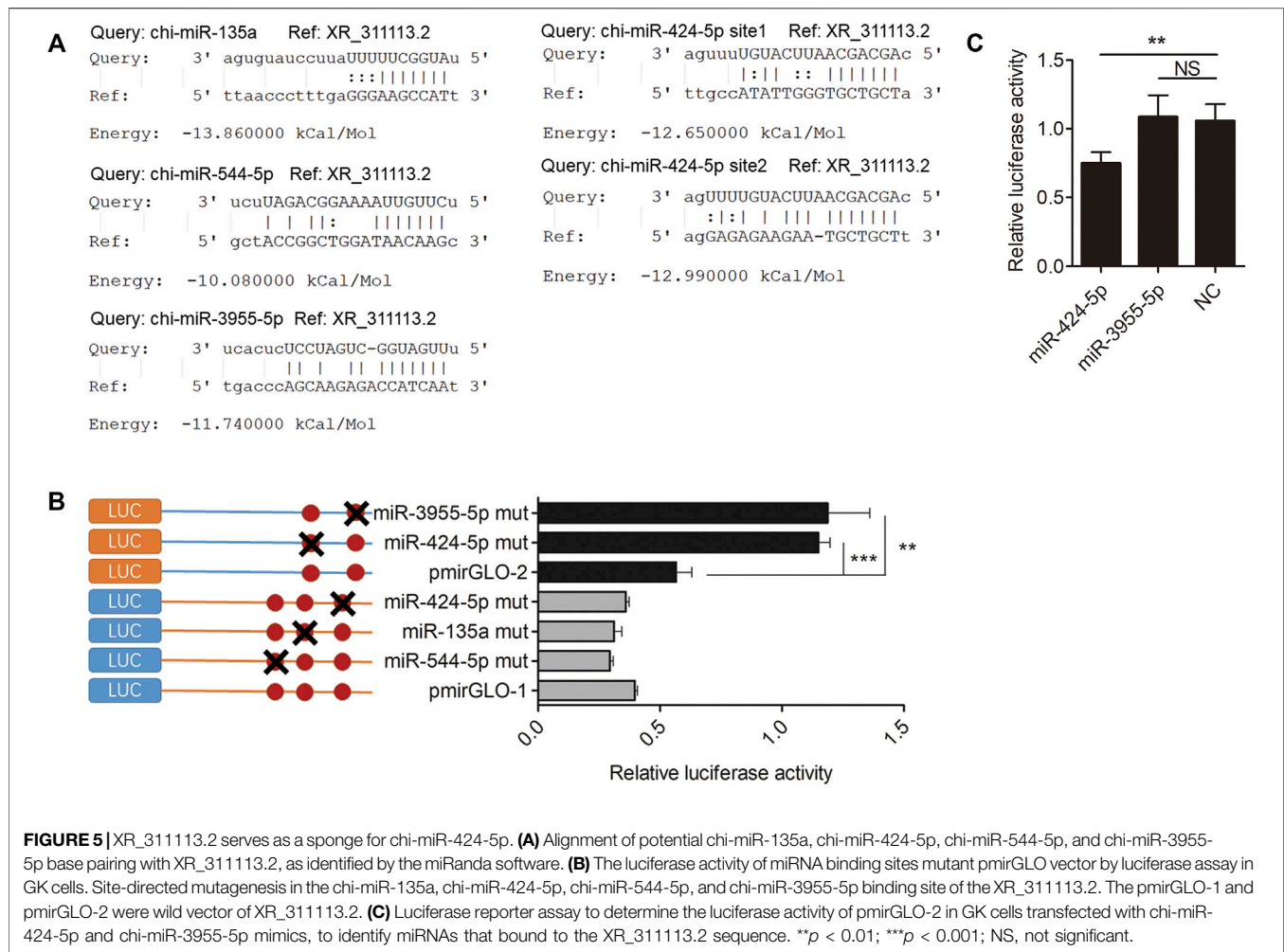
Folliculogenesis is a complex process that includes primordial follicle recruitment, granulosa cell (GC) proliferation, oocyte maturation, steroidogenesis, and ovulation (McLaughlin and McIver, 2009; Field et al., 2014). GCs play a crucial role in each stage of folliculogenesis, provide nutritional support for the development of oocytes through the gap junction, and enhance the maturation of oocytes. The interaction between GCs and theca cells is an important condition for maintaining follicular function and promoting normal development. The growth and differentiation processes of GCs are key criteria for the normal initiation and growth of primordial follicles, and mediate the development and atresia process of growth-period follicles, thereby playing an important regulatory role in the development of follicles (Uyar et al., 2013; Johnson, 2015; Sugiyama et al., 2016; Nguyen et al., 2019; Kim et al., 2020).

lncRNAs are newly discovered ncRNAs that have important regulatory functions in a range of biological events (Peng et al., 2017), especially their role as therapeutic targets and diagnostic biomarkers of various diseases (Ji et al., 2003; Yan et al., 2015; Huang et al., 2017; Huang, 2018; Simion et al., 2019; Yan et al., 2019). A growing body of evidence has shown that lncRNAs are widely distributed in mammals and plants (Necsulea et al., 2014; Wang et al., 2014; Zhao et al., 2018; Sarropoulos et al., 2019), but little has been reported regarding the folliculogenesis process of

goats. In the field of molecular biology, the genetic research of goats is subject to many restrictions compared with the research of other species, such as humans and mice. The database of lncRNAs mainly targets common research subjects (humans, mice, etc). The lack of various databases can hamper the analysis of lncRNA data.

In the current study, we used high-throughput sequencing to analyze the lncRNA profiles in goat ovarian follicle samples. In addition, we investigated the manner in which lncRNA and mRNA networks contribute to follicle development, and confirmed the regulatory function of lncRNA XR_311113.2. We predicted a total of 895 lncRNAs in samples of pre-ovulatory follicles from Macheng black goats and Boer goats, which was less than that detected in the developmental skeletal muscle of fetal goat (3981) (Zhan et al., 2016) and the fetal skin of goats (1336) (Ren et al., 2016), and more than that detected in the anagen-phase skin samples of cashmere goats (437) (Zheng et al., 2019). These results revealed that the expression of lncRNAs was tissue-specific in different goat breeds.

A recent study has revealed that lncRNAs can function as miRNA sponges to further affect the expression of miRNA target genes (Du et al., 2016). For example, lncRNA TRPM2-AS acts as the sponge of miR-612 to promote gastric cancer progression (Xiao et al., 2020), and lncRNA MDNCR binds to miR-133a, thus



promoting cell differentiation in bovine primary myoblasts (Li et al., 2018). In the present study, we found that lncRNA XR_311113.2 was differentially expressed in the pre-ovulatory follicles of the Boer and Macheng goat groups. The bioinformatics analyses showed that four miRNAs potentially interacted with XR_311113.2 (as shown in **Figure 4**). In addition, we validated the interacting relationship between XR_311113.2 and chi-miR-424-5p using luciferase activity assays (as shown in **Figure 5**). Therefore, our results suggest that XR_311113.2 targets chi-miR-424-5p directly by functioning as a sponge in the pre-ovulatory follicles of goats. Moreover, the previous study results demonstrated that miR-424-5p plays an important role in diseases of the reproductive system by directly targeting several key functional genes (Xu et al., 2013; Liu et al., 2018). In addition, miR-424 suppresses the proliferation and promotes the apoptosis of human ovarian granulosa cells (Du et al., 2020). These observations suggest that lncRNAs regulate goat ovarian follicular development by binding to miRNAs that regulate target gene expression.

One result we find out is really interesting. In **Figure 5B**, when we mutated the binding site of miR-3955-5p in XR_311113.2 sequence, the luciferase activity of the mutant pmirGLO vector significantly increased compared with wild type pmirGLO vector.

The results showed that the mutant site might prevent miR-3955-5p bind to pmirGLO, so that the luciferase activity of pmirGLO will not be inhibited by miR-3955-5p. To clarify further the binding activity of miR-3955-5p, the chi-miR-3955-5p mimic was synthesized and co-transfected with the wild pmirGLO plasmid. Compared with the NC group, chi-miR-3955-5p did not significantly reduce luciferase activity of wild pmirGLO. Why? The reason may be that the mutated binding site is not necessarily the factual site of miR-3955-5p, but may be the binding site of other unknown miRNAs. When this fake site is mutated, the unknown miRNA cannot inhibit the luciferase activity. This is why we continue to use the results of **Figure 5C** to further eliminate the fake results of **Figure 5B** and ensure the reliability of the results.

MiRNAs have emerged as key post-transcriptional regulators of target expression through complementary base pairing with the target (Filipowicz et al., 2008). A large number of computational prediction tools for the prediction of putative miRNA targets have been developed, and commonly used tools in mammals include miRWalk, PicTar, miRanda, Targetscan, RNAhybrid, PITA, miRmap, DIANA-microT and RNA22V2 (Riffo-Campos et al., 2016). However, the described

tools are mainly developed for a few species such as humans, mice, pigs etc. Some of these tools can only set a few parameters and cannot enter sequences. For obscure livestock such as goats, it is impossible to accurately predict the targeting relationship between miRNA and target. The lack of various databases and tools can make it difficult to carry out the study of goats.

CONCLUSION

In conclusion, we analyzed the lncRNA and mRNA expression profiles of goat ewe pre-ovulatory follicles to predict the interactions between lncRNAs and miRNAs. Overall, 895 lncRNAs were identified, 88 of which were preliminarily determined to show a marked differential expression, indicating potentially substantial effects on goat ovarian follicles. The tissue expression profile of lncRNA showed tissue specificity. lncRNA XR_311113.2 may function as a sponge of chi-miR-424-5p. Our RNA-seq data contributed to the known types of lncRNA species. lncRNAs may represent a promising new field of research in the area of ovarian follicular development.

DATA AVAILABILITY STATEMENT

Sequencing data were deposited in the NCBI Sequence Read Archive (SRA) under accession number PRJNA648013.

ETHICS STATEMENT

The animal study was reviewed and approved by The ethics committee of Hubei Academy of Agricultural Sciences. Written

informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: HT and MC. Wrote the manuscript: HT. Performed the experiments: HT and JY. Collection of the sample: NZ and PZ. RNA extraction: XS. Critically revised the manuscript: XL and YL. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (NSFC, 31972995 and 31501932), Hubei Province Key R&D Program (2021BBA086) and Hubei Announcement System Technology Project (2020BED028).

ACKNOWLEDGMENTS

We thank the reviewers for their critical reading and discussion of the manuscript. The authors were grateful to Online English (<https://www.oleng.com.au/>) for manuscript editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.760416/full#supplementary-material>

REFERENCES

- Choi, Y., and Rajkovic, A. (2006). Genetics of Early Mammalian Folliculogenesis. *Cell. Mol. Life Sci.* 63 (5), 579–590. doi:10.1007/s00018-005-5394-7
- Cui, H. X., Zhao, S. M., Cheng, M. L., Guo, L., Ye, R. Q., Liu, W. Q., et al. (2009). Cloning and Expression Levels of Genes Relating to the Ovulation Rate of the Yunling Black Goat. *Biol. Reprod.* 80 (2), 219–226. doi:10.1095/biolreprod.108.069021
- Du, J., Lin, X., Wu, R., Gao, Z., Du, Y., Liao, Y., et al. (2020). miR-424 Suppresses Proliferation and Promotes Apoptosis of Human Ovarian Granulosa Cells by Targeting Apelin and APJ Expression. *Am. J. Transl. Res.* 12 (7), 3660–3673.
- Du, Z., Sun, T., Hacisuleyman, E., Fei, T., Wang, X., Brown, M., et al. (2016). Integrative Analyses Reveal a Long Noncoding RNA-Mediated Sponge Regulatory Network in Prostate Cancer. *Nat. Commun.* 7, 10982. doi:10.1038/ncomms10982
- Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA Targets in *Drosophila*. *Genome Biol.* 5 (1), R1. doi:10.1186/gb-2003-5-1-r1
- Field, S. L., Dasgupta, T., Cummings, M., and Orsi, N. M. (2014). Cytokines in Ovarian Folliculogenesis, Oocyte Maturation and Luteinisation. *Mol. Reprod. Dev.* 81 (4), 284–314. doi:10.1002/mrd.22285
- Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional Regulation by microRNAs: Are the Answers in Sight? *Nat. Rev. Genet.* 9 (2), 102–114. doi:10.1038/nrg2290
- Han, P., and Chang, C.-P. (2015). Long Non-coding RNA and Chromatin Remodeling. *RNA Biol.* 12 (10), 1094–1098. doi:10.1080/15476286.2015.1063770
- Huang, Y. (2018). The Novel Regulatory Role of lncRNA-miRNA-mRNA axis in Cardiovascular Diseases. *J. Cel. Mol. Med.* 22 (12), 5768–5775. doi:10.1111/jcmm.13866
- Huang, Y., Zhang, J., Hou, L., Wang, G., Liu, H., Zhang, R., et al. (2017). lncRNA AK023391 Promotes Tumorigenesis and Invasion of Gastric Cancer through Activation of the PI3K/Akt Signaling Pathway. *J. Exp. Clin. Cancer Res.* 36 (1), 194. doi:10.1186/s13046-017-0666-2
- Hung, T., and Chang, H. Y. (2010). Long Noncoding RNA in Genome Regulation. *RNA Biol.* 7 (5), 582–585. doi:10.4161/rna.7.5.13216
- Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., et al. (2003). MALAT-1, a Novel Noncoding RNA, and Thymosin β 4 Predict Metastasis and Survival in Early-Stage Non-small Cell Lung Cancer. *Oncogene* 22 (39), 8031–8041. doi:10.1038/sj.onc.1206928
- Jin, J. J., Lv, W., Xia, P., Xu, Z. Y., Zheng, A. D., Wang, X. J., et al. (2018). Long Noncoding RNAs YSL1 Regulates Myogenesis by Interacting with Polycomb Repressive Complex 2. *Proc. Natl. Acad. Sci. USA* 115 (42), E9802–E9811. doi:10.1073/pnas.1801471115
- Johnson, A. L. (2015). Ovarian Follicle Selection and Granulosa Cell Differentiation. *Poult. Sci.* 94 (4), 781–785. doi:10.3382/ps/peu008
- Kim, B. H., Ju, W. S., Kim, J.-S., Kim, S.-U., Park, S. J., Ward, S. M., et al. (2020). Effects of Gangliosides on Spermatozoa, Oocytes, and Preimplantation Embryos. *Int. J. Mol. Sci.* 21 (1), 106. doi:10.3390/ijms21010106

- Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., et al. (2007). CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine. *Nucleic Acids Res.* 35, W345–W349. doi:10.1093/nar/gkm391
- Li, H., Yang, J., Jiang, R., Wei, X., Song, C., Huang, Y., et al. (2018). Long Non-coding RNA Profiling Reveals an Abundant MDNCR that Promotes Differentiation of Myoblasts by Sponging miR-133a. *Mol. Ther. - Nucleic Acids* 12, 610–625. doi:10.1016/j.omtn.2018.07.003
- Li, R., Harvey, A. R., Hodgetts, S. I., and Fox, A. H. (2017). Functional Dissection of NEAT1 Using Genome Editing Reveals Substantial Localization of the NEAT1_1 Isoform outside Paraspeckles. *RNA* 23 (6), 872–881. doi:10.1261/rna.059477.116
- Li, Y., Liu, Y.-d., Chen, S.-l., Chen, X., Ye, D.-s., Zhou, X.-y., et al. (2018a). Down-regulation of Long Non-coding RNAMALAT1 inhibits Granulosa Cell Proliferation in Endometriosis by Up-Regulating P21 via Activation of the ERK/MAPK Pathway. *Mol. Hum. Reprod.* 25 (1), 17–29. doi:10.1093/molehr/gay045
- Li, Y., Wang, H., Zhou, D., Shuang, T., Zhao, H., and Chen, B. (2018b). Up-regulation of Long Noncoding RNA SRA Promotes Cell Growth, Inhibits Cell Apoptosis, and Induces Secretion of Estradiol and Progesterone in Ovarian Granular Cells of Mice. *Med. Sci. Monit.* 24, 2384–2390. doi:10.12659/MSM.907138
- Liu, J., Gu, Z., Tang, Y., Hao, J., Zhang, C., and Yang, X. (2018). Tumour-suppressive microRNA-424-5p Directly Targets CCNE1 as Potential Prognostic Markers in Epithelial Ovarian Cancer. *Cell Cycle* 17 (3), 309–318. doi:10.1080/15384101.2017.1407894
- Liu, Y.-d., Li, Y., Feng, S.-x., Ye, D.-s., Chen, X., Zhou, X.-y., et al. (2017). Long Noncoding RNAs: Potential Regulators Involved in the Pathogenesis of Polycystic Ovary Syndrome. *Endocrinology* 158 (11), 3890–3899. doi:10.1210/en.2017-00605
- Lu, W., Huang, S. Y., Su, L., Zhao, B. X., and Miao, J. Y. (2016). Long Noncoding RNA LOC100129973 Suppresses Apoptosis by Targeting MIR-4707-5p and MIR-4767 in Vascular Endothelial Cells. *Sci. Rep.* 6 (1), 21620. doi:10.1038/srep21620
- Malan, S. W. (2000). The Improved Boer Goat. *Small Ruminant Res.* 36 (2), 165–170. doi:10.1016/S0921-4488(99)00160-1
- McHugh, C. A., Chen, C.-K., Chow, A., Surka, C. F., Tran, C., McDonel, P., et al. (2015). The Xist lncRNA Interacts Directly with SHARP to Silence Transcription through HDAC3. *Nature* 521 (7551), 232–236. doi:10.1038/nature14443
- McLaughlin, E. A., and McIver, S. C. (2009). Awakening the Oocyte: Controlling Primordial Follicle Development. *Reproduction* 137 (1), 1–11. doi:10.1530/REP-08-0118
- Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521
- Naing, S. W., Wahid, H., Mohd Azam, K., Rosnina, Y., Zuki, A. B., Kazhal, S., et al. (2010). Effect of Sugars on Characteristics of Boer Goat Semen after Cryopreservation. *Anim. Reprod. Sci.* 122 (1–2), 23–28. doi:10.1016/j.anireprosci.2010.06.006
- Nakagawa, S., Shimada, M., Yanaka, K., Mito, M., Arai, T., Takahashi, E., et al. (2014). The lncRNA Neat1 Is Required for Corpus Luteum Formation and the Establishment of Pregnancy in a Subpopulation of Mice. *Dev* 141 (23), 4618–4627. doi:10.1242/dev.110544
- Necsulea, A., Soumilion, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., et al. (2014). The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods. *Nature* 505 (7485), 635–640. doi:10.1038/nature12943
- Nguyen, X. P., Nakamura, T., Osuka, S., Bayasula, B., Nakanishi, N., Kasahara, Y., et al. (2019). Effect of the Neuropeptide Phoenixin and its Receptor GPR173 during Folliculogenesis. *Reproduction* 158 (1), 25–34. doi:10.1530/REP-19-0025
- Pan, Z., Zhang, J., Li, Q., Li, Y., Shi, F., Xie, Z., et al. (2012). Current Advances in Epigenetic Modification and Alteration during Mammalian Ovarian Folliculogenesis. *J. Genet. Genomics* 39 (3), 111–123. doi:10.1016/j.jgg.2012.02.004
- Paraskevopoulou, M. D., and Hatzigeorgiou, A. G. (2016). Analyzing MiRNA-lncRNA Interactions. *Methods Mol. Biol.* 1402 (1), 271–286. doi:10.1007/978-1-4939-3378-5_21
- Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., et al. (2016). Genome-wide Changes in lncRNA, Splicing, and Regional Gene Expression Patterns in Autism. *Nature* 540 (7633), 423–427. doi:10.1038/nature20612
- Peng, W.-X., Koirala, P., and Mo, Y.-Y. (2017). lncRNA-mediated Regulation of Cell Signaling in Cancer. *Oncogene* 36 (41), 5661–5667. doi:10.1038/onc.2017.184
- Punta, M., Coghill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam Protein Families Database. *Nucleic Acids Res.* 40 (1), D290–D301. doi:10.1093/nar/gkr1065
- Ren, H., Wang, G., Chen, L., Jiang, J., Liu, L., Li, N., et al. (2016). Genome-wide Analysis of Long Non-coding RNAs at Early Stage of Skin Pigmentation in Goats (*Capra hircus*). *BMC Genomics* 17, 67. doi:10.1186/s12864-016-2365-3
- Riffo-Campos, Á., Riquelme, I., and Brebi-Mieville, P. (2016). Tools for Sequence-Based miRNA Target Prediction: What to Choose? *Int. J. Mol. Sci.* 17 (12), 1987. doi:10.3390/ijms17121987
- Saharrea, A., Valencia, J., Balcázar, A., Mejía, O., Cerbón, J. L., Caballero, V., et al. (1998). Premature Luteal Regression in Goats Superovulated with PMSG: Effect of hCG or GnRH Administration during the Early Luteal Phase. *Theriogenology* 50 (7), 1039–1052. doi:10.1016/S0093-691X(98)00206-4
- Sarropoulos, I., Marin, R., Cardoso-Moreira, M., and Kaessmann, H. (2019). Developmental Dynamics of lncRNAs across Mammalian Organs and Species. *Nature* 571 (7766), 510–514. doi:10.1038/s41586-019-1341-x
- Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation Coefficients. *Anesth. Analgesia* 126 (5), 1763–1768. doi:10.1213/ANE.0000000000002864
- Simion, V., Haemmig, S., and Feinberg, M. W. (2019). lncRNAs in Vascular Biology and Disease. *Vasc. Pharmacol.* 114, 145–156. doi:10.1016/j.vph.2018.01.003
- Sugiyama, M., Sumiya, M., Shirasuna, K., Kuwayama, T., and Iwata, H. (2016). Addition of Granulosa Cell Mass to the Culture Medium of Oocytes Derived from Early Antral Follicles Increases Oocyte Growth, ATP Content, and Acetylation of H4K12. *Zygote* 24 (6), 848–856. doi:10.1017/S0967199416000198
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing Sequence Intrinsic Composition to Classify Protein-Coding and Long Non-coding Transcripts. *Nucleic Acids Res.* 41 (17), e166. doi:10.1093/nar/gkt646
- Tao, H., Xiong, Q., Zhang, F., Zhang, N., Liu, Y., Suo, X., et al. (2018). Circular RNA Profiling Reveals Chi_circ_0008219 Function as microRNA Sponges in Pre-ovulatory Ovarian Follicles of Goats (*Capra hircus*). *Genomics* 110 (4), 257–266. doi:10.1016/j.ygeno.2017.10.005
- Torner, H., Brüssow, K.-P., Alm, H., Rätty, J., and Kanitz, W. (1998). Morphology of Porcine Cumulus-Oocyte-Complexes Depends on the Stage of Preovulatory Maturation. *Theriogenology* 50 (1), 39–48. doi:10.1016/S0093-691X(98)00111-3
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks. *Nat. Protoc.* 7 (3), 562–578. doi:10.1038/nprot.2012.016
- Uyar, A., Torrealday, S., and Seli, E. (2013). Cumulus and Granulosa Cell Markers of Oocyte and Embryo Quality. *Fertil. Sterility* 99 (4), 979–997. doi:10.1016/j.fertnstert.2013.01.129
- Wang, J.-J., Zhang, T., Chen, Q.-M., Zhang, R.-Q., Li, L., Cheng, S.-F., et al. (2020). Genomic Signatures of Selection Associated with Litter Size Trait in Jining Gray Goat. *Front. Genet.* 11, 286. doi:10.3389/fgene.2020.00286
- Wang, J., Gong, X., Tian, G. G., Hou, C., Zhu, X., Pei, X., et al. (2018). Long Noncoding RNA Growth Arrest-specific 5 Promotes Proliferation and Survival of Female Germline Stem Cells *In Vitro*. *Gene* 653 (4), 14–21. doi:10.1016/j.gene.2018.02.021
- Wang, K., Jin, W., Song, Y., and Fei, X. (2017). lncRNA RP11-436H11.5, Functioning as a Competitive Endogenous RNA, Upregulates BCL-W Expression by Sponging miR-335-5p and Promotes Proliferation and Invasion in Renal Cell Carcinoma. *Mol. Cancer* 16 (1), 166. doi:10.1186/s12943-017-0735-3
- Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., et al. (2014). Arabidopsis Noncoding RNA Mediates Control of Photomorphogenesis by Red Light. *Proc. Natl. Acad. Sci.* 111 (28), 10359–10364. doi:10.1073/pnas.1409457111
- Xiao, J., Lin, L., Luo, D., Shi, L., Chen, W., Fan, H., et al. (2020). Long Noncoding RNA TRPM2-AS Acts as a microRNA Sponge of miR-612 to Promote Gastric Cancer Progression and Radioresistance. *Oncogenesis* 9 (3), 29. doi:10.1038/s41389-020-0215-2
- Xu, J., Li, Y., Wang, F., Wang, X., Cheng, B., Ye, F., et al. (2013). Suppressed miR-424 Expression via Upregulation of Target Gene Chk1 Contributes to the Progression of Cervical Cancer. *Oncogene* 32 (8), 976–987. doi:10.1038/onc.2012.121

- Yan, L., Zhou, J., Gao, Y., Ghazal, S., Lu, L., Bellone, S., et al. (2015). Regulation of Tumor Cell Migration and Invasion by the H19/let-7 axis Is Antagonized by Metformin-Induced DNA Methylation. *Oncogene* 34 (23), 3076–3084. doi:10.1038/onc.2014.236
- Yan, S., Wang, P., Wang, J., Yang, J., Lu, H., Jin, C., et al. (2019). Long Non-coding RNA HIX003209 Promotes Inflammation by Sponging miR-6089 via TLR4/NF-Kb Signaling Pathway in Rheumatoid Arthritis. *Front. Immunol.* 10, 2218. doi:10.3389/fimmu.2019.02218
- Zhan, S., Dong, Y., Zhao, W., Guo, J., Zhong, T., Wang, L., et al. (2016). Genome-wide Identification and Characterization of Long Non-coding RNAs in Developmental Skeletal Muscle of Fetal Goat. *BMC Genomics* 17 (1), 666. doi:10.1186/s12864-016-3009-3
- Zhao, X., Li, J., Lian, B., Gu, H., Li, Y., and Qi, Y. (2018). Global Identification of Arabidopsis lncRNAs Reveals the Regulation of MAF4 by a Natural Antisense RNA. *Nat. Commun.* 9 (1), 5056. doi:10.1038/s41467-018-07500-7
- Zheng, Y. Y., Sheng, S. D., Hui, T. Y., Yue, C., Sun, J. M., Guo, D., et al. (2019). An Integrated Analysis of cashmere Fineness lncRNAs in cashmere Goats. *Genes* 10 (4), 266. doi:10.3390/genes10040266

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Tao, Yang, Zhang, Zhang, Suo, Li, Liu and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Function Identification of Bovine *ACSF3* Gene and Its Association With Lipid Metabolism Traits in Beef Cattle

Wei He^{1†}, Xibi Fang^{1†}, Xin Lu¹, Yue Liu¹, Guanghui Li¹, Zhihui Zhao², Junya Li³ and Runjun Yang^{1*}

¹ College of Animal Science, Jilin University, Changchun, China, ² College of Coastal Agricultural Sciences, Guangdong Ocean University, Zhanjiang, China, ³ Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and
Technology, India

Reviewed by:

Jiangjiang Zhu,
Southwest Minzu University, China
Nedenia Stafuzza,
Animal Science Institute, São Paulo,
Brazil

*Correspondence:

Runjun Yang
yrj@jlu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Veterinary Science

Received: 30 August 2021

Accepted: 13 December 2021

Published: 06 January 2022

Citation:

He W, Fang X, Lu X, Liu Y, Li G,
Zhao Z, Li J and Yang R (2022)
Function Identification of Bovine
ACSF3 Gene and Its Association With
Lipid Metabolism Traits in Beef Cattle.
Front. Vet. Sci. 8:766765.
doi: 10.3389/fvets.2021.766765

Acyl-CoA synthetase family member 3 (*ACSF3*) carries out the first step of mitochondrial fatty acid synthesis II, which is the linkage of malonate and, to a lesser extent, methylmalonate onto CoA. Malonyl-coenzyme A (malonyl-CoA) is a central metabolite in mammalian fatty acid biochemistry that is generated and utilized in the cytoplasm. In this research, we verified the relationship between expression of the *ACSF3* and the production of triglycerides (TGs) at the cellular level by silencing and over-expressing *ACSF3*. Subsequently, through Sanger sequencing, five polymorphisms were found in the functional domain of the bovine *ACSF3*, and the relationship between *ACSF3* polymorphism and the economic traits and fatty acid composition of Chinese Simmental cattle was analyzed by a means of variance analysis and multiple comparison. The results illustrated that the expression of *ACSF3* promoted triglyceride synthesis in bovine mammary epithelial cells and bovine fetal fibroblast cells. Further association analysis also indicated that individuals with the AG genotype (g.14211090 G > A) of *ACSF3* were significantly associated with the fatty acid composition of intramuscular fat (higher content of linoleic acid, α -linolenic acid, and arachidonic acid), and that CTCAG haplotype individuals were significantly related to the fatty acid composition of intramuscular fat (higher linoleic acid content). Individuals with the AA genotypes of g.14211055 A > G and g.14211090 G > A were substantially associated with a larger eye muscle area in the Chinese Simmental cattle population. *ACSF3* played a pivotal role in the regulation of cellular triacylglycerol and long-chain polyunsaturated fatty acid levels, and polymorphism could serve as a useful molecular marker for future marker-assisted selection in the breeding of intramuscular fat deposition traits in beef cattle.

Keywords: Acyl-CoA synthetase family member 3, Chinese Simmental cattle, fat deposition traits, single nucleotide polymorphism, triglyceride (TG)

INTRODUCTION

As there are large numbers of Chinese Simmental cattle breeding groups, verifying the gene function of this breed's high-quality traits is an area in need of urgent research. The Chinese Simmental has become a large-scale dairy and meat breed in China since it was successfully selected and bred in 2001. In recent years, with the maturity of sequencing technology, new association studies have been conducted to assess whether genetic polymorphisms affect the economic traits and fatty acid composition in cattle (1, 2). Molecular marker information can be used as a useful tool for evaluating breeding value and selecting animal carcass and meat traits.

Fatty acid synthesis occurs through two pathways, one of which takes place in cellular structures called mitochondria. Mitochondria convert the energy from food into a form that cells can use, and fatty acid synthesis in these structures is essential for their proper functioning. The ACSF3 enzyme is found only in the mitochondria and is involved in mitochondrial fatty acid synthesis. The ACSF3 provides instructions for making an enzyme involved in the formation (synthesis) of fatty acids, which are the building blocks used to make fats (lipids) (3). The ACSF3 enzyme performs a chemical reaction that converts malonic acid to malonyl-CoA, which is the first step of fatty acid synthesis (4). Based on this activity, the enzyme is classified as a malonyl-CoA synthetase (5). The ACSF3 enzyme also converts methylmalonic acid to methylmalonyl-CoA, making it a methylmalonyl-CoA synthetase as well (6). Studies have shown that the single nucleotide polymorphism of ACSF3 may be related to meat quality traits, such as backfat thickness (7, 8). The bovine ACSF3 gene is located on bovine chromosome 18:14,209,800–14,248,433 and has 10 exons and 9 introns. The ACSF3 mRNA sequence contains a 1,761 bp coding region (CDS) and a 276 bp untranslated region. The open reading frame (ORF) encodes 586 amino acids and has a 94 and 84% homology with the coding regions of sheep and swine, respectively. The transcriptome analysis results of our previous study showed that ACSF3 was a candidate gene related to bovine fat deposits. However, the ACSF3 gene has been rarely studied in bovine lipid metabolism.

This study aims to reveal the regulatory effect of ACSF3 on adipogenesis by intracellular gene RNA interference (RNAi) and gene over-expression. Moreover, we investigate the polymorphisms of the bovine ACSF3 gene functional domain to explore the relationship between genotypes and fat deposition, and the fatty acid composition traits of Chinese Simmental cattle.

MATERIALS AND METHODS

Ethics Statements

All the animal experiments in the present study strictly complied with the relevant regulations regarding the care and use of experimental animals issued by the Animal Protection and Use Committee of Jilin University [permit no. SYXK(Ji) pzpx20181227083].

Experimental Materials

The BMECs (bovine mammary epithelial cells) in this study were purified and cultured according to the previous work carried out by the Laboratory of Bovine Genetic Resources and Functional Genomics (9). Briefly, bovine mammary tissues were cut into 1 mm³ nubbles and washed again with PBS solution until the tissue was clean. The smaller pieces of tissue were transferred onto cell culture dishes. A basal media was prepared in advance: DMEM/F12 with 10% fetal bovine serum, 1% penicillin and streptomycin (Hyclone, Logan, UT, USA), and 1% epithelial growth factor. Next, the cell culture dishes with the tissue were incubated at 37°C in a 5% CO₂ incubator. After 6 h, 5 mL of basal media were added to each dish, ensuring that the tissue would not float and separate from the bottom of the culture dish. The basal media was replaced with fresh media every 48 h until the

culture dish was full of cells. The cells were detached with 0.25% trypsin-0.02%EDTA (Hyclone, Logan, UT, USA) and transferred to new culture dishes that were used to remove the fibroblasts. Subsequently, pure mammary epithelial cells were isolated after 3–5 passages.

The BFFs (bovine fetal fibroblast cells) were isolated and preserved by the Genetic Breeding and Reproduction Laboratory of Jilin University. Briefly, the BFFs were obtained by isolating fetal bovine ear tip tissue using a tissue block apposition method similar to that described above.

This study involved 135 Chinese Simmental cattle (28-month-old bulls) from a Baolongshan cattle farm in Inner Mongolia. These cattle were randomly selected from the offspring of a Simmental cattle population with ~2,000 cows and 25 bulls. Blood samples (10 mL each) were collected from jugular vein with anticoagulant (Acid citrate dextrose, ACD) and stored in –70°C. DNA was extracted from 1 mL whole blood with the DNA extraction kit (Tiangen, Beijing, China) according to the manufacturer's protocol.

Trait Analysis

In this study, 36 traits were measured and 14 fatty acid compositions of the *Longissimus dorsi* muscle were analyzed (10, 11).

Before the measurements of carcass and fat deposition traits, all the carcasses were stored in refrigerating chambers at the temperature between 0 and 4 °C for 24 h. All the measurements complied with the criterion GB/T17238-1998 cutting standard of fresh and chilled beef of China (China Standard Publish). Before slaughter, ultima live backfat thickness, body weight and *longissimus* muscle area (by ultrasound) were recorded. The weights of carcass, omental fat, mesenteric fat, kidney fat (stripping kidney fat and weighting) was recorded at the slaughter plant. And we also record fat coverage rate, marbling score, fat color score, muscle color score, rib eye area, and backfat thickness after slaughter. Weight of kidney fat was expressed as an absolute value and also as a percentage of hot carcass weight. Carcass yield was calculated as (hot carcass weight × 100)/final weight. Marbling was measured by video image analysis and expressed as the percentage of visible fat area over the total area of a steak (*Longissimus thoracis* muscle taken between the 9 th and 10 th vertebrae from the right side of the carcass). Moreover, the carcass length (measuring the shortest length between the bun midpoint to the sciatic trailing edge), carcass chest depth (measuring the shortest distance of the test cattle between bun posterior border to pectoral by using a ruler), hind leg circumference (measuring the surrounded degree in the junction of femur tibia and fibula), hind leg width (measuring the horizontal width since the end of the medial sag to thigh front), thigh meat thickness (the vertical distance from the surface to the midpoint of the autologous femoral body) and other carcass traits were recorded at the slaughter plant.

Briefly, intramuscular fat was obtained from the *Longissimus dorsi* muscle. The analysis was performed in accordance with the ISO 5,509 (2,000) norms, and AOAC procedures (Association of Official Analytical Chemists), and expressed as g/100 g of fresh tissue. The intramuscular fat was collected in a glass

methylation tube, mixed with 4.0 mL of MeOH, 2.0 mL of chloroform (capillary GC; Sigma-Aldrich, MO, USA), and 1.5 mL of ddH₂O, and placed in a rotating platform shaker for 10 s, then left at room temperature for 15 min. Chloroform and ddH₂O were added to the wells and mixed again, and left at room temperature for 5 min. The mixture was centrifuged at 1,500 g/min for 30 min, then cooled to 16°C, the upper layer removed, and dried under nitrogen. Next, 1.5 mL of 14% BF₃/MeOH reagent were added and mixed vigorously. The mixture was heated at 90°C for 30 min, cooled to room temperature, and 4.0 mL of hexane was added, then methyl esters were extracted in the hexane phase following the addition of 1.5 mL of ddH₂O. The samples were centrifuged for 1 min, and then the upper hexane layer was removed and concentrated under nitrogen. Fatty acid methyl esters were analyzed by gas chromatography using a fully automated HP5890 system (Agilent, Santa Clara, CA, USA) equipped with a flame-ionization detector. An SP-2,560 column (Supelco, PA, USA) 100 m × 0.25 mm × 0.20 μm was used for gas chromatographic detection. The detection procedure was in an oven at 140°C for 5 min, 10°C/min to 220°C, and held for 50 min. An injector/detector was used at 260°C with helium as a carrier gas at about 5 psi. The identification of components was carried out by a comparison of the retention times with those of authentic standards (Sigma-Aldrich, MO, USA).

Primer Design

Primer Premier 6 software (PREMIER Biosoft, San Francisco, CA, USA) was utilized to design the ACSF3 gene coding sequence primers and the SNP primers, based on the bovine ACSF3 gene's existing published sequences (ENSBTAG00000015968).

Construction of pGPU6/GFP/NEO-shACSF3 Vector and pBI-CMV3-ACSF3 Vector

The constructed RNAi-vector pGPU6/GFP/NEO-shACSF3 with a size of 5276 bp and containing the shRNA sequence of bovine ACSF3 gene. The enhanced green fluorescent protein (EGFP) gene was used for transient expression in cells, and kanamycin was used as a selection marker for prokaryotic cell (DH5α) amplification (Figures 2A-II). The shRNA target sequences of bovine ACSF3 mRNA were screened and designed by BLOCK-iTTM RNAi Designer (<https://rnaidesigner.thermofisher.com/rnaexpress/>, Table 1). The shRNA sequence was cloned into a pGPU6/GFP/Neo vector (GenePharma Corporation, Shanghai, China) using the *Bbs*I and *Bam*HI (New England Biolabs, MA, USA) restriction enzyme digestion method. The eukaryotic expression vector pGPU6/GFP/Neo was double digested with *Bbs*I and *Bam*HI in a 10 μL reaction system: *Bbs*I 0.75 μL, *Bam*HI 0.75 μL, NEBuffer 2.1TM 1 μL, pGPU6/GFP/Neo plasmid 2 μL, Nuclease-free water 4.5 μL, and incubated at 37°C 5 h. Recovered and ligated by T4 DNA Ligase (Takara, Dalian, China): T4 DNA Ligase Buffer 2 μL, T4 DNA Ligase 1 μL, pGPU6/GFP/Neo vector 1 μL, shRNA fragment 1 μL, Nuclease-free water 15 μL and incubated at 4°C overnight. The products were transformed into competent *E. coli* (DH5α) cells (Tiangen, Beijing, China), and then prepared the pGPU6/GFP/NEO-shACSF3 plasmid.

The kanamycin-positive clones were picked out and transferred to 200 mL of liquid Luria-Bertani (LB) medium and shaken overnight at 37°C. The pGPU6/GFP/NEO-shACSF3 plasmids were extracted using the EndoFree Maxi Plasmid Kit (Tiangen, Beijing, China) according to the manufacturer's protocol. Briefly, the bacterial broth was centrifuged to obtain the precipitate, fully lysed and removed the endotoxin, and rinsed with enzyme-free water to obtain the plasmid. The concentration and purity of the plasmids were assessed using the agarose gel and spectrophotometer (Nanodrop 2,000, Thermo Scientific, Waltham, MA, USA).

The constructed overexpression vector pBI-CMV3-ACSF3, which is 5,539 bp in size and contains the sequence of the CDS region of bovine ACSF3 gene. The reef coral *Zoanthus sp.* green fluorescent protein (ZsGreen) gene was used for transient expression in cells and ampicillin was used as a selection marker (Figures 2A-IV). The CDS region of ACSF3 synthesized by Sangon Biotech (Shanghai, China) was ligated into a pBI-CMV3 vector (Clontech Laboratories, Mountain View, CA, USA) using the *Mlu*I and *Hind*III (New England Biolabs, MA, USA) enzyme digestion method. Briefly, using the method similar to that described above, the resistance was replaced with ampicillin. The ampicillin-positive clones were picked out and transferred to 200 mL of liquid Luria-Bertani medium and shaken overnight at 37°C. The pBI-CMV3-ACSF3 plasmid was extracted using the EndoFree Maxi Plasmid Kit (Tiangen, Beijing, China) according to the manufacturer's protocol.

The Culture and Transfection of BMECs and BFFs

The BMECs and BFFs were proliferated for 24 h in six-well plates (Jet Bio-Filtration Co., Ltd, Guangzhou, China), and the final concentration of cells was 1.2×10^6 cells per well. Each well was supplemented with 2 mL of growth medium containing DMEM/F12 (Corning, NY, USA) and 10% fetal bovine serum (FBS; Tian Hang, Zhejiang, China), then incubated at 37°C in a 5% CO₂ incubator (Thermo Fisher Scientific, Massachusetts, USA).

In this study, the cells transfected with pGPU6/GFP/NEO-shACSF3 vector were used as RNA interference group, and cells transfected with pGPU6/GFP/NEO vector were used as the control. While the cells transfected with pBI-CMV3-ACSF3 vector were used as the ACSF3 gene over-expression group, cells transfected with pBI-CMV3 vector were used as the control. After referring to similar experimental methods in published articles, this experiment adopted a transient transfection method, using the cells 48 h after the vector was transfected into the cells for subsequent experiments (12, 13).

For transfection, each vector DNA (3.0 μg) and 7.5 μL of FuGENE HD transfection reagent (Promega, Madison, Wisconsin, USA) were diluted in 150 μL of DMEM/F12 media and mixed lightly. The mixture was incubated at room temperature for 15 min and then added to the pores of each six-well plate. The cell culture medium was replaced

TABLE 1 | The primer sequences.

Primer	Forward sequences (5'-3')	Reverse sequences (5'-3')	Target sequences
SNPs primer for <i>ACSF3</i>	TGTGACCTCAGTGCCTTCTT	CCTACATTTCTGGGTGCTTCC	GTATAAGGACCTCTACTTGCG
shRNA of bovine <i>ACSF3</i>	AGAGGTATAAGGACCTCTACTTGC GTTCAAGAGACGCAAGTAGAGGTCCTAT ACTTTTTTG	GATCCAAAAAGTATAAGGACCTCTACT TGCCTCTCTTGAACGCAAGTAGAG GTCCTTATAC	
q-PCR primer of <i>ACSF3</i>	GTGGCTGTGATTGGAGTT	TCTCGCTTGTGACCTTC	
<i>ACACA</i>	GAAGGAGCGAGAGGAGTT	GTAGAAGAAGGTGCGTGAA	
<i>ACACB</i>	CAAGATTGCCTCCACCAT	CCTCCTGTAGACTGTGTTC	
<i>MCAT</i>	AAGGCGATTGATGTCAAGA	CTTCCTCCTCTCGTATATGG	
<i>FASN</i>	CACGAACAACAGCCTCTT	GCCTCCAGCACTCTACTA	
<i>CEBPα</i>	TGGACAAGAACAGCAACGAGT	GGTCATTGTCACTGGTCAGCT	
<i>β-actin</i>	AGAGCAAGAGAGGCATCC	TCGTTGTAGAAGGTGTGGT	

after 6 h of transfection. After 36 h of transfection, green fluorescent protein (GFP) expression was detected with a fluorescence microscope (NikonTE2000, Tokyo, Japan). The efficiency of transfection was also determined by calculating the ratio of cells positive for green fluorescent protein expression to the total number of cells. The experiment was repeated three times.

Analysis of mRNA Levels of *ACSF3* in Transfected BMECs

The transfected cells were collected for analysis of the mRNA expression levels. RNAiso Plus reagent (Takara, Dalian, China) was used to extract the total RNA from cultured cells for reverse transcription. cDNA was synthesized using a reverse transcription kit (TransGen Biotech, Beijing, China), 1 μ g of total RNA, 4 μ L of 5 \times EasyScript[®] All-in-One SuperMix for qRT-PCR, and 1 μ L of gDNA remover, and RNA-free water was added to make the total volume of 20 μ L. The cDNA synthesis conditions were: incubate at 42°C for 15 min, and at 85°C for 5 s. Quantitative real-time PCR (qRT-PCR) was followed through with SYBR Green Real-Time PCR Master Mix (Takara, Dalian, China) utilizing the specific primers shown in **Table 1**. qRT-PCR was performed in a 10 μ L reaction with 5 μ moles in 0.5 μ L of forward primer, 5 μ moles in 0.5 μ L of reverse primer, 5 μ L of SYBR Green Real-Time PCR Master Mix (Roche, Basel, Switzerland), 1 μ L of cDNA, and 3 μ L of ddH₂O, using the following procedure: 95°C for 30 s and 45 cycles of 95°C for 5 s, and 60°C for 30 s in a PCRmax (Eco, Staffordshire, UK). Experiments were repeated three times. Three technical replicates were analyzed for each sample and *β -actin* was used as an internal standard to normalize the mRNA expression level using the $2^{-\Delta\Delta CT}$ method.

Determination of Triglyceride Content in Cells of *ACSF3* Interference and Overexpression

In BMECs after interference or the overexpression of *ACSF3*, the detection of TGs was performed with a tissue and cell TG assay kit (Appligen Technologies Inc., Beijing, China) according

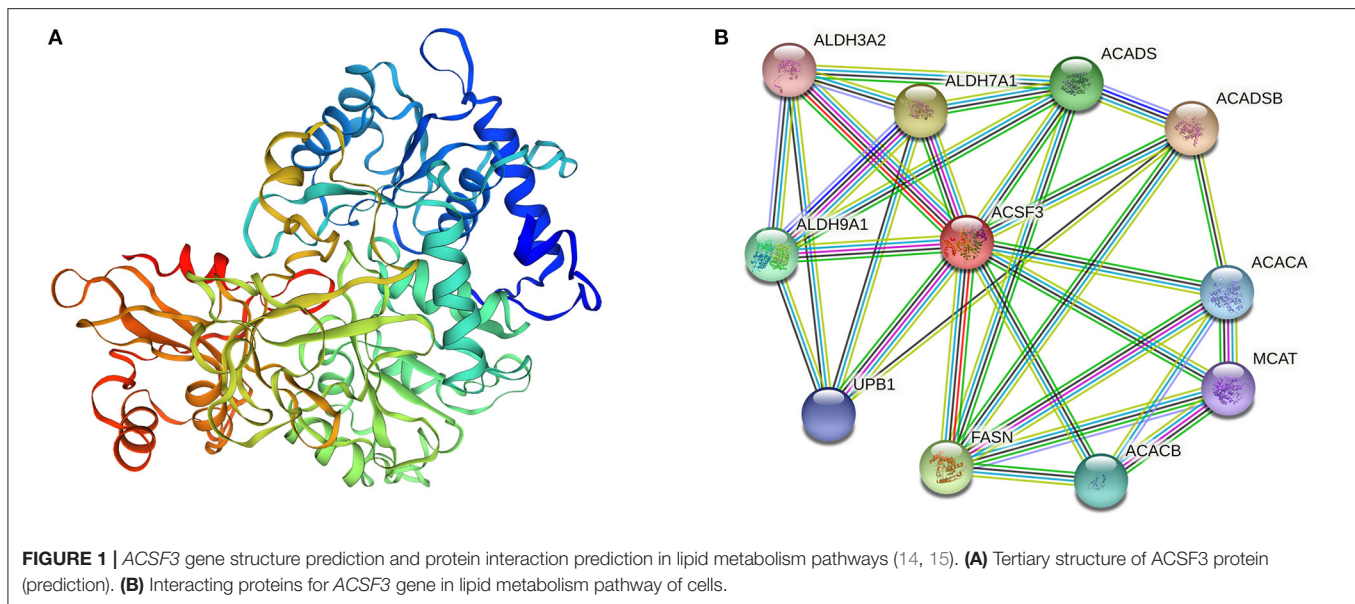
to the manufacturer's protocol. Meanwhile, *ACSF3* was over-expressed in the BFFs and the TG content was examined. The optical density of each sample was determined using a microplate reader (Yong Chuang SM600, Shanghai, China). The experiment was repeated three times and three technical replicates were performed for each sample. The data regarding the triglyceride in the cells were adjusted based on the quantity of the protein, and the cellular content of triglyceride was corrected for the protein concentration of each μ g.

Polymorphic Loci Detection in *ACSF3* Gene and Genotyping

We used a 20 μ L system for PCR amplification, combining 10 pmol/ μ L of each primer, 140 ng of bovine genomic DNA, and 10 μ L of green Taq mix, according to the manufacturer's protocol (Vazyme, Nanjing, China). The PCR amplification conditions were as follows: incubation of the PCR mixture at 95°C for 5 min, 35 cycles of 95°C for 30 s, the annealing temperature of each fragment for 30 s, 72°C for 1,000 bp/min, and a final extension at 72°C for 10 min. The annealing temperature of the polymorphism fragments of *ACSF3* were 60°C. The DNA samples of 135 Chinese Simmental were amplified by PCR. Next, PCR amplification was performed with the primer pairs of SNPs of *ACSF3* as per **Table 1**. The specificity of PCR products was detected by 2% agarose gel electrophoresis. Nucleotide sequences of PCR products were determined by Sanger sequencing (Sangon Biotech Shanghai, China). Briefly, the nucleotide sequence of the amplified product was determined by double-stranded sequencing using the Sanger di-deoxy chain termination method. Finally, the genotypes of the cattle at each SNP were verified by the sequencing results. Haplotype analysis of the SNPs was performed using Haploview v.4.2 (<https://www.broadinstitute.org/haploview/haploview>).

Statistical Analysis

Some relevant data relating to SNPs on the *ACSF3* were calculated according to the genotyping results, which were the genotypic frequency, Hardy-Weinberg test, linkage disequilibrium analysis, and polymorphism information content. The genotype frequencies and allele frequencies were calculated for the



examined Chinese Simmental-cross steers and were analyzed by the significance test. An analysis of the genotypic effects of the *ACSF3* was carried out using the GLM procedure of SPSS 13.0 for Windows.

$$Y_{ijk} = u + y_{si} + m_j + e_{ijk}$$

where Y_{ijk} was the phenotypic observation of the k th individual from the Simmental breed of genotype j in the i th-year season, u was the population mean, y_{si} was the year effect of the i th-year season, m_j was the genotype effect of the genotype j , and e_{ijk} was the random residual effect corresponding to the observed value (10).

RESULTS

ACSF3 Gene Structure Prediction and Protein Interaction Prediction in Lipid Metabolism Pathways

SWISS-MODEL (<https://swissmodel.expasy.org/>) predicted the tertiary structure of the ACSF3 protein (**Figure 1A**) (14). Protein interaction analysis was performed with STRING (<https://cn.string-db.org/>) (15). Bioinformatic data showed that 10 proteins, such as FASN, ACADSB, ACADS, ACACA, and ACACB, can interact with the ACSF3 protein (**Figure 1B**). The fatty acid biosynthesis and metabolism process GO terms in the biological process were enriched, implying that ACSF3 may be involved in the fatty acid biosynthetic process (GO:0006633) and the fatty acid metabolic process (GO:0006631).

The Effect of ACSF3 Gene Expression Level on Triglyceride Content in Cells

According to the sequencing results, the siRNA target oligonucleotide sequences for *ACSF3* were successfully cloned and framed into the *Bam*HI/*Bbs*I sites of the pGPU6/GFP/NEO vector. Meanwhile, the CDS fragments of *ACSF3* were cloned

into the multiple cloning site (*Mlu*I/*Hind*III) of pBI-CMV3 plasmid. The green fluorescence of the transfected cells was observed in each group under fluorescent microscopy after 24 h of transfection (**Figure 2**), indicating that plasmids were successfully transfected into BMECs.

To investigate the effect of the pGPU6/GFP/NEO-shACSF3 and pBI-CMV3-ACSF3 vectors on the bovine *ACSF3* expression level, the relative mRNA expression level of *ACSF3* was analyzed by qRT-PCR. Compared with the control group, the mRNA expression of *ACSF3* in the pGPU6/GFP/NEO-shACSF3 group was significantly reduced ($p < 0.01$, **Figures 2C-I**). Moreover, the expression level of *ACSF3* mRNA in the pBI-CMV3-ACSF3 group was significantly increased compared with the control group ($p < 0.01$, **Figures 2C-II**).

The triglyceride contents of BMECs after transfection with the pGPU6/GFP/NEO-shACSF3 and pBI-CMV3-ACSF3 vectors were investigated. The results demonstrated that, compared with the control group, the content of TGs in the pGPU6/GFP/NEO-shACSF3 group had a decreasing trend (**Figures 2D-I**). In addition, the content of TGs in the pBI-CMV3-ACSF3 group increased significantly ($p < 0.01$, **Figures 2D-II**). Similarly, the results of *ACSF3* on TG content in BFFs are consistent with our validation on BMECs, where an over-expression of the *ACSF3* elevated triglyceride content in both types of cells ($p < 0.01$, **Figures 2E-III**). These results indicated that *ACSF3* could promote the synthesis of triglyceride in the lipid metabolism pathway.

The Effect of ACSF3 Gene Expression on Lipid Metabolism-Related Genes

To further investigate the regulation of *ACSF3* on BMECs lipid metabolism-related genes, the expression levels of the related genes that interacted with the *ACSF3* gene in the lipid metabolism pathway were analyzed. The mRNA expression level of the CCAAT/enhancer binding protein α (*CEBP* α), fatty acid synthase

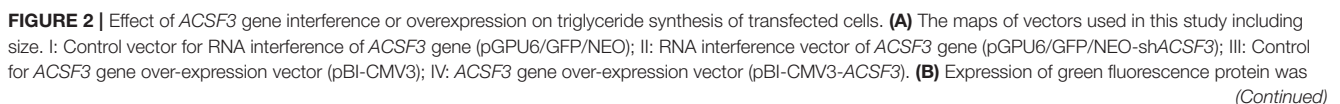


FIGURE 2 | observed in two vector groups under fluorescent microscopy. I: Cells transfected with pGPU6/GFP/Neo vectors; II: Cells transfected with pGPU6/GFP/Neo-sh*ACSF3* vectors; III: Cells transfected with pBI-CMV3 vectors; IV: Cells transfected with pBI-CMV3-*ACSF3* vectors. **(C)** The relative mRNA expression level of *ACSF3* mRNA in the two groups transfected with BMECs. I: Cells transfected with pGPU6/GFP/Neo or pGPU6/GFP/Neo-sh*ACSF3* vectors; II: Cells transfected with pBI-CMV3 or pBI-CMV3-*ACSF3* vectors. **(D)** The two vector groups caused changes in triglyceride content after transfection of BMECs. I: Cells transfected with pGPU6/GFP/Neo-sh*ACSF3* vector caused a decrease in triglyceride content; II: Cells transfected with pBI-CMV3-*ACSF3* vector caused an increase in triglyceride content. **(E)** Overexpression vectors transfected with BFFs caused changes in triglyceride content. I: Cells transfected with pBI-CMV3 vectors; II: Cells transfected with pBI-CMV3-*ACSF3* vectors. III: Cells transfected with pBI-CMV3-*ACSF3* vector caused an increase in triglyceride content. *** means $p < 0.001$, ** means $p < 0.01$.

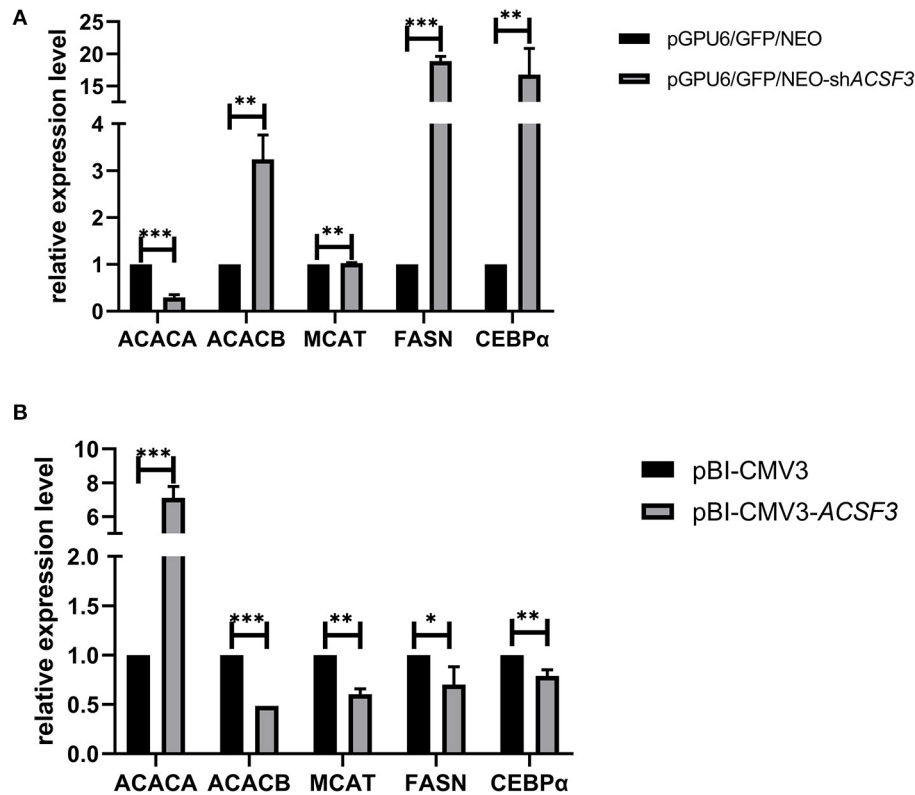


FIGURE 3 | The mRNA expression of lipid metabolism-related genes in BMECs after interference **(A)** or overexpression of **(B)** the *ACSF3* gene. *** means $p < 0.001$, ** means $p < 0.01$, * means $p < 0.05$.

(*FASN*), and acetyl-CoA carboxylase beta (*ACACB*) gene in the pGPU6/GFP/Neo-sh*ACSF3* group was significantly increased compared with the control group ($p < 0.01$, **Figure 3A**). The mRNA expression level of the malonyl-CoA-acyl carrier protein transacylase (*MCAT*) gene in the pGPU6/GFP/Neo-sh*ACSF3* group increased. Nevertheless, the mRNA levels of the acetyl-CoA carboxylase alpha (*ACACA*) gene of the pGPU6/GFP/Neo-sh*ACSF3* group were significantly lower than the control group ($p < 0.01$). In addition, in comparison with the control group, the mRNA expression levels of the above genes in the pBI-CMV3-*ACSF3* group exhibited an opposite trend (**Figure 3B**).

Genetic Diversity of the Functional Domain of *ACSF3* Gene in Chinese Simmental Cattle

Meanwhile, we screened for SNPs in the key functional domain of the bovine *ACSF3* in the Chinese Simmental cattle population.

According to the PCR product sequencing results, there were five polymorphisms (g.14210566 C > T, g.14210668 C > T, g.14210887 T > C, g.14211055 G > A, and g.14211090 A > G) screened in the second exon of the *ACSF3* gene in Chinese Simmental cattle (**Figure 4A**). Among the five loci screened, the g.14211090 A > G locus was a missense mutation (arginine to glutamine), while the other four sites were synonymous mutations. The allele frequency and genotype frequency of five *ACSF3* SNPs in Chinese Simmental cattle are presented in **Figure 4**. Allele C had frequencies of 0.934, 0.798, and 0.954 at the g.14210566 C > T, g.14210668 C > T, and g.14210887 T > C polymorphism sites. Allele A had a frequency of 0.926 at the g.14211055 G > A polymorphism site. Allele G had a frequency of 0.892 at the g.14211090 A > G polymorphism site (**Figure 4B**).

A haplotype analysis was performed on five SNPs loci in this population. Among the different genotypes, the frequency of the three haplotypes (CCCAG, CTCAG, and

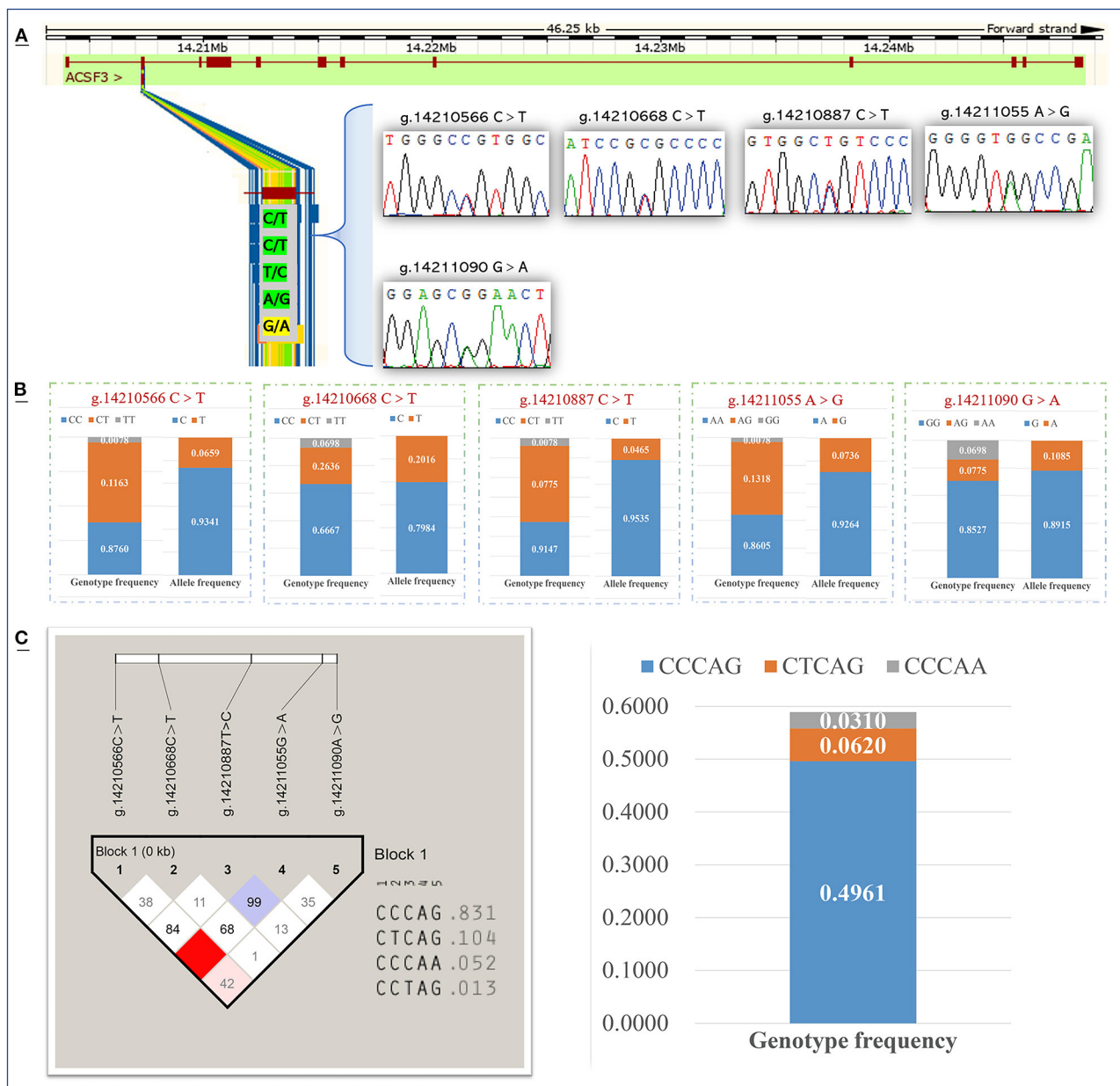


FIGURE 4 | Detection and genotyping of Single-nucleotide polymorphism sites of the *ACSF3* gene in Chinese Simmental cattle. **(A)** Identify five SNPs in the second exon of the *ACSF3* gene. **(B)** Gene frequency and genotype frequency of five SNPs of *ACSF3* gene. **(C)** The haplotypes composed of five SNPs of *ACSF3* gene and the frequency analysis of haplotypes.

CCCAA) was higher than 0.03, and the frequency of the haplotype CCCAG was 0.496, which was the major haplotype (Figure 4C).

Association Analysis of *ACSF3* Gene SNP With Economic Traits and Fatty Acids in Chinese Simmental Cattle

Regarding economic traits, the association analysis results revealed that, at the g.14211090 G > A locus, for the individuals

with the G allele homozygous, a higher value of the eye muscle area, spleen weight, and oxtail weight ($71.33 \pm 8.80 \text{ cm}^2$, $0.75 \pm 0.14 \text{ kg}$, and $1.17 \pm 0.19 \text{ kg}$) were observed compared with the heterozygous ($65.20 \pm 5.90 \text{ cm}^2$, $0.65 \pm 0.10 \text{ kg}$, and $1.03 \pm 0.17 \text{ kg}$) individuals ($p = 0.038$, 0.040 , 0.037 , Table 2). Furthermore, the individuals with the TT genotype (5.99 ± 0.29) at g.14210668 C > T had a significantly higher pH (24h) than individuals with the CT genotype (5.80 ± 0.17 , $p = 0.008$). The values of the dressing percentage in individuals of the TT genotype ($51.82 \pm 1.70\%$) were also significantly higher than

TABLE 2 | Association analysis of five SNPs in the second exon of *ACSF3* gene and economic traits in Chinese Simmental Cattle.

Traits	g.14210566 C>T		g.14210668 C>T			g.14210887 C>T		g.14211055 A>G		g.14211090 G>A		
	CC (113)	CT (15)	CC (86)	CT (34)	TT (9)	CC (118)	TC (10)	AA (111)	GA (17)	AA (9)	AG (10)	GG (110)
LW (kg)	460.77 ± 55.53 ^a	429.47 ± 59.05 ^b	457.94 ± 53.00	458.65 ± 65.46	441.00 ± 55.83	454.84 ± 56.03	464.40 ± 31.88	461.26 ± 55.87 ^a	430.00 ± 55.57 ^b	472.67 ± 86.53	432.25 ± 56.52	457.90 ± 53.38
CW (kg)	234.02 ± 32.07 ^a	214.63 ± 34.54 ^b	232.67 ± 31.29	230.09 ± 36.42	229.18 ± 34.69	230.91 ± 32.84	232.70 ± 16.94	234.24 ± 32.24 ^a	215.48 ± 32.88 ^b	234.02 ± 48.14	214.28 ± 35.29	233.15 ± 30.86
DP (%)	50.73 ± 2.13	49.87 ± 2.35	50.73 ± 2.07	50.13 ± 2.42 ^b	51.82 ± 1.70 ^a	50.69 ± 2.18	50.14 ± 2.22	50.73 ± 2.13	50.01 ± 2.32	49.48 ± 3.29	49.42 ± 2.80	50.86 ± 1.94
BW (kg)	21.19 ± 3.03	19.37 ± 2.39	21.10 ± 3.00	20.70 ± 2.90	21.04 ± 3.72	20.84 ± 2.98	22.20 ± 2.79	21.32 ± 3.04	19.29 ± 2.25	21.62 ± 3.37	19.59 ± 2.67	21.06 ± 2.99
FW (kg)	23.65 ± 2.67	22.61 ± 3.21	23.61 ± 2.52	23.61 ± 3.13	22.50 ± 3.15	23.47 ± 2.76	23.53 ± 1.19	23.68 ± 2.68	22.56 ± 3.00	23.60 ± 3.86	21.98 ± 2.67	23.67 ± 2.62
FHW (kg)	5.71 ± 0.58	5.39 ± 0.70	5.68 ± 0.56	5.68 ± 0.70	5.61 ± 0.64	5.66 ± 0.61	5.90 ± 0.47	5.71 ± 0.58	5.40 ± 0.66	5.76 ± 0.98	5.51 ± 0.69	5.68 ± 0.56
HHW (kg)	4.81 ± 0.56	4.61 ± 0.48	4.76 ± 0.57	4.83 ± 0.53	4.79 ± 0.46	4.77 ± 0.55	4.95 ± 0.54	4.81 ± 0.56	4.61 ± 0.46	4.94 ± 0.73	4.55 ± 0.40	4.79 ± 0.54
TaW (kg)	48.05 ± 5.87	44.32 ± 5.18	47.48 ± 5.88	48.37 ± 6.02	46.00 ± 5.55	47.46 ± 5.81	47.55 ± 4.12	48.09 ± 5.92	44.50 ± 4.88	47.92 ± 7.61	44.40 ± 6.17	47.88 ± 5.67
RRAW (kg)	8.24 ± 0.91	8.04 ± 0.73	8.20 ± 0.83	8.36 ± 0.99	7.71 ± 1.02	8.14 ± 0.86	8.76 ± 0.87	8.26 ± 0.91	7.96 ± 0.72	8.73 ± 1.23	8.03 ± 0.92	8.18 ± 0.86
OmW (kg)	4.57 ± 0.67	4.39 ± 0.71	4.54 ± 0.65	4.50 ± 0.72	4.68 ± 0.75	4.55 ± 0.69	4.55 ± 0.47	4.58 ± 0.67	4.32 ± 0.70	4.93 ± 0.90	4.48 ± 0.74	4.52 ± 0.65
HW (kg)	1.51 ± 0.20	1.38 ± 0.19	1.49 ± 0.21	1.50 ± 0.21	1.47 ± 0.17	1.48 ± 0.20	1.57 ± 0.14	1.51 ± 0.20	1.38 ± 0.18	1.53 ± 0.28	1.39 ± 0.23	1.50 ± 0.19
LW (kg)	4.74 ± 0.65	4.56 ± 0.67	4.72 ± 0.64	4.77 ± 0.70	4.45 ± 0.59	4.72 ± 0.64	4.55 ± 0.70	4.74 ± 0.66	4.57 ± 0.64	4.74 ± 0.87	4.55 ± 0.40	4.72 ± 0.66
LTW (kg)	2.78 ± 0.37	2.53 ± 0.26	2.71 ± 0.36	2.81 ± 0.39	2.90 ± 0.33	2.74 ± 0.37	2.81 ± 0.38	2.79 ± 0.37	2.52 ± 0.25	2.71 ± 0.55	2.59 ± 0.27	2.77 ± 0.36
KW (kg)	1.00 ± 0.13	0.92 ± 0.15	0.98 ± 0.12	1.02 ± 0.15	1.00 ± 0.16	0.99 ± 0.14	0.97 ± 0.09	1.01 ± 0.13	0.91 ± 0.14	1.05 ± 0.12	0.95 ± 0.11	0.99 ± 0.14
RAW (kg)	1.59 ± 0.54	1.54 ± 0.63	1.59 ± 0.57	1.56 ± 0.53	1.60 ± 0.49	1.58 ± 0.54	1.40 ± 0.40	1.60 ± 0.54	1.50 ± 0.59	1.53 ± 0.39	1.50 ± 0.65	1.59 ± 0.55
CPW (kg)	0.53 ± 0.07	0.49 ± 0.07	0.53 ± 0.07	0.51 ± 0.07	0.53 ± 0.05	0.52 ± 0.07	0.55 ± 0.06	0.53 ± 0.07	0.50 ± 0.07	0.50 ± 0.05	0.51 ± 0.09	0.53 ± 0.07
TeW (kg)	0.61 ± 0.12	0.56 ± 0.13	0.62 ± 0.12	0.60 ± 0.13	0.55 ± 0.09	0.61 ± 0.12	0.55 ± 0.11	0.62 ± 0.12	0.55 ± 0.13	0.51 ± 0.10 ^b	0.60 ± 0.14	0.62 ± 0.12 ^a
GFW (kg)	1.10 ± 0.25	0.98 ± 0.36	1.07 ± 0.26	1.17 ± 0.30	1.06 ± 0.25	1.09 ± 0.27	1.05 ± 0.24	1.10 ± 0.25	1.00 ± 0.35	1.17 ± 0.31	0.96 ± 0.39	1.10 ± 0.25
SW (kg)	0.75 ± 0.15	0.67 ± 0.12	0.74 ± 0.14	0.75 ± 0.15	0.73 ± 0.15	0.74 ± 0.15	0.72 ± 0.09	0.75 ± 0.15	0.69 ± 0.12	0.76 ± 0.22	0.65 ± 0.10 ^b	0.75 ± 0.14 ^a
OxW (kg)	1.17 ± 0.19	1.07 ± 0.22	1.18 ± 0.20	1.13 ± 0.19	1.05 ± 0.17	1.15 ± 0.19	1.21 ± 0.16	1.17 ± 0.19	1.07 ± 0.21	1.15 ± 0.22	1.03 ± 0.17 ^b	1.17 ± 0.19 ^a
pH (0 h)	6.81 ± 0.25	6.84 ± 0.18	6.83 ± 0.26	6.80 ± 0.21	6.73 ± 0.21	6.81 ± 0.25	6.82 ± 0.22	6.81 ± 0.25	6.82 ± 0.18	6.71 ± 0.17	6.84 ± 0.23	6.82 ± 0.25
pH (24 h)	5.85 ± 0.19	5.90 ± 0.20	5.87 ± 0.18	5.80 ± 0.17 ^B	5.99 ± 0.29 ^A	5.86 ± 0.19	5.80 ± 0.13	5.85 ± 0.19	5.88 ± 0.20	5.93 ± 0.16	5.92 ± 0.20	5.85 ± 0.19
CL (cm)	146.50 ± 5.84	142.47 ± 6.12	145.98 ± 5.81	146.38 ± 6.35	144.56 ± 6.78	145.87 ± 6.03	146.10 ± 4.53	146.55 ± 5.85	142.59 ± 5.95	147.44 ± 6.29	143.20 ± 6.20	146.12 ± 5.93
CD (cm)	65.54 ± 3.04	64.93 ± 4.23	65.44 ± 3.02	65.68 ± 3.67	65.33 ± 3.00	65.48 ± 3.24	65.50 ± 2.59	65.55 ± 3.03	64.94 ± 4.10	67.78 ± 3.46 ^a	65.40 ± 4.58	65.31 ± 2.97 ^b
CBD (cm)	66.53 ± 3.60	65.87 ± 3.60	66.64 ± 3.46	66.09 ± 3.93	66.44 ± 3.81	66.44 ± 3.67	66.35 ± 2.33	66.54 ± 3.62	65.88 ± 3.44	66.94 ± 3.71	66.00 ± 3.77	66.49 ± 3.60
HLC (cm)	48.20 ± 3.97	47.00 ± 4.39	48.49 ± 3.99	46.93 ± 3.95	48.53 ± 4.08	48.03 ± 4.06	48.53 ± 3.81	48.19 ± 4.00	47.18 ± 4.17	46.94 ± 3.43	47.85 ± 4.61	48.20 ± 4.02
HLW (cm)	44.68 ± 2.62	44.13 ± 3.20	44.74 ± 2.65	44.00 ± 2.87	45.61 ± 1.82	44.56 ± 2.77	45.10 ± 1.56	44.70 ± 2.64	44.06 ± 3.00	44.56 ± 3.39	44.85 ± 3.16	44.59 ± 2.60
HLL (cm)	84.39 ± 2.76	84.37 ± 2.58	84.43 ± 2.79	84.54 ± 2.53	83.48 ± 2.85	84.31 ± 2.75	85.40 ± 2.38	84.41 ± 2.87	84.24 ± 2.46	84.11 ± 2.76	83.90 ± 2.51	84.46 ± 2.75
TMT (cm)	17.44 ± 1.54	17.41 ± 1.57	17.47 ± 1.64	17.42 ± 1.45	17.42 ± 1.01	17.49 ± 1.53	17.11 ± 1.80	17.44 ± 1.55	17.36 ± 1.51	17.20 ± 1.56	17.92 ± 1.53	17.43 ± 1.55
WMT (cm)	6.02 ± 0.44	5.81 ± 0.39	6.01 ± 0.43	5.96 ± 0.48	5.93 ± 0.33	5.99 ± 0.43	6.00 ± 0.42	6.02 ± 0.44	5.81 ± 0.37	6.04 ± 0.55	5.81 ± 0.45	6.00 ± 0.42
BFT (cm)	0.27 ± 0.11	0.27 ± 0.12	0.28 ± 0.11	0.26 ± 0.12	0.26 ± 0.10	0.28 ± 0.12	0.21 ± 0.07	0.27 ± 0.11	0.28 ± 0.13	0.29 ± 0.15	0.29 ± 0.10	0.27 ± 0.12
FCR (%)	22.24 ± 8.05	22.20 ± 7.48	22.66 ± 7.99	20.74 ± 7.76	24.44 ± 8.08	22.34 ± 8.00	20.80 ± 7.44	22.19 ± 8.08	22.53 ± 7.27	21.67 ± 5.66	21.10 ± 8.86	22.44 ± 8.06
MS	5.88 ± 0.35	5.87 ± 0.35	5.90 ± 0.34	5.82 ± 0.39	6.00 ± 0.00	5.88 ± 0.35	6.00 ± 0.00	5.88 ± 0.35	5.88 ± 0.33	5.89 ± 0.33	5.90 ± 0.32	5.88 ± 0.35
EMA (cm ²)	71.82 ± 9.04 ^a	64.53 ± 5.38 ^b	71.62 ± 8.71	69.15 ± 9.58	72.11 ± 8.87	70.86 ± 8.87	72.10 ± 10.71	71.99 ± 9.02 ^A	64.29 ± 5.16 ^B	73.44 ± 11.84 ^a	65.20 ± 5.90 ^{ab}	71.33 ± 8.80 ^b
MC	4.95 ± 0.83	4.93 ± 0.88	4.99 ± 0.85	4.76 ± 0.78	5.33 ± 0.87	4.97 ± 0.86	4.90 ± 0.57	4.96 ± 0.83	4.82 ± 0.88	4.89 ± 0.60	4.90 ± 0.88	4.96 ± 0.86
FCS	3.59 ± 0.62 ^a	3.13 ± 0.64 ^b	3.47 ± 0.65	3.59 ± 0.56	3.89 ± 0.93	3.53 ± 0.66	3.50 ± 0.53	3.59 ± 0.62 ^a	3.18 ± 0.64 ^b	3.56 ± 0.73	3.20 ± 0.92	3.55 ± 0.61

^{a,b}Means with different letters were significant difference ($p < 0.05$). ^{A,B}Means with different letters were significant difference ($p < 0.01$). SD, standard deviation; LW, liveweight; CW, carcass weight; DP, dressing percentage; BW, bone weight; FW, front weight; FHW, front hoof weight; HHW, hind hoof weight; TaW, tare weight; RRAW, rumen, reticulum and abomasum weight; OmW, omasum weight; HW, heart weight; LW, liver weight; LTW, lung and trachea weight; KW, kidney weight; RAW, renal adipose weight; CPW, cow penis weight; TeW, testicular weight; GFW, genital fat weight; SW, spleen weight; OxW, oxtail weight; CL, carcass length; CD, carcass depth; CBD, carcass breast depth; HLC, hind legs circumference; HLW, hind legs width; HLL, hind legs length; TMT, thigh meat thickness; WMT, waist meat thickness; BFT, back-fat thickness; FCR, carcass fat coverage rate; MS, marbling score; EMA, eye muscle area; MC, meat color; FCS, fat color score.

in that of the CT genotype ($50.13 \pm 2.42\%$, $p = 0.038$). The g.14210566 C > T and g.14211055 A > G loci showed a strong linkage relationship ($r^2 = 0.887$, LOD = 19.92).

Regarding fatty acid composition, as shown in **Table 3**, the linoleic acid content of the AG genotype (0.131 ± 0.056 g/100g) at the g.14211090 G > A locus was significantly higher than that of the GG genotype (0.101 ± 0.032 g/100g, $p = 0.008$, **Table 3**). The individuals with the TT genotype (0.130 ± 0.054 g/100g) at the g.14210668 C > T locus had higher linoleic acid than the other two genotypes (0.102 ± 0.035 g/100g, 0.100 ± 0.029 g/100g, $p = 0.023$, $p = 0.022$). Similarly, individuals with the TC genotype (0.129 ± 0.054 g/100g) at the g.14210887 C > T locus had higher levels of linoleic acid than the other genotype (0.102 ± 0.032 g/100g, $p = 0.017$).

Correlation Analysis of Polymorphic Locus Haplotypes, Economic Traits and Fatty Acids in Chinese Simmental Cattle

The haplotype CTCAG was significantly correlated with pH (24h) and fat color score (6.03 ± 0.27 , 5.87 ± 0.18 , 4.00 ± 0.93 , 3.51 ± 0.59 , $p = 0.022$, and $p = 0.042$, **Table 4**), and the weight of the lung and trachea was significantly higher than in other haplotype individuals (2.90 ± 0.35 kg, 2.23 ± 0.30 kg, $p = 0.002$). The front hoof weight of the haplotype CCCAG (5.71 ± 0.53 kg) was significantly higher than that of CCCAA (5.13 ± 0.52 kg, $p = 0.044$), the weight of oxtail (1.20 ± 0.19 kg) was significantly higher than that of CTCAG (1.03 ± 0.16 kg, $p = 0.015$), and the weight of the testis (0.63 ± 0.12 kg) was significantly higher than that of CCCAA (0.46 ± 0.10 kg, $p = 0.004$). Similar to the results of the correlation analysis between SNPs and fatty acid content, the content of linoleic acid in haplotype CTCAG (0.130 ± 0.058 g/100g) was significantly higher than that of haplotype CCCAG (0.097 ± 0.028 g/100g, $p < 0.01$, **Table 5**).

DISCUSSION

The candidate genes associated with carcass and meat quality traits have been extensively validated in the livestock industry in the expectation of obtaining better economic characteristics. Single nucleotide polymorphisms as DNA markers for genetic and molecular breeding preferences, combining gene function studies with association analysis is more conducive to fully explore the relationships and characteristics of genes. We performed transcriptome sequencing analysis in cattle with different fat deposition traits and obtained differentially expressed genes, including ACSF3. Acyl-CoA synthetase family member 3 (ACSF3), an essential enzyme that activates fatty acids through the formation of thioester bonds to form acyl-CoA, serves as the substrate for both *de novo* fatty acid synthesis and oxidation (16). ACSF3 is valuable to be studied as a gene that plays a role in the process of lipid metabolism.

In this study, we successfully constructed the RNA interference and over-expression vectors of ACSF3, and found that ACSF3 mRNA levels were positively correlated with

triglyceride content in cells. Five SNP loci were screened in a population of 135 Chinese Simmental cattle, and haplotype analysis and association analysis were performed on these five SNP loci. The association between SNP loci, haplotypes and economic traits and fatty acid composition was investigated, and the results showed that these five SNP loci and haplotypes were significantly associated with economic traits and fatty acid composition of Chinese Simmental cattle.

Both milk lipid metabolism and meat lipid metabolism are of interest in livestock industry research, and in the present study, we were surprised to find that increased mRNA levels of ACSF3 caused an increase in triglyceride content in both BMECs and BFFs. This suggested to us that ACSF3 may play a similar role in lipid metabolism, without tissue specificity. This may be due to main role of ACSF3 is to activate the toxic, endogenous antimetabolite malonate into malonyl-CoA that can be decarboxylated to acetyl-CoA and therefore fully oxidized within the TCA cycle. And one function of ACSF3 that has been proposed was that it generated malonyl-CoA in the matrix of the mitochondria to enable mitochondrial type II (mtFASII) fatty acid synthesis (17, 18).

We found that it was co-annotated with genes such as FASN and ACACA in the fatty acid biosynthesis and fatty acid metabolism pathway of KEGG by predicting the ACSF3 protein interaction network. We hypothesized that ACSF3 may affect genes related to lipid metabolism and thus affect lipid metabolism. In the mitochondria, ACSF3 links malonate to CoA, producing malonyl-CoA. In addition, when ACACA is inhibited, TG synthesis will also be inhibited, and at the same time, fatty acid synthesis is inhibited and fatty acid oxidation is stimulated (19). The results of our qRT-PCR showed that the mRNA expression levels of ACACA and ACSF3 changed consistently. This may be due to the coordination of mammalian ACACA and ACSF3 to produce the malonyl-CoA required for mitochondrial fatty acid synthesis, indicating that ACACA also plays an essential role in acetyl-CoA sensing. In mitochondria, malonyl-CoA is the main signal of energy metabolism (20). It has been proved that FASN inhibition leads to a rapid increase in malonyl-CoA in cells (21). The result may be the reason why we found that ACSF3 and FASN have a negative regulatory relationship at the mRNA level.

We screened five SNP loci in Chinese Simmental cattle population, where the SNP of the g.14211090 A > G loci triggered missense mutations that resulted in the conversion of the encoded amino acid from arginine to glutamine. In human metabolic diseases research, ACSF3 as a cause of combined malonic and methylmalonic aciduria (CMAMMA) has been studied. Mutations in ACSF3 were identified, encoding the putative methylmalonyl-CoA and malonyl-CoA synthase as the cause of CMAMMA (16). A similar study found a homozygous missense allele in the non-classical CMAMMA candidate gene ACSF3 in patients (22). Hence, it is worth investigating whether the mutation of ACSF3 could be used as an early molecular marker for the detection of lipid metabolism in bovine.

For the carcass traits of Chinese Simmental cattle, the g.14210566 C > T and g.14211055 A > G loci were significantly associated with the liveweight and carcass weight. These results suggest that these two SNPs may be associated with growth

TABLE 3 | Association analysis of five SNPs in the second exon of *ACSF3* gene and fatty acids in Chinese Simmental Cattle.

Types of fatty acids (g/100g)	g.14210566 C>T		g.14210668 C>T			g.14210887 C>T		g.14211055 A>G		g.14211090 G>A		
	CC (113)	CT (15)	CC (86)	CT (34)	TT (9)	CC (118)	TC (10)	AA (111)	GA (17)	AA (9)	AG (10)	GG (110)
Myristic acid	0.020 ± 0.017	0.024 ± 0.018	0.020 ± 0.017	0.021 ± 0.019	0.023 ± 0.013	0.020 ± 0.017	0.025 ± 0.019	0.020 ± 0.017	0.025 ± 0.017	0.025 ± 0.022	0.029 ± 0.021	0.020 ± 0.016
Myristoleic acid	0.002 ± 0.005	0.003 ± 0.004	0.002 ± 0.005	0.003 ± 0.004	0.002 ± 0.002	0.002 ± 0.005	0.004 ± 0.005	0.002 ± 0.005	0.002 ± 0.004	0.003 ± 0.005	0.003 ± 0.004	0.002 ± 0.005
Palmitic acid	0.258 ± 0.191	0.300 ± 0.184	0.265 ± 0.198	0.248 ± 0.181	0.296 ± 0.149	0.257 ± 0.190	0.313 ± 0.189	0.254 ± 0.191	0.318 ± 0.180	0.306 ± 0.211	0.347 ± 0.210	0.251 ± 0.185
Palmitoleic acid	0.027 ± 0.031	0.031 ± 0.021	0.028 ± 0.034	0.026 ± 0.020	0.024 ± 0.013	0.027 ± 0.030	0.032 ± 0.023	0.027 ± 0.031	0.031 ± 0.020	0.029 ± 0.021	0.037 ± 0.023	0.027 ± 0.031
Margaric acid	0.011 ± 0.007	0.013 ± 0.007	0.011 ± 0.007	0.012 ± 0.008	0.014 ± 0.007	0.011 ± 0.007	0.014 ± 0.008	0.011 ± 0.007	0.014 ± 0.007	0.015 ± 0.010	0.015 ± 0.009	0.011 ± 0.007
Heptadecenoic acid	0.005 ± 0.007	0.007 ± 0.005	0.006 ± 0.007	0.005 ± 0.006	0.003 ± 0.005	0.005 ± 0.007	0.005 ± 0.006	0.005 ± 0.007	0.007 ± 0.005	0.005 ± 0.006	0.009 ± 0.005	0.005 ± 0.007
Stearic acid	0.187 ± 0.112	0.221 ± 0.111	0.187 ± 0.104	0.188 ± 0.126	0.235 ± 0.131	0.186 ± 0.110	0.238 ± 0.140	0.184 ± 0.110	0.240 ± 0.117	0.241 ± 0.178	0.250 ± 0.123	0.181 ± 0.103
Oleic acid	0.365 ± 0.378	0.394 ± 0.224	0.379 ± 0.414	0.329 ± 0.230	0.404 ± 0.201	0.363 ± 0.372	0.424 ± 0.237	0.362 ± 0.381	0.410 ± 0.215	0.397 ± 0.271	0.462 ± 0.245	0.357 ± 0.377
Linoleic acid	0.101 ± 0.033 ^b	0.120 ± 0.050 ^a	0.102 ± 0.035 ^b	0.100 ± 0.029 ^b	0.130 ± 0.054 ^a	0.102 ± 0.032 ^b	0.129 ± 0.054 ^a	0.101 ± 0.032 ^b	0.122 ± 0.047 ^a	0.106 ± 0.036	0.131 ± 0.056 ^A	0.101 ± 0.032 ^B
α-linolenic acid	0.006 ± 0.006	0.009 ± 0.013	0.007 ± 0.008	0.005 ± 0.005	0.010 ± 0.010	0.006 ± 0.006 ^B	0.013 ± 0.015 ^A	0.006 ± 0.006	0.009 ± 0.012	0.005 ± 0.005	0.011 ± 0.016 ^a	0.006 ± 0.006 ^b
Arachic acid	0.001 ± 0.003	0.001 ± 0.002	0.000 ± 0.001	0.001 ± 0.005	0.001 ± 0.001	0.001 ± 0.003	0.001 ± 0.002	0.001 ± 0.003	0.001 ± 0.002	0.001 ± 0.001	0.002 ± 0.002	0.001 ± 0.003
Eicosanic acid	0.001 ± 0.002	0.000 ± 0.001	0.000 ± 0.002	0.001 ± 0.002	0.000 ± 0.001	0.000 ± 0.002	0.001 ± 0.001	0.001 ± 0.002	0.000 ± 0.001	0.000 ± 0.001	0.001 ± 0.001	0.000 ± 0.002
Dihomo-γ-linolenic acid	0.010 ± 0.003	0.010 ± 0.003	0.010 ± 0.003	0.010 ± 0.003	0.009 ± 0.002	0.010 ± 0.003	0.011 ± 0.002	0.010 ± 0.003	0.010 ± 0.003	0.010 ± 0.001	0.011 ± 0.003	0.010 ± 0.003
Arachidonic acid	0.049 ± 0.013	0.055 ± 0.019	0.050 ± 0.014	0.051 ± 0.013	0.049 ± 0.011	0.050 ± 0.013	0.056 ± 0.018	0.049 ± 0.013	0.053 ± 0.018	0.051 ± 0.012	0.059 ± 0.020 ^a	0.049 ± 0.013 ^b

^{a,b} Means with different letters were significant difference ($p < 0.05$), ^{A,B} Means with different letters were significant difference ($p < 0.01$). SD, standard deviation.

TABLE 4 | Association analysis of three haplotypes and economic traits in Chinese Simmental Cattle.

Types of fatty acids (g/100g)	Haplotype (Mean \pm SD)		
	Hap1 CCCAA (4)	Hap2 CCCAG (63)	Hap3 CTCAG (8)
Liveweight (kg)	423.63 \pm 21.73	461.65 \pm 50.83	442.94 \pm 59.36
Carcass weight (kg)	213.00 \pm 18.72	235.35 \pm 30.33	229.88 \pm 37.02
Dressing percentage	50.22 \pm 2.00	50.91 \pm 2.01	51.73 \pm 1.80
Bone weight (kg)	18.68 \pm 0.84	21.26 \pm 3.04	20.88 \pm 3.94
Front weight (kg)	21.44 \pm 1.56	23.90 \pm 2.52	22.63 \pm 3.35
Front hoof weight (kg)	5.13 \pm 0.52 ^b	5.71 \pm 0.53 ^a	5.59 \pm 0.68
Hind hoof weight (kg)	4.38 \pm 0.48	4.82 \pm 0.59	4.83 \pm 0.47
Tare weight (kg)	43.20 \pm 6.94	48.13 \pm 5.59	46.55 \pm 5.67
Rumen, reticulum and abomasum weight (kg)	7.79 \pm 0.18	8.25 \pm 0.83	7.81 \pm 1.05
Omasum weight (kg)	4.75 \pm 0.35	4.59 \pm 0.66	4.79 \pm 0.73
Heart weight (kg)	1.35 \pm 0.09	1.50 \pm 0.20	1.47 \pm 0.18
Liver weight (kg)	4.19 \pm 0.41	4.78 \pm 0.62	4.48 \pm 0.62
Lung and trachea weight (kg)	2.23 \pm 0.30 ^{AB}	2.79 \pm 0.35 ^B	2.90 \pm 0.35 ^A
Kidney weight (kg)	0.98 \pm 0.04	1.00 \pm 0.13	1.02 \pm 0.16
Renal adipose weight (kg)	1.45 \pm 0.20	1.63 \pm 0.57	1.70 \pm 0.40
Cow penis weight (kg)	0.48 \pm 0.04	0.53 \pm 0.07	0.52 \pm 0.05
Testicular weight (kg)	0.46 \pm 0.10 ^B	0.63 \pm 0.12 ^A	0.56 \pm 0.09
Genital fat weight (kg)	1.07 \pm 0.33	1.10 \pm 0.22	1.07 \pm 0.27
Spleen weight (kg)	0.63 \pm 0.09	0.76 \pm 0.14	0.73 \pm 0.16
Oxtail weight (kg)	1.05 \pm 0.12	1.20 \pm 0.19 ^a	1.03 \pm 0.16 ^b
pH(0 h)	6.64 \pm 0.18	6.84 \pm 0.27	6.75 \pm 0.22
pH(24 h)	5.95 \pm 0.15	5.87 \pm 0.18 ^b	6.03 \pm 0.27 ^a
Carcass length (cm)	143.00 \pm 3.74	146.70 \pm 5.71	144.25 \pm 7.19
Carcass depth (cm)	66.00 \pm 2.31	65.48 \pm 3.01	65.25 \pm 3.20
Carcass breast depth (cm)	64.88 \pm 0.85	66.78 \pm 3.69	66.25 \pm 4.03
Hind legs circumference (cm)	44.75 \pm 1.50	48.66 \pm 4.11	47.50 \pm 2.83
Hind legs width (cm)	43.25 \pm 3.20	44.76 \pm 2.63	45.50 \pm 1.91
Hind legs length (cm)	82.13 \pm 2.46	84.61 \pm 2.89	83.54 \pm 3.04
Thigh meat thickness (cm)	16.80 \pm 0.89	17.54 \pm 1.64	17.31 \pm 1.02
Waist meat thickness (cm)	5.75 \pm 0.50	6.04 \pm 0.44	5.95 \pm 0.35
Back-fat thickness (cm)	0.28 \pm 0.22	0.28 \pm 0.11	0.26 \pm 0.11
Carcass fat coverage rate (%)	21.50 \pm 2.38	22.70 \pm 8.09	25.63 \pm 7.76
Marbling score	5.75 \pm 0.50	5.92 \pm 0.33	6.00 \pm 0.00
Eye muscle area (cm ²)	67.00 \pm 9.63	73.14 \pm 8.57	71.25 \pm 9.07
Meat color	4.50 \pm 0.58	5.11 \pm 0.86	5.38 \pm 0.92
Fat color score	3.50 \pm 0.58	3.51 \pm 0.59 ^b	4.00 \pm 0.93 ^a

^{a,b}Means with different letters were significant difference ($p < 0.05$), ^{A,B}Means with different letters were significant difference ($p < 0.01$). SD, standard deviation.

and development in cattle. For meat quality traits of Chinese Simmental cattle, the g.14210566 C > T, g.14211055 A > G, and g.14211090 G > A loci were found to be significantly associated with the eye muscle area. In Angus bulls, eye muscle area had positive genetic (57%) and phenotypic (56%) correlations with liveweight (23). The reason why the g.14210566 C > T and g.14211055 A > G loci were found to be associated with live weight and eye muscle area was attributed to the fact that eye muscle area is genetically associated with live weight and meat quality.

Similarly, the results of an association analysis of pork quality traits and whole genomes showed that ACSF3 was a candidate gene related to intramuscular fat (24). We examined the fatty acid composition and content of intramuscular fat, rather than measuring the percentage of intramuscular fat. We were surprised to find that all the five SNPs obtained from the screening were significantly associated with linoleic acid. Linoleic acid is one of the essential fatty acids in the human body. It plays a meaningful role in maintaining many physiological functions of the human body, such as participating

TABLE 5 | Association analysis of three haplotypes and fatty acids in Chinese Simmental Cattle.

Types of fatty acids (g/100g)	Haplotype (Mean \pm SD)		
	Hap1 CCCAA (4)	Hap2 CCCAG (63)	Hap3 CTCAG (8)
Myristic acid	0.013 \pm 0.005	0.020 \pm 0.017	0.023 \pm 0.014
Myristoleic acid	0.000 \pm 0.001	0.002 \pm 0.006	0.002 \pm 0.002
Palmitic acid	0.180 \pm 0.014	0.261 \pm 0.212	0.302 \pm 0.159
Palmitoleic acid	0.017 \pm 0.005	0.029 \pm 0.039	0.024 \pm 0.014
Margaric acid	0.009 \pm 0.002	0.010 \pm 0.007	0.015 \pm 0.008
Heptadecenoic acid	0.002 \pm 0.004	0.006 \pm 0.008	0.003 \pm 0.005
Stearic acid	0.149 \pm 0.034	0.178 \pm 0.102	0.243 \pm 0.138
Oleic acid	0.247 \pm 0.017	0.387 \pm 0.475	0.409 \pm 0.214
Linoleic acid	0.098 \pm 0.022	0.097 \pm 0.028 ^B	0.130 \pm 0.058 ^A
α -linolenic acid	0.006 \pm 0.005	0.005 \pm 0.006	0.010 \pm 0.011
Arachic acid	0.000 \pm 0.001	0.000 \pm 0.001	0.001 \pm 0.002
Eicosanic acid	0.000 \pm 0.001	0.000 \pm 0.003	0.001 \pm 0.001
Dihomo- γ -linolenic acid	0.010 \pm 0.000	0.010 \pm 0.003	0.009 \pm 0.002
Arachidonic acid	0.053 \pm 0.012	0.049 \pm 0.013	0.047 \pm 0.011

^{A,B}Means with different letters were significant difference ($p < 0.01$). SD, standard deviation.

in the synthesis of phospholipids and the metabolism of other lipids, and has the effect of significantly lowering serum cholesterol. Studies have shown that *ACSF3* is significantly up-regulated in the process of alcoholic liver disease, participates in fatty acid and lipid metabolism, and accelerates liver damage (25). Therefore, the SNPs of *ACSF3* can be used as a molecular marker for breeding cattle with high linoleic acidity. The effect of bovine *ACSF3* on lipid metabolism is deserving of a follow-up in-depth study. Meanwhile, two SNPs (g.14210887 C > T and g.14211090 G > A) were significantly associated with α -linolenic acid. Linoleic acid and α -linolenic acid are both beneficial to human health as polyunsaturated fatty acids. Breeding individuals with a high polyunsaturated fatty acid content in muscle is the goal of specialized beef breed selection and high-quality beef production. As a consequence, it is of great significance to screen the genes and genetic markers that determine the polyunsaturated fatty acids and to select high-quality beef cattle through the application of molecular breeding techniques such as marker-assisted selection.

REFERENCES

1. Ferraz JB, Pinto LF, Meirelles FV, Eler JP, De Rezende FM, Oliveira EC, et al. Association of single nucleotide polymorphisms with carcass traits

The haplotype consisting of five SNPs was also associated with linoleic acid. In an association analysis between SNPs or haplotypes and meat quality traits, only the two traits of fat color score and pH (24 h) were significantly associated. The traits that were considered to be significant in the association analysis of SNPs with traits did not exactly match the results of haplotypes, which may be due to the relatively small population of cattle collected in the current experiment.

In conclusion, the function of the *ACSF3* in bovine lipid metabolism was preliminarily analyzed and we proposed an association analysis report combining SNPs with economic traits and fatty acid composition to support molecular marker-assisted selection to predict the association analysis of SNPs with economic traits of beef cattle.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Protection and Use Committee of Jilin University [permit no. SYXK (Ji) pzp20181227083].

AUTHOR CONTRIBUTIONS

WH and XF: conceptualization, validation, and writing—original draft. WH and XL: formal analysis. RY: funding acquisition, project administration, and writing—review & editing. WH, YL, and JL: methodology. ZZ: resources. GL: software. WH and GL: visualization. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31972993), the Jilin Scientific and Technological Development Program (20180101275JC), and Graduate Innovation Fund of Jilin University (101832020CX333).

ACKNOWLEDGMENTS

The authors would like to thank the Jiaqi Mi and Xiuqi Zhang for technical support, also thank the Meat Laboratory of the Chinese Academy of Agricultural Sciences for providing assistance with carcass and meat quality determinations.

in nellore cattle. *Genet Mol Res.* (2009) 8:1360–6. doi: 10.4238/vol8-4 Vgmr650

2. Sasazaki S, Akiyama K, Narukami T, Matsumoto H, Oyama K, Mannen H. UTS2R gene polymorphisms are associated with fatty acid composition

- in Japanese beef cattle. *Anim Sci J.* (2014) 85:499–505. doi: 10.1111/asj.12167
3. Witkowski A, Thweatt J, Smith S. Mammalian ACSF3 protein is a Malonyl-CoA synthetase that supplies the chain extender units for mitochondrial fatty acid synthesis. *J Biol Chem.* (2011) 286:33729–36. doi: 10.1074/jbc.M111.291591
 4. Lombard DB, Zhao YM. ACSF3 and Mal(onate)-Adapted mitochondria. *Cell Chem Biol.* (2017) 24:649–50. doi: 10.1016/j.chembiol.2017.06.004
 5. Chen H, Kim HU, Weng H, Browse J. Malonyl-CoA synthetase, encoded by ACYL ACTIVATING ENZYME13, is essential for growth and development of arabidopsis. *Plant Cell.* (2011) 23:2247–62. doi: 10.1105/tpc.111.086140
 6. Monteuis G, Suomi F, Keratar JM, Masud AJ, Kastaniotis AJ. A conserved mammalian mitochondrial isoform of acetyl-CoA carboxylase ACC1 provides the malonyl-CoA essential for mitochondrial biogenesis in tandem with ACSF3. *Biochem J.* (2017) 474:3783–97. doi: 10.1042/BCJ20170416
 7. Bovo S, Bertolini F, Schiavo G, Mazzoni G, Dall'olio S, Fontanesi L. Reduced representation libraries from DNA pools analysed with next generation semiconductor based-sequencing to identify SNPs in extreme and divergent pigs for back fat thickness. *Int J Genomics.* (2015) 2015:950737. doi: 10.1155/2015/950737
 8. Edea Z, Jung KS, Shin SS, Yoo SW, Choi JW, Kim KS. Signatures of positive selection underlying beef production traits in Korean cattle breeds. *J Anim Sci Technol.* (2020) 62:293–305. doi: 10.5187/jast.2020.62.3.293
 9. Lu CY, Yang RJ, Liu BY, Li ZZ, Shen BL, Yan SQ, et al. Establishment of two types of mammary epithelial cell lines from chinese holstein dairy cow. *J Animal Vet Adv.* (2012) 11:1166–72. doi: 10.3923/javaa.2012.1166.1172
 10. Fang XB, Zhang LP, Yu XZ, Li JY, Lu CY, Zhao ZH, et al. Association of HSL gene E1-c.276C > T and E8-c.51C > T mutation with economical traits of Chinese simmental cattle. *Mol Biol Rep.* (2014) 41:105–12. doi: 10.1007/s11033-013-2842-6
 11. Yu HB, Zhao ZH, Yu XZ, Li JY, Lu CY, Yang RJ. Bovine lipid metabolism related gene GPAM: molecular characterization, function identification, and association analysis with fat deposition traits. *Gene.* (2017) 609:9–18. doi: 10.1016/j.gene.2017.01.031
 12. Liao L, Zheng R, Wang C, Gao J, Ying Y, Ning Q, et al. The influence of down-regulation of suppressor of cellular signaling proteins by RNAi on glucose transport of intrauterine growth retardation rats. *Pediatr Res.* (2011) 69:497–503. doi: 10.1203/PDR.0b013e31821769bd
 13. Shu K, Xiao Z, Long S, Yan J, Yu X, Zhu Q, et al. Expression of DJ-1 in endometrial cancer: close correlation with clinicopathological features and apoptosis. *Int J Gynecol Cancer.* (2013) 23:1029–35. doi: 10.1097/IGC.0b013e3182959182
 14. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* (2018) 46:W296–303. doi: 10.1093/nar/gky427
 15. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* (2021) 49:D605–12. doi: 10.1093/nar/gkaa1074. Erratum in: *Nucleic Acids Res.* (2021) 49:10800.
 16. Sloan JL, Johnston JJ, Manoli I, Chandler RJ, Krause C, Carrillo-Carrasco N, et al. Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. *Nature Genetics.* (2011) 43:883–U96. doi: 10.1038/ng.908
 17. Guan X, Nikolau BJ. AAE13 encodes a dual-localized malonyl-CoA synthetase that is crucial for mitochondrial fatty acid biosynthesis. *Plant J.* (2016) 85:581–93. doi: 10.1111/tpj.13130
 18. Bowman CE, Wolfgang MJ. Role of the malonyl-CoA synthetase ACSF3 in mitochondrial metabolism. *Adv Biol Regul.* (2019) 71:34–40. doi: 10.1016/j.jbior.2018.09.002
 19. Harwood HJ, Petras SE, Shelly LD, Zaccaro LM, Perry DA, Makowski MR, et al. Isozyme-nonselective N-substituted bipiperidylcarboxamide acetyl-CoA carboxylase inhibitors reduce tissue malonyl-CoA concentrations, inhibit fatty acid synthesis, and increase fatty acid oxidation in cultured cells and in experimental animals. *J Biol Chem.* (2003) 278:37099–111. doi: 10.1074/jbc.M304481200
 20. Bowman CE, Rodriguez S, Alperin ESS, Acoba MG, Zhao L, Hartung T, et al. The mammalian Malonyl-CoA synthetase ACSF3 is required for mitochondrial protein malonylation and metabolic efficiency. *Cell Chem Biol.* (2017) 24:673. doi: 10.1016/j.chembiol.2017.04.009
 21. Pizer ES, Thupari J, Han WF, Pinn ML, Chrest FJ, Townsend CA, et al. Malonyl-coenzyme-A is a potential mediator of cytotoxicity induced by fatty acid synthase inhibition in human breast cancer cells and xenografts. *Clin Cancer Res.* (1999) 5:3767s.
 22. Alfares A, Nunez LD, Al-Thihli K, Mitchell J, Melancon S, Anastasio N, et al. Combined malonic and methylmalonic aciduria: exome sequencing reveals mutations in the ACSF3 gene in patients with a non-classic phenotype. *J Med Genet.* (2011) 48:602–5. doi: 10.1136/jmedgenet-2011-100230
 23. Robinson DL, Cafe LM, Mckiernan WA. Heritability of muscle score and genetic and phenotypic relationships with weight, fatness and eye muscle area in beef cattle. *Anim Prod Sci.* (2014) 54:1443–8. doi: 10.1071/AN14347
 24. Bernal Rubio YL, Gualdrón Duarte JL, Bates RO, Ernst CW, Nonneman D, Rohrer GA, et al. Implementing meta-analysis from genome-wide association studies for pork quality traits. *J Anim Sci.* (2015) 93:5607–17. doi: 10.2527/jas.2015-9502
 25. Liu Y, Chen SH, Jin X, Li YM. Analysis of differentially expressed genes and microRNAs in alcoholic liver disease. *Int J Mol Med.* (2013) 31:547–54. doi: 10.3892/ijmm.2013.1243

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 He, Fang, Lu, Liu, Li, Zhao, Li and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Effects of Heat Stress on the Ruminal Epithelial Barrier of Dairy Cows Revealed by Micromorphological Observation and Transcriptomic Analysis

Zitai Guo, Shengtao Gao, Jun Ding, Junhao He, Lu Ma and Dengpan Bu*

State Key Laboratory of Animal Nutrition, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Sayyed Mahmoud Nasrollahi,
University of Tehran, Iran
Amanda K. Lindholm-Perry,
U.S. Meat Animal Research Center,
United States

*Correspondence:

Dengpan Bu
budengpan@caas.cn

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 31 August 2021

Accepted: 07 December 2021

Published: 13 January 2022

Citation:

Guo Z, Gao S, Ding J, He J, Ma L and
Bu D (2022) Effects of Heat Stress on
the Ruminal Epithelial Barrier of Dairy
Cows Revealed by
Micromorphological Observation and
Transcriptomic Analysis.
Front. Genet. 12:768209.
doi: 10.3389/fgene.2021.768209

Heat stress (HS) alters the rumen fermentation of dairy cows thereby affecting the metabolism of rumen papillae and thus the epithelial barrier function. The aim of the present study was to investigate if HS damages the barrier function of ruminal epithelia. Eight multiparous Holstein dairy cows with rumen cannula were randomly equally allocated to two replicates ($n = 4$), with each replicate being subjected to heat stress or thermal neutrality and pair-feeding in four environmental chambers. Micromorphological observation showed HS aggravated the shedding of the corneum and destroyed the physical barrier of the ruminal epithelium to a certain extent. Transcriptomics analysis of the rumen papillae revealed pathways associated with DNA replication and repair and amino acid metabolism were perturbed, the biological processes including sister chromatid segregation, etc. were up-regulated by HS, while the MAPK and NF- κ B cell signaling pathways were downregulated. However, no heat stress-specific change in the expression of tight junction protein or TLR4 signaling was found, suggesting that HS negatively affected the physical barrier of the ruminal epithelium to some extent but did not break the ruminal epithelium. Heat stress invoked mechanisms to maintain the integrity of the rumen epithelial barrier by upregulating the expression of heat shock protein and repairs in rumen papillae. The increase in amino acid metabolism in rumen papillae might affect the nutrient utilization of the whole body. The findings of this study may inform future research to better understand how heat stress affects the physiology and productivity of lactating cows and the development of mitigation strategies.

Keywords: heat stress, dairy cow, ruminal epithelium, milk protein, rumen fermentation

INTRODUCTION

Heat stress (HS) has been a major concern for dairy producers in tropical and subtropical areas, especially in summer, as HS can not only decrease milk yield but also decrease the content of milk protein (Gao et al., 2017; Guo et al., 2018). Recently, perturbation of the inflammatory response during HS was reported to be responsible, at least partially, for the declined milk protein synthesis (Gao et al., 2019). The host inflammatory response was shown to be related to some of the abnormal metabolites in the rumen, especially lipopolysaccharide (LPS) (Mani et al., 2012). In a previous study, we revealed that HS increased the concentration of volatile fatty acid (VFA) in the rumen fluid before

feeding (Gao et al., 2017). The alterations of rumen fermentation in HS cows are similar to those observed in cows that suffer from subacute ruminal acidosis (SARA). Indeed, the decline of rumen pH during SARA increases the lysis of Gram-negative bacteria, resulting in a rapid increase in LPS (Danscher et al., 2015). However, it is difficult to determine the specific changes in rumen fermentation in cows under HS.

It has been shown that HS can increase the intestinal permeability of monogastric animals (Sakurada and Hales, 1998; Hall et al., 2001; Leon and Helwig, 2010). However, it is not known if HS also increases the permeability of the rumen epithelia, which are structured differently than the intestinal epithelia. Unlike the intestinal epithelia of single monolayer cells connected with tight junctions (Shen and Turner, 2006), the ruminal epithelia have a multi-layered structure of stratified squamous epithelium, the granulosum that has tight junctions, spinosum, and basale (Graham and Simmons, 2005). The complex structure of the rumen papillae also plays an important role in defending against harmful substances in rumen fluid, however, the blood diversion from the viscera to the periphery might alter the ruminal epithelial morphology under heat stress (Kregel, 2002; Lambert et al., 2002). Though a recent study showed that mild HS did not induce barrier dysfunction of the rumen papillae in lactating dairy cows probably owing to a defense mechanism and feeding adaptation (Eslamizad et al., 2020). The damages of HS to the barrier function of ruminal epithelia in lactating dairy cows remain elusive. We hypothesized that HS might damage the barrier function of ruminal epithelia and induce tissue inflammation, which could eventually decrease milk protein synthesis. To this end, we evaluated the effect of HS on the micromorphology and gene expression of the rumen epithelia in lactating cows.

MATERIALS AND METHODS

Animals and Study Design

Eight multiparous Holstein dairy cows (238 ± 10 DIM; 618 ± 100 kg of BW; 23 ± 2.8 kg of milk/d) each with a permanent rumen cannula were used in the current study. Due to environmental chamber availability, the study was carried out in two replicates with four different cows in each replicate as reported in a previous study (Sun et al., 2019). When four cows were being used in a replicate experiment, the other cows were kept in a free-stall barn cooled with running fans until they were required for another replicate experiment. The four cows in one replicate were randomly allocated to four individual environmental chambers (Beijing Kooland Technologies Co., Ltd.) that had 12 h light (0600–1800) and 12 h dark (1800–0600) cycles. For the first 7 days, all cows were maintained at thermal neutral conditions (20°C and 55% relative humidity) and fed *ad libitum* for adaptation. The experiment period lasted for 18 days including 9 days of the control phase and 9 days of the trial phase. In the control phase, all cows continued to be in thermal neutral conditions [20°C and 55% of RH as configuration; 0600–1800 h of light]

TABLE 1 | Ingredients and nutrients of experimental diet (DM basis).

Item	Value
Ingredients (% of DM)	
Bean meal	10.42
Cotton meal	5.03
Rapeseed meal	2.18
DDGS ^a	5.45
Feeding corn meal ^b	1.15
Steam-flaked corn	23.98
Limestone	0.91
Salt	0.55
Magnesium Oxide	0.36
Dicalcium Phosphate	0.42
Fat powder	1.15
Sodium bicarbonate	0.97
Supplement ^c	0.67
Corn silage	28.77
Alfalfa hay	17.99
Chemical analysis (% of DM)	
NDF ^d	27.69
ADF ^e	18.57
CP ^f	15.31
Ash	7.88
Organic matter	92.12
Ether extract	2.1
NE _L ^g (Mcal/kg of DM)	1.69

^aDistillers dried grains with solubles.

^bFlour made with corn.

^cContained (per kg of DM) a minimum of 250,000 IU, of vitamin A; 65,000 IU, of vitamin D; 2,100 IU, of vitamin E; 400 mg of Fe; 540 mg of Cu; 2,100 mg of Zn; 560 mg of Mn; 15 mg of Se; 35 mg of I; and 68 mg of Co.

^dNeutral detergent fiber.

^eAcid detergent fiber.

^fCrude protein.

^gNet energy of lactation.

and fed *ad libitum*. While in the trial phase, two of the four cows were exposed to cyclical heat stress conditions (HS, 0600–1,800 h at 36°C, 1,800 to 0600 h at 32°C, and 40% of RH) and fed *ad libitum*, whereas the other two cows were maintained at the same thermal neutral conditions as mentioned above but pair-fed (PFTN). The amount of feed provided to the PFTN cows was calculated based on the average feed intake of the HS cows 1 day earlier as previously described by Wheelock et al. (2010); thus the trial for the PFTN cows started 1 day after the HS cows. All the cows had free access to drink water and were fed a total mixed ration (TMR) formulated to meet the predicted requirements of NRC in energy, protein, minerals, and vitamins (Table 1). The cows were individually fed twice a day (0500 and 1,700 h). The cows were milked twice a day (0500 and 1700 h) and the milk yield was recorded at each milking time. After the previous four cows exited the chamber, the next four cows were randomly allocated to the four chambers to repeat the experiment as described above. In total, HS and PFTN each had four cows ($n = 4$).

Sampling and Measurements

In both the control and HS trial phases, rectal temperature (RT), skin temperature (ST), and respiratory rate (RR) were recorded for each cow four times daily (0100, 0700, 1,300, and

1900 h). To calculate the precise temperature-humidity-index (THI) inside the chamber, ambient temperature (AT) and RH were recorded four times daily (0100, 0700, 1300, 1900 h) automatically with an electronic thermometer; THI was calculated using the equation below as previously described (Buffington et al., 1981).

The weight of orts from each cow was recorded daily before the morning feeding. On d 2, 4, 6, and 8 of the control and HS trial phases, samples of rumen fluid were collected *via* rumen cannula before morning feeding. The pH of the rumen fluid was immediately measured after sampling using an electronic pH meter. Then the rumen fluid samples were filtered through four layers of gauze and divided into two aliquots (10 ml each), with one aliquot being acidified with 0.1 ml of 6 M HCl for ammonia concentration determination, while the other aliquot being preserved following the addition of 1 ml of 25% metaphosphoric acid for analysis of VFA. All samples of rumen fluid were stored at -20°C until analysis. Milk samples were collected daily from morning and afternoon milking (25 ml for each milking) and mixed equally and stored at 4°C until analysis after bronopol tablet (D&F Control System, San Ramon, CA) was added as a preservative.

Before morning feeding on d 9 of the trial phase in each replicate, six samples of rumen papillae were collected using forceps *via* the rumen cannula from each cow. Three papillae samples were gently rinsed in 0.9% NaCl solution as described by Dieho et al. (2016), immediately frozen in liquid nitrogen, and then stored at -80°C until RNA isolation, while the other three papillae samples were fixed in buffered 4% paraformaldehyde for micromorphological observation.

The fixed papillae samples were rinsed and dehydrated in a series of ethanol baths and then deparaffinized in xylene. Each papillae sample was stained with hematoxylin and eosin (H&E) and then observed under a Leica S9 Stereo microscope (Leica Microsystems Inc., Buffalo Grove, United States) as described previously (Nishihara et al., 2019).

Analysis

Feed samples were dried at 65°C for 48 h and ground with a Wiley mill (Arthur H. Thomas Co., Philadelphia, PA) for analysis of ash, DM, CP, NDF, and ADF content. The NDF content was measured using a fiber analyzer (Ankom Technology, A200, Macedon, NY) using the method of Van Soest et al. (1991), but α -amylase and sodium sulfide were used. The Ash, DM, CP, and ADF contents were analyzed according to the AOAC Official Method (Horwitz et al., 2010, 942.05 for ash; 2001.12-2005 for dry matter; 988.05 for CP, and 973.18 for ADF).

Milk samples were analyzed with the method of infrared spectrophotometry using an automatic analyzer of milk composition (MilkoScan Type 78,110, Foss Electric, Hillerød, Denmark). The concentration of VFA in rumen fluid samples was determined using gas chromatography (Agilent 6890A, Agilent Technologies, California, United States) as described by Hu et al. (2005). Ammonia-N in the rumen fluid samples was measured as described by Broderick and Kang (Broderick and Kang, 1980).

TABLE 2 | The vital signs, lactation performance and rumen fermentation of dairy cows.

Items	PFTN ^a	HS	SEM	p-value	
				Treatment	Period
THI ^b	68.92	83.11	0.7996	0.0002	0.0145
RR ^c (counts/min)	27.42	72.62	3.7667	0.0006	0.4904
RT ^d , $^{\circ}\text{C}$	38.49	39.82	0.4505	0.0484	0.3095
ST ^e , $^{\circ}\text{C}$	32.30	36.89	0.2374	0.0002	0.3138
DMI ^f , kg/d	10.25	10.01	1.0783	0.8293	0.1206
Milk yield, kg/d	17.94	11.80	2.2121	0.0480	0.2112
Protein, %	3.64	3.26	0.1387	0.0436	0.6024
Fat, %	4.88	5.41	0.9828	0.6295	0.8379
Lactose, %	4.74	4.92	0.5317	0.7408	0.7902
SCS ^g	7.31	8.40	1.8059	0.5783	0.0199
pH	6.71	6.34	—	0.0335	0.5840
LPS ^h , EU/mL $\times 10^5$	0.1817	0.1861	0.0285	0.8780	0.5218
NH ₃ -N, mg/dL	17.04	19.07	2.2663	0.4707	0.2407
Total VFA ⁱ	58.01	76.34	2.3998	0.0001	0.9284
Acetate, mmol/L	32.65	45.55	4.2921	0.0189	0.7827
Propionate, mmol/L	14.25	20.99	1.9927	0.0244	0.9488
Isobutyrate, mmol/L	0.68	0.58	0.0579	0.1875	0.6024
Butyrate, mmol/L	7.96	9.58	0.8364	0.0671	0.9911
Isovalerate, mmol/L	1.29	1.15	0.1585	0.4265	0.9848
Valerate, mmol/L	0.65	1.05	0.0676	0.0107	0.6748

^aPair-feeding thermal neutral.

^bTemperature-humidity index.

^cRespiration rate.

^dRectal temperature.

^eSkin temperature.

^fDry matter intake.

^gSomatic cell score.

^hLipopolysaccharide.

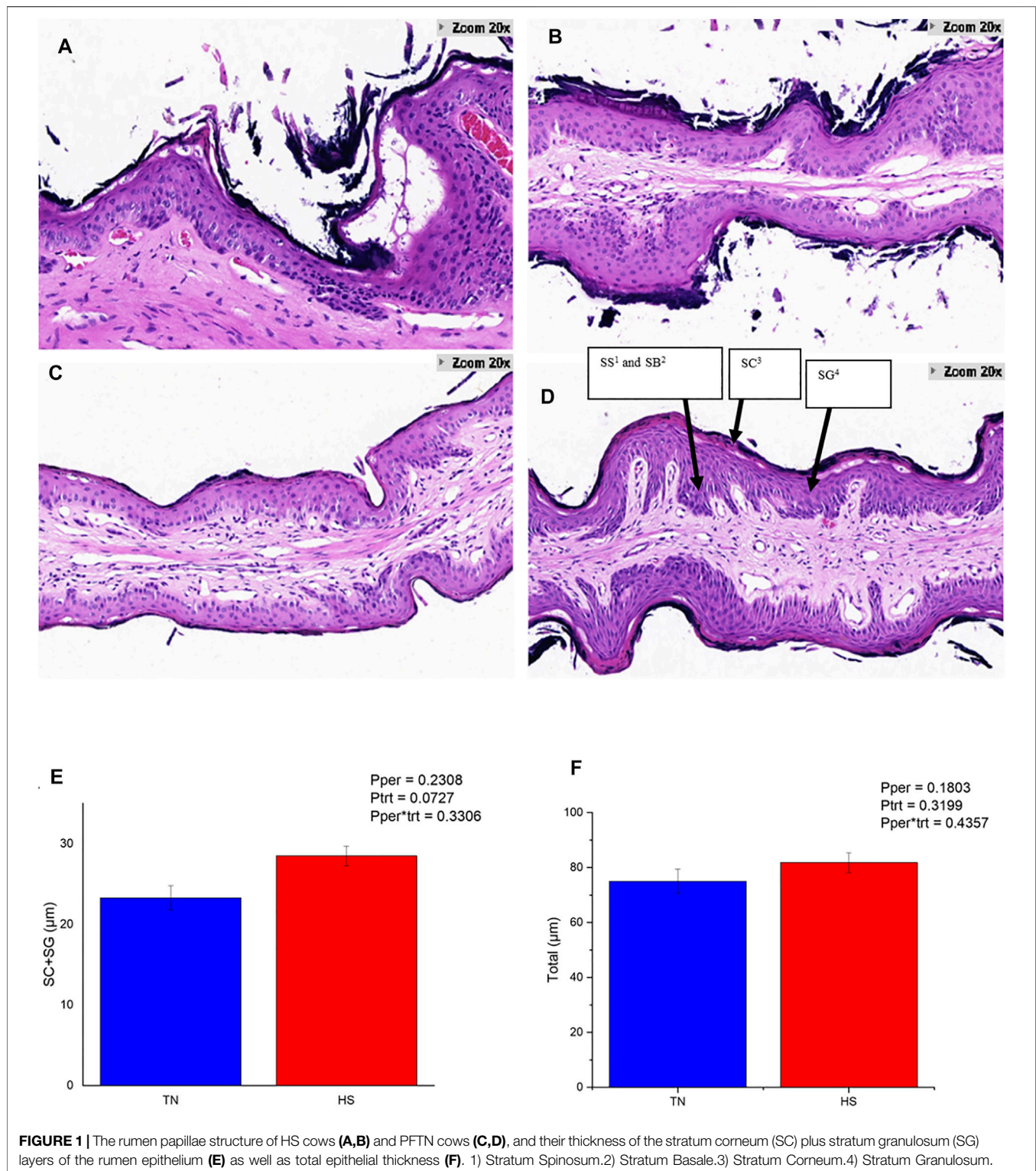
ⁱVolatile fatty acid.

Specimens of rumen papillae were microscopically examined and analyzed using a panoramic scanner (3DHISTECH Ltd., H-1141 Budapest, Hungary), meanwhile the results were assembled and viewed. For each specimen, four clear areas were selected and used to measure the total thickness of ruminal epithelia and the thickness of the corneum and granulosa. To acquire the distance, a segment perpendicular to the screen section was made and measured *via* the ruler tool in the application.

The effect of the HS on the gene expression in papillae was evaluated using transcriptomics. The RNA extraction and sequencing, analysis of differently expressed genes (DEGs) and GO and KEGG enrichment analysis were conducted following our former research (Gao et al., 2021), during which the Dynamic Impact Approach (DIA), Database for Annotation, Visualization, and Integrated Discovery (DAVID) were used.

Statistical Analysis

Data of lactation performance, rumen fermentation, parameters of rumen papillae, and vital signs were statistically analyzed using SAS v. 9.4 (SAS Institute, Cary, NC) with all data tested for normality. All data were analyzed using PROC MIXED. Significance was declared at $p \leq 0.05$, and the tendency was declared at $0.05 < p \leq 0.10$.



Analysis of differentially expressed genes (DEGs) was conducted with the DESeq2 package (1.26.0) in R (3.6.1). The mapped read count tables of individual samples and genes were used as the

standard workflow instructed. The false discovery rate (FDR) obtained by the method of Benjamin and Hochberg was used to correct the *p*-value. Genes with *p* < 0.05 were regarded as DEGs.

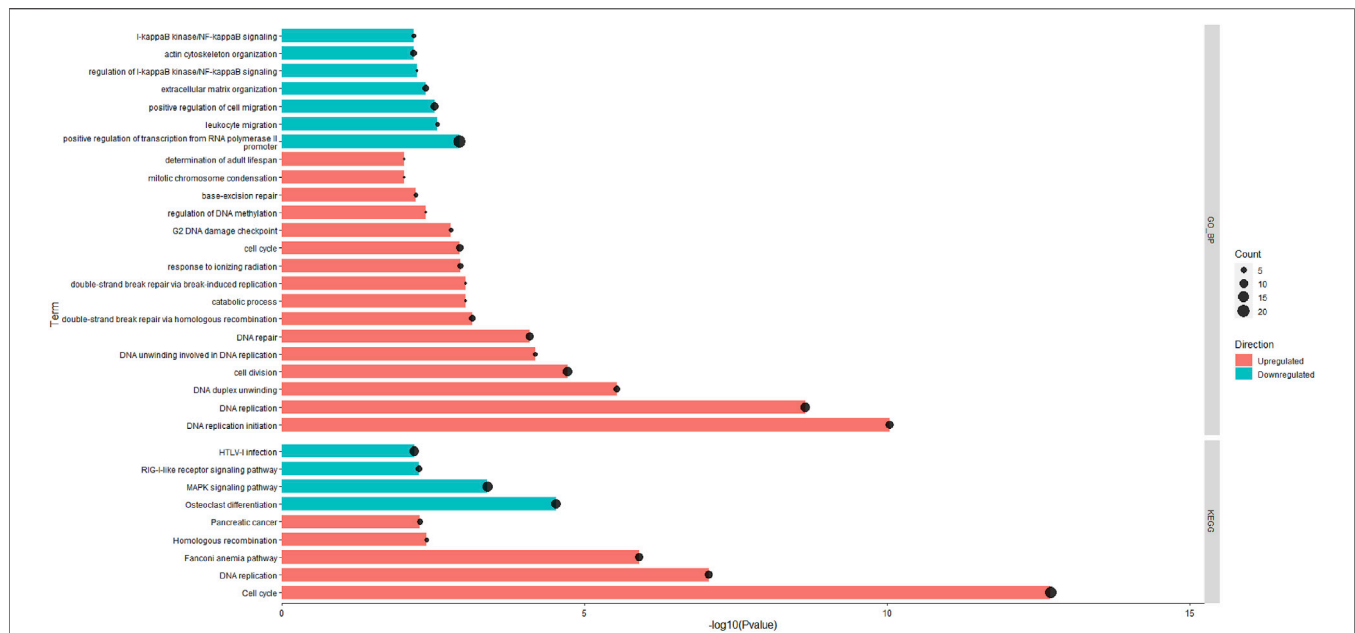


FIGURE 2 | The significantly enriched pathways in rumen epithelial tissues of the HS and the PFTN cows as determined using Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis against the GO_BP and KEGG databases.

RESULTS

The Effect of HS on Vital Signs, Lactation Performance, and Rumen Fermentation of Dairy Cows

The effects of HS on vital signs, lactation performance, and rumen fermentation are reported in **Table 2**. In brief, HS increased the RR, RT, and ST (2.65-fold, 1.33°C and 4.59°C, respectively; $p < 0.01$) compared with the PFTN conditions, and reduced milk yield (by 34.22%, $p < 0.05$) and the content of milk protein (by 10.44%, $p < 0.05$) compared with the PFTN conditions. In addition, HS increased the concentration of total VFA (by 31.60%, $p < 0.01$), acetate (by 39.51%, $p < 0.05$), propionate (by 47.29%, $p < 0.05$), and valerate (by 61.54%, $p < 0.05$) in rumen fluid, while decreased the rumen liquor pH compared with PFTN. The HS cows tended to have increased concentration of butyrate (by 20.35%, $0.05 < p < 0.1$) but not isobutyrate or isovalerate.

The Effect of HS on Rumen Papillae

The HS cows showed obvious damage and slough off of the corneum (**Figures 1A,B**), while the PFTN cows had intact corneum (**Figures 1C,D**). Heat stress enlarged the intercellular space of the granulosum and spinosum and induced obvious separations of layers inside the rumen papillae (**Figure 1B**). Furthermore, HS tended to increase the thickness of corneum and granulosum (23.20 vs. 28.39 μ m; $0.05 < p < 0.1$; **Figure 1E**) but did not affect the thickness of the whole rumen papillae (**Figure 1F**).

The Effect of HS on the Gene Expression in Rumen Papillae

A total of 15,654 unique expressed genes were detected in the tissue of rumen papillae. Between the HS and PFTN cows, 501 genes were DEGs, including 238 up-regulated and 263 down-regulated ($\text{Padj} \leq 0.05$). The methods of DIA and DAVID were used in the functional analysis of the DEGs. The whole DIA output, the results of perturbations on the main categories and subcategories of the KEGG in rumen papillae between HS and PFTN cows, and the top 20 most impacted pathways, as uncovered by the DIA, in the rumen papillae of HS cows compared with PFTN cows are available in supplement File S2. All the results of DAVID analysis using the KEGG and GO Biological Process (GO_BP) are shown in **Figure 2**. In the HS cows, the GO_BP analysis revealed 7 and 16 different terms that were downregulated DEGs and upregulated DEGs ($p \leq 0.05$), respectively.

DISCUSSION

Heat stress significantly decreases feed intake and milk performance (Beede and Collier, 1986; West, 2003; Ranjitkar et al., 2020). A THI of 68–69 was considered to cause some HS when evaluated with respect to rising respiratory rates and rectal temperature (Brügemann et al., 2012; Collier et al., 2012; Umar et al., 2021). However, in this study, the HS cows suffered a more severe reduction in milk yield compared with PFTN cows, which is consistent with our previous study (Gao et al., 2017). The results of vital signs, milk yield, and THI indicated that the HS and pair-fed model was

successfully performed. Evidently, HS probably directly decreased milk synthesis through the regulation of the growth hormone axis as proposed previously (Rhoads et al., 2010).

As the major products of rumen fermentation, VFA are the main energy source for ruminants (Sutton et al., 1988; Owens and Basalan, 2016). However, slight perturbations in rumen fermentation can change the concentration of individual VFAs. Heat stress can increase the frequency of feed ingestion and drinking (Eslamizad et al., 2020; Herbut et al., 2021). It is believed that such adaptations were the results of self-regulation in response to HS (West, 2003). Robles et al. (2007) found that increasing the frequency of daily feeding (2 or 4 times) increased the concentration of total VFA compared with cows fed only once daily. In the current study, the amount of ingested fermentable organic matter was similar between the HS and the PFTN cows. There was a small chance that feeding changes would affect VFA concentration by reducing or increasing DMI. Faster dilution in the rumen caused by increased water intake was proved to promote microbial growth in an early study (Nocek and Braund, 1985). The rumen microbes can provide moderate concentrations of fermentation end product source of the ingesta provided by the ruminant (Calamari et al., 2013), which were ultimately metabolized to VFAs and absorbed through the rumen wall (Hyder et al., 2017). Thus, the higher rumen liquid turnover rate in the HS cows caused by feeding changes might be attributable to their increased VFA concentration in the rumen.

Elevated concentrations of propionate and butyrate in the rumen can stimulate the development of the rumen papillae, hence enhancing VFA absorption across the ruminal epithelium (Kristensen and Harmon, 2004; Storm et al., 2011). Butyrate is likely to be oxidized into β -hydroxybutyrate after being absorbed in the ruminal epithelium cells, directly providing energy to the rumen papillae (Ślusarczyk et al., 2010). Steele et al. (2012) showed that butyrate promotes the length of rumen papillae by perfusion trials. Moreover, a low concentration of butyrate was shown to inhibit cell apoptosis (Xu et al., 2018). Therefore, elevated rumen concentrations of propionate and butyrate corroborate the increased thickness of the corneum and granulosa observed in the HS cows.

The corneum of rumen papillae in the HS cows was shed and the inside layers appeared separated. The corneum is in direct contact with rumen content and is colonized with rumen microorganisms and thus the epithelial tissues are updated regularly (Lavker and Sun, 1983). The update frequency is related to diet ingredients, with high-concentrate diets significantly reducing the renewal frequency. Tajima et al. (2001) indicated that long-term feeding of high-concentrate diets decreased the acetate:propionate ratio in the rumen fluid whereas it increased the concentration of total VFA, suggesting that ruminal corneum exfoliation in HS cows may also be attributed to increased total rumen VFA concentration. In addition, the transport and absorption of VFA by the ruminal epithelium depend on the integrity of the corneum and the degree of keratinization (Galfi et al., 1991; Del Bianco Benedetti et al., 2018). Yohe et al. (2019) reported that the exfoliation of the corneum and the hyperkeratosis of the epithelium reduce the ability of the ruminal epithelium to transport and absorb VFA. Therefore, the increased VFA concentration in the HS cows might be a major reason for the exfoliation of the corneum.

The HS treatment affected the expression of some genes in the rumen papillae. Among the 20 most impacted pathways (as detected with the DIA analysis), HS activated one pathway related to membrane transport (ABC Transporter). The ABC transporters transport various substrates including amino acids, peptides, and cellular metabolites across the cell membrane (Schneider and Hunke, 1998; Dean et al., 2001). Except for the upregulation of ABC transporters, HS upregulated the expression of HSPA5 and DNAJB9, which encode heat shock proteins (Hsp) 70 and 40, respectively. Hsp has functions in maintaining physiological and stabilizing the structure of cells (Sharp et al., 1999), and among all the known Hsp, Hsp70 has the most prominent functions in most animals under stress (Kiang and Tsokos, 1998; Bhat et al., 2016). Considerable increases in the expression of Hsp are usually directly associated with stress (Feder and Hofmann, 1999). Herein, the upregulation of Hsp observed in the rumen papillae of the HS cows is beneficial to the repair and maintenance of the cells. As a result, the absorption and utilization of amino acids in the rumen papillae increased as indicated by the upregulation of six amino acid metabolism pathways by HS. The up-regulation of the amino acids metabolism indicates that amino acid utilization in the ruminal epithelium was enhanced by Hsp in the HS cows, suggesting the amino acid concentrations in rumen fluid were increased. The source of amino acids in rumen fluid was mainly from diet protein as well as the degradation of rumen microorganisms. Some rumen microorganisms could directly synthesize some amino acids from VFA and ammonia. The increased rumen concentration of VFA could provide the carbon source for amino acid synthesis, meanwhile causing the death of gram-negative bacteria in rumen. Furthermore, the concentrations of the amino acids whose metabolism was up-regulated by HS were also increased in the rumen by SARA, and SARA can also exacerbate the breakdown of bacteria and increase the rumen concentration of amino acids. Thus, the activation of amino acids metabolism in the papillae in the HS cows might be similar to that observed in cows suffering from SARA, which is mostly related to the changing of rumen fermentation.

Besides up-regulating the metabolism of amino acids, HS also influenced the replication of DNA and the repair of ruminal epithelial cells. Four of the 20 most impacted pathways are relevant to the replication and repair of genetic information processing, among which homologous recombination, Fanconi anemia pathway is related to the repair process (San Filippo et al., 2008; Krokan and Bjørås, 2013; Rodríguez and D'Andrea, 2017). While homologous repair is the most important mechanism of repair of DNA double-strand breaks (Baumann and West, 1998). HS leads to excessive production of reactive oxygen species (ROS) in dairy cows, inducing oxidative stress (OS) in their body (Guo et al., 2021). The mitochondrial dysfunction following OS can induce pro-apoptotic factors in the mitochondrial inner membrane and activate endogenous apoptosis and ultimately lead to tissue apoptosis (Kannan and Jain, 2000; Antonsson, 2001). Propionate and butyrate were believed to play an important role in cell growth, but several studies have contradictory opinions on the effects of these two VFAs

(Baldwin, 1999; Shen et al., 2005). Liu et al. (2014) indicated that butyrate needed to be used in combination with insulin-like growth factor-1 (IGF-1) to significantly promote DNA replication, and otherwise it can inhibit DNA replication. In the present study, we found up-regulation of the gene encoding insulin-like growth factor binding protein 5 in the HS cows, which might enhance the biological effects of IGF-1. Therefore, the OS induced by HS could aggravate the damage to the epithelial cells in rumen papillae. Such damage might activate the repairing of intracellular DNA. To protect the epithelial cells, Hsp synthesis needs to increase, and ruminal epithelial cells need to upregulate their amino acid metabolism. Because nutrients are partitioned between the liver and mammary gland in cows under HS (Bu et al., 2017), the increase in amino acid metabolism in rumen papillae would decrease the supply of precursors for milk production by the mammary glands. However, further correlation analysis of nutrient utilization among the rumen epithelia, mammary gland, and other tissues is still needed to support this premise.

The results of our DAVID analysis verified the activation of replication and repair in rumen papillae. All the up-regulated DEGs are involved in DNA replication and cell division, while the down-regulated DEGs are involved in the MAPK signaling pathway and the NF- κ B cell signaling pathway in rumen papillae. Mitogen-activated protein kinases (MAPKs) are a type of protein kinases that are widely present in animal cells. They are activated in cells in which DNA damage or oxidative damage occurs, inducing a pro-apoptotic effect (Seeger and Krebs, 1995; Zumsande and Gross, 2010). The TLR4/NF- κ B pathway is closely related to the anti-inflammatory immunity of the body (Zusso et al., 2019). Moreover, one study has found that TLR4 in epithelial cells could recognize LPS and activate related pathways (Bäckhed et al., 2002). Taken together, the rumen papillae in the HS cows did not exhibit leukocyte migration or up-regulation of the NF- κ B pathway, suggesting that TLR4 related upstream receptors might have not been activated. Furthermore, the expressions of tight junction protein including claudin, occludin, and ZO-1 showed no difference between the HS and PFTN cows. The tight junctions among ruminal epithelial cells are the most important connection between cells and serve as the barrier preventing harmful substances from the rumen (Zhang et al., 2019; Ma et al., 2021). They can also assist in the transport of ions and nutrients. The results of the present study suggest that the expression of Hsp may participate in the protection of tight junctions.

CONCLUSION

Heat stress exhibited direct impacts on rumen fermentation and metabolism of rumen papillae. Heat stress promoted the proliferation of the rumen papillae but aggravated the shedding of the corneum and may negatively affect the physical barrier of the ruminal epithelium to a certain extent. The increase in VFA concentration induced by heat stress might stimulate the development of rumen papillae. However, heat stress did not break the intact barrier function, since neither a change in tight junctions nor perturbation to inflammatory

response was observed in ruminal papillae, and heat stress did not alter the expression of TLR4-related upstream receptors. Heat stress up-regulated the expression of heat shock protein and activated the repair of damaged epithelial cells in rumen papillae. These mechanisms contribute to the maintenance of the integrity of rumen tissue. The up-regulated metabolism of amino acids along with Hsp synthesis may affect the supply of the precursors for milk protein synthesis, but correlation analysis of the utilization of amino acids in organs and the whole body is needed in future studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: <https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA778529>. The supplementary material are available online at <https://doi.org/10.6084/m9.figshare.17041181>, File S1: Dataset of differently expressed genes; File S2: The output of DIA analysis; File S3: The output of DAVID analysis.

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care and Use Committee of Institute of Animal Science, Chinese Academy of Agricultural Sciences (No. IAS20180115). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, ZG, SG, and DB; Software operation, SG; Statistical analysis, ZG and SG; Drafting, review, and editing of the manuscripts, ZG, LM, JD, JH, and DB; All authors have read and agreed to the published version of the manuscript.

FUNDING

The present study was financially supported by the National Natural Science Foundation of China (31872383), the Key Research and Development Program of the Ningxia Hui Autonomous Region (2021BEF02018), the Scientific Research Project for Major Achievements of The Agricultural Science and Technology Innovation Program (ASTIP) (ASTIP-IAS07, CAAS-XTCX2016011-01) and Beijing Dairy Industry Innovation Team (BAIC06-2021).

ACKNOWLEDGMENTS

We especially thank our former and current students who have contributed to this study and all the staff of the State Key Laboratory of Animal Nutrition (Beijing, China) for the use of environmental chambers and assistance with sample analyses.

REFERENCES

- Antonsson, B. (2001). Bax and Other Pro-apoptotic Bcl-2 Family "Killer-Proteins" and Their Victim the Mitochondrion. *Cell Tissue Res* 306 (3), 347–361. doi:10.1007/s00441-001-0472-0
- Bäckhed, F., Meijer, L., Normark, S., and Richter-Dahlfors, A. (2002). TLR4-dependent Recognition of Lipopolysaccharide by Epithelial Cells Requires sCD14. *Cell. Microbiol.* 4 (8), 493–501. doi:10.1046/j.1462-5822.2002.00208.x
- Baldwin, R. L. (1999). The Proliferative Actions of Insulin, Insulin-like Growth Factor-I, Epidermal Growth Factor, Butyrate and Propionate on Ruminal Epithelial Cells *In Vitro*. *Small Ruminant Res.* 32 (3), 261–268. doi:10.1016/s0921-4488(98)00188-6
- Baumann, P., and West, S. C. (1998). Role of the Human RAD51 Protein in Homologous Recombination and Double-Stranded-Break Repair. *Trends Biochem. Sci.* 23 (7), 247–251. doi:10.1016/s0968-0004(98)01232-8
- Beede, D. K., and Collier, R. J. (1986). Potential Nutritional Strategies for Intensively Managed Cattle during thermal Stress. *J. Anim. Sci.* 62 (2), 543–554. doi:10.2527/jas1986.622543x
- Bhat, S., Kumar, P., Kashyap, N., Deshmukh, B., Dige, M. S., Bhushan, B., et al. (2016). Effect of Heat Shock Protein 70 Polymorphism on Thermotolerance in Tharparkar Cattle. *Vet. World* 9 (2), 113–117. doi:10.14202/vetworld.2016.113-117
- Broderick, G. A., and Kang, J. H. (1980). Automated Simultaneous Determination of Ammonia and Total Amino Acids in Ruminal Fluid and *In Vitro* media. *J. Dairy Sci.* 63 (1), 64–75. doi:10.3168/jds.s0022-0302(80)82888-8
- Brügemann, K., Gernand, E., von Borstel, U. K., and König, S. (2012). Defining and Evaluating Heat Stress Thresholds in Different Dairy Cow Production Systems. *Arch. Anim. Breed.* 55 (1), 13–24. doi:10.5194/aab-55-13-2012
- Bu, D., Bionaz, M., Wang, M., Nan, X., Ma, L., and Wang, J. (2017). Transcriptome Difference and Potential Crosstalk between Liver and Mammary Tissue in Mid-lactation Primiparous Dairy Cows. *PloS one* 12 (3), e0173082. doi:10.1371/journal.pone.0173082
- Buffington, D., Collazo-Arocho, A., Canton, G., Pitt, D., Thatcher, W., and Collier, R. (1981). Black globe-humidity index (BGHI) as comfort Equation for Dairy Cows. *Trans. ASAE* 24 (3), 711–714. doi:10.13031/2013.34325
- Calamari, L., Petrera, F., Stefanini, L., and Abeni, F. (2013). Effects of Different Feeding Time and Frequency on Metabolic Conditions and Milk Production in Heat-Stressed Dairy Cows. *Int. J. Biometeorol.* 57 (5), 785–796. doi:10.1007/s00484-012-0607-x
- Collier, R. J., Hall, L. W., Rungruang, S., and Zimbleman, R. B. (2012). Quantifying Heat Stress and its Impact on Metabolism and Performance. *Department of Animal Sciences University of Arizona*. 68.
- Dansch, A. M., Li, S., Andersen, P. H., Khafipour, E., Kristensen, N. B., and Plaizier, J. C. (2015). Indicators of Induced Subacute Ruminal Acidosis (SARA) in Danish Holstein Cows. *Acta Vet. Scand.* 57 (1), 39–44. doi:10.1186/s13028-015-0128-9
- Dean, M., Hamon, Y., and Chimini, G. (2001). The Human ATP-Binding Cassette (ABC) Transporter Superfamily. *J. Lipid Res.* 42 (7), 1007–1017. doi:10.1016/s0022-2275(20)31588-1
- Del Bianco Benedetti, P., Silva, B. d. C., Pacheco, M. V. C., Serão, N. V. L., Carvalho Filho, I., Lopes, M. M., et al. (2018). Effects of Grain Processing Methods on the Expression of Genes Involved in Volatile Fatty Acid Transport and pH Regulation, and Keratinization in Ruminal Epithelium of Beef Cattle. *PloS one* 13 (6), e0198963. doi:10.1371/journal.pone.0198963
- Dieho, K., Bannink, A., Geurts, I. A. L., Schoneville, J. T., Gort, G., and Dijkstra, J. (2016). Morphological Adaptation of Ruminal Papillae during the Dry Period and Early Lactation as Affected by Rate of Increase of Concentrate Allowance. *J. Dairy Sci.* 99 (3), 2339–2352. doi:10.3168/jds.2015-9837
- Eslamizad, M., Albrecht, D., and Kuhla, B. (2020). The Effect of Chronic, Mild Heat Stress on Metabolic Changes of Nutrition and Adaptations in Ruminal Papillae of Lactating Dairy Cows. *J. Dairy Sci.* 103 (9), 8601–8614. doi:10.3168/jds.2020-18417
- Feder, M. E., and Hofmann, G. E. (1999). Heat-shock Proteins, Molecular Chaperones, and the Stress Response: Evolutionary and Ecological Physiology. *Annu. Rev. Physiol.* 61 (1), 243–282. doi:10.1146/annurev.physiol.61.1.243
- Gálf, P., Neogrady, S., and Sakata, T. (1991). "Effects of Volatile Fatty Acids on the Epithelial Cell Proliferation of the Digestive Tract and its Hormonal Mediation," in *Physiological Aspects of Digestion and Metabolism in Ruminants* (Amsterdam: Elsevier), 49–59. doi:10.1016/b978-0-12-702290-1.50010-2
- Gao, S. T., Guo, J., Quan, S. Y., Nan, X. M., Fernandez, M. V. S., Baumgard, L. H., et al. (2017). The Effects of Heat Stress on Protein Metabolism in Lactating Holstein Cows. *J. Dairy Sci.* 100 (6), 5040–5049. doi:10.3168/jds.2016-11913
- Gao, S. T., Ma, L., Zhang, Y., Wang, J., Loo, J., and Bu, D. J. B. G. (2021). Hepatic Transcriptome Perturbations in Dairy Cows Fed Different Forage Resources. *BMC Genomics* 22 (1), 1–13. doi:10.1186/s12864-020-07332-0
- Gao, S. T., Ma, L., Zhou, Z., Zhou, Z. K., Baumgard, L. H., Jiang, D., et al. (2019). Heat Stress Negatively Affects the Transcriptome Related to Overall Metabolism and Milk Protein Synthesis in Mammary Tissue of Midlactating Dairy Cows. *Physiol. Genomics* 51 (8), 400–409. doi:10.1152/physiolgenomics.00039.2019
- Graham, C., and Simmons, N. L. (2005). Functional Organization of the Bovine Ruminal Epithelium. *Am. J. Physiology-Regulatory, Integr. Comp. Physiol.* 288 (1), R173–R181. doi:10.1152/ajpregu.00425.2004
- Guo, J., Gao, S., Quan, S., Zhang, Y., Bu, D., and Wang, J. (2018). Blood Amino Acids Profile Responding to Heat Stress in Dairy Cows. *Asian-australas. J. Anim. Sci.* 31 (1), 47–53. doi:10.5713/ajas.16.0428
- Guo, Z., Gao, S., Ouyang, J., Ma, L., and Bu, D. (2021). Impacts of Heat Stress-Induced Oxidative Stress on the Milk Protein Biosynthesis of Dairy Cows. *Animals* 11 (3), 726. doi:10.3390/ani11030726
- Hall, D. M., Buettner, G. R., Oberley, L. W., Xu, L., Matthes, R. D., and Gisolfi, C. V. (2001). Mechanisms of Circulatory and Intestinal Barrier Dysfunction during Whole Body Hyperthermia. *Am. J. Physiology-Heart Circulatory Physiol.* 280 (2), H509–H521. doi:10.1152/ajpheart.2001.280.2.h509
- Herbut, P., Hoffmann, G., Angrecka, S., Godyń, D., Vieira, F. M. C., Adamczyk, K., et al. (2021). The Effects of Heat Stress on the Behaviour of Dairy Cows—A Review. *Annu. Rev. Physiol.* 21 (2), 385–402. doi:10.2478/aoas-2020-0116
- Horwitz, W. (2010). *Official Methods of Analysis of AOAC International. Volume I, Agricultural Chemicals, Contaminants, Drugs*. Editor W. Horwitz (Maryland: AOAC International).
- Hu, W. L., Liu, J. X., Ye, J. A., Wu, Y. M., and Guo, Y. Q. (2005). Effect of Tea Saponin on Ruminal Fermentation *In Vitro*. *Anim. Feed Sci. Technol.* 120 (3–4), 333–339. doi:10.1016/j.anifeeds.2005.02.029
- Hyder, I., Ravi Kanth Reddy, P., Raju, J., Manjari, P., Srinivasa Prasad, C., Aswani Kumar, K., et al. (2017). "Alteration in Ruminal Functions and Diet Digestibility during Heat Stress in Sheep," in *Sheep Production Adapting to Climate Change* (Berlin: Springer), 235–265. doi:10.1007/978-981-10-4714-5_11
- Kannan, K., and Jain, S. K. (2000). Oxidative Stress and Apoptosis. *Pathophysiology* 7 (3), 153–163. doi:10.1016/s0928-4680(00)00053-5
- Kiang, J., and Tsokos, G. C. (1998). Heat Shock Protein 70 kDa Molecular Biology, Biochemistry, and Physiology. *Pharmacol. Ther.* 80 (2), 183–201. doi:10.1016/s0163-7258(98)00028-x
- Kregel, K. C. (2002). Invited Review: Heat Shock Proteins: Modifying Factors in Physiological Stress Responses and Acquired Thermotolerance. *J. Appl. Physiol.* 92 (5), 2177–2186. doi:10.1152/japplphysiol.01267.2001
- Kristensen, N. B., and Harmon, D. L. (2004). Effect of Increasing Ruminal Butyrate Absorption on Splanchnic Metabolism of Volatile Fatty Acids Absorbed from the Washed Reticulorum of Steers. *J. Anim. Sci.* 82 (12), 3549–3559. doi:10.2527/2004.82123549x
- Krokan, H. E., and Björås, M. (2013). Base Excision Repair. *Cold Spring Harbor Perspect. Biol.* 5 (4), a012583. doi:10.1101/cshperspect.a012583
- Lambert, G. P., Gisolfi, C. V., Berg, D. J., Moseley, P. L., Oberley, L. W., and Kregel, K. C. (2002). Selected Contribution: Hyperthermia-Induced Intestinal Permeability and the Role of Oxidative and Nitrosative Stress. *J. Appl. Physiol.* 92 (4), 1750–1761. doi:10.1152/japplphysiol.00787.2001
- Lavker, R. M., and Sun, T.-T. (1983). Epidermal Stem Cells. *J. Invest. Dermatol.* 81 (1), S121–S127. doi:10.1111/1523-1747.ep12540880
- Leon, L. R., and Helwig, B. G. (2010). Heat Stroke: Role of the Systemic Inflammatory Response. *J. Appl. Physiol.* 109 (6), 1980–1988. doi:10.1152/japplphysiol.00301.2010
- Liu, D.-c., Zhou, X.-l., Liu, G.-j., Gao, M., and Hu, H.-l. (2014). Promotion and Inhibition of Ruminal Epithelium Growth by Butyric Acid and Insulin-like

- Growth Factor-1 (IGF-1) in Dairy Goats. *J. Integr. Agric.* 13 (9), 2005–2009. doi:10.1016/s2095-3119(13)60603-6
- Ma, Y., Zhang, Y., Zhang, H., and Wang, H. (2021). Thiamine Alleviates High-Concentrate-Diet-Induced Oxidative Stress, Apoptosis, and Protects the Rumen Epithelial Barrier Function in Goats. *Front. Vet. Sci.* 8, 663698. doi:10.3389/fvets.2021.663698
- Mani, V., Weber, T. E., Baumgard, L. H., and Gabler, N. K. (2012). GROWTH and DEVELOPMENT SYMPOSIUM: Endotoxin, Inflammation, and Intestinal Function in Livestock1,2. *J. Anim. Sci.* 90 (5), 1452–1465. doi:10.2527/jas.2011-4627
- Nishihara, K., Suzuki, Y., Kim, D., and Roh, S. (2019). Growth of Rumen Papillae in Weaned Calves Is Associated with Lower Expression of Insulin-like Growth Factor-binding Proteins 2, 3, and 6. *Anim. Sci. J.* 90 (9), 1287–1292. doi:10.1111/asj.13270
- Nocek, J. E., and Braund, D. G. (1985). Effect of Feeding Frequency on Diurnal Dry Matter and Water Consumption, Liquid Dilution Rate, and Milk Yield in First Lactation. *J. Dairy Sci.* 68 (9), 2238–2247. doi:10.3168/jds.s0022-0302(85)81096-1
- Owens, F. N., and Basalan, M. (2016). “Ruminal Fermentation,” in *Rumenology* (Berlin: Springer), 63–102. doi:10.1007/978-3-319-30533-2_3
- Ranjitkar, S., Bu, D., Van Wijk, M., Ma, Y., Ma, L., Zhao, L., et al. (2020). Will Heat Stress Take its Toll on Milk Production in China? *Climatic Change* 161 (4), 637–652. doi:10.1007/s10584-020-02688-4
- Rhoads, M. L., Kim, J. W., Collier, R. J., Crooker, B. A., Boisclair, Y. R., Baumgard, L. H., et al. (2010). Effects of Heat Stress and Nutrition on Lactating Holstein Cows: II. Aspects of Hepatic Growth Hormone Responsiveness. *J. Dairy Sci.* 93 (1), 170–179. doi:10.3168/jds.2009-2469
- Robles, V., Gonza'lez, L. A., Ferret, A., Manteca, X., and Calsamiglia, S. (2007). Effects of Feeding Frequency on Intake, Ruminal Fermentation, and Feeding Behavior in Heifers Fed High-Concentrate Diets1. *J. Anim. Sci.* 85 (10), 2538–2547. doi:10.2527/jas.2006-739
- Rodríguez, A., and D'Andrea, A. (2017). Fanconi Anemia Pathway. *Curr. Biol.* 27 (18), R986–R988. doi:10.1016/j.cub.2017.07.043
- Sakurada, S., and Hales, J. R. S. (1998). A Role for Gastrointestinal Endotoxins in Enhancement of Heat Tolerance by Physical Fitness. *J. Appl. Physiol.* 84 (1), 207–214. doi:10.1152/jappl.1998.84.1.207
- San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. *Annu. Rev. Biochem.* 77, 229–257. doi:10.1146/annurev.biochem.77.061306.125255
- Schneider, E., and Hunke, S. (1998). ATP-binding-cassette (ABC) Transport Systems: Functional and Structural Aspects of the ATP-Hydrolyzing Subunits/domains. *FEMS Microbiol. Rev.* 22 (1), 1–20. doi:10.1111/j.1574-6976.1998.tb00358.x
- Seger, R., and Krebs, E. G. (1995). The MAPK Signaling cascade. *FASEB j.* 9 (9), 726–735. doi:10.1096/fasebj.9.9.7601337
- Sharp, F. R., Massa, S. M., and Swanson, R. A. (1999). Heat-shock Protein protection. *Trends Neurosciences* 22 (3), 97–99. doi:10.1016/s0166-2236(98)01392-7
- Shen, L., and Turner, J. R. (2006). Role of Epithelial Cells in Initiation and Propagation of Intestinal Inflammation. Eliminating the Static: Tight Junction Dynamics Exposed. *Am. J. Physiology-Gastrointestinal Liver Physiol.* 290 (4), G577–G582. doi:10.1152/ajpgi.00439.2005
- Shen, Z., Kuhla, S., Zitnan, R., Seyfert, H.-M., Schneider, F., Hagemeister, H., et al. (2005). Intraruminal Infusion of N-Butyric Acid Induces an Increase of Ruminal Papillae Size Independent of IGF-1 System in Castrated Bulls. *Arch. Anim. Nutr.* 59 (4), 213–225. doi:10.1080/17450390500216894
- Ślusarczyk, K., Strzetelski, J., and Furgał-Dierżuk, I. (2010). The Effect of Sodium Butyrate on Calf Growth and Serum Level of β -hydroxybutyric Acid. *J. Anim. Feed Sci.* 19 (3), 348–357. doi:10.22358/jafs/66298/2010
- Steele, M. A., Dionissopoulos, L., AlZahal, O., Doelman, J., and McBride, B. W. (2012). Rumen Epithelial Adaptation to Ruminal Acidosis in Lactating Cattle Involves the Coordinated Expression of Insulin-like Growth Factor-Binding Proteins and a Cholesterolgenic Enzyme. *J. Dairy Sci.* 95 (1), 318–327. doi:10.3168/jds.2011-4465
- Storm, A. C., Hanigan, M. D., and Kristensen, N. B. (2011). Effects of Ruminal Ammonia and Butyrate Concentrations on Reticulorumen Epithelial Blood Flow and Volatile Fatty Acid Absorption Kinetics under Washed Reticulorumen Conditions in Lactating Dairy Cows. *J. Dairy Sci.* 94 (8), 3980–3994. doi:10.3168/jds.2010-4091
- Sun, L. L., Gao, S. T., Wang, K., Xu, J. C., Sanz-Fernandez, M. V., Baumgard, L. H., et al. (2019). Effects of Source on Bioavailability of Selenium, Antioxidant Status, and Performance in Lactating Dairy Cows during Oxidative Stress-Inducing Conditions. *J. Dairy Sci.* 102 (1), 311–319. doi:10.3168/jds.2018-14974
- Sutton, J. D., Broster, W. H., Schuller, E., Napper, D. J., Broster, V. J., and Bines, J. A. (1988). Influence of Plane of Nutrition and Diet Composition on Rumen Fermentation and Energy Utilization by Dairy Cows. *J. Agric. Sci.* 110 (2), 261–270. doi:10.1017/s0021859600081284
- Tajima, K., Aminov, R. I., Nagamine, T., Matsui, H., Nakamura, M., and Benno, Y. (2001). Diet-dependent Shifts in the Bacterial Population of the Rumen Revealed with Real-Time PCR. *Appl. Environ. Microbiol.* 67 (6), 2766–2774. doi:10.1128/aem.67.6.2766-2774.2001
- Umar, S. I. U., Konwar, D., Khan, A., Bhat, M. A., Javid, F., Jeelani, R., et al. (2021). Delineation of Temperature-Humidity index (THI) as Indicator of Heat Stress in Riverine Buffaloes (Bubalus Bubalis) of a Sub-tropical Indian Region. *Cell Stress and Chaperones* 26, 657–669. doi:10.1007/s12192-021-01209-1
- Van Soest, P. J., Robertson, J. B., and Lewis, B. A. (1991). Methods for Dietary Fiber, Neutral Detergent Fiber, and Nonstarch Polysaccharides in Relation to Animal Nutrition. *J. Dairy Sci.* 74 (10), 3583–3597. doi:10.3168/jds.s0022-0302(91)78551-2
- West, J. W. (2003). Effects of Heat-Stress on Production in Dairy Cattle. *J. Dairy Sci.* 86 (6), 2131–2144. doi:10.3168/jds.s0022-0302(03)73803-x
- Wheelock, J. B., Rhoads, R. P., VanBaale, M. J., Sanders, S. R., and Baumgard, L. H. (2010). Effects of Heat Stress on Energetic Metabolism in Lactating Holstein Cows. *J. Dairy Sci.* 93 (2), 644–655. doi:10.3168/jds.2009-2295
- Xu, L., Wang, Y., Liu, J., Zhu, W., and Mao, S. (2018). Morphological Adaptation of Sheep's Rumen Epithelium to High-Grain Diet Entails Alteration in the Expression of Genes Involved in Cell Cycle Regulation, Cell Proliferation and Apoptosis. *J. Anim. Sci. Biotechnol.* 9 (1), 32–12. doi:10.1186/s40104-018-0247-z
- Yohe, T. T., Schramm, H., White, R. R., Hanigan, M. D., Parsons, C. L. M., Tucker, H. L. M., et al. (2019). Form of Calf Diet and the Rumen. II: Impact on Volatile Fatty Acid Absorption. *J. Dairy Sci.* 102 (9), 8502–8512. doi:10.3168/jds.2019-16450
- Zhang, K., Meng, M., Gao, L., Tu, Y., and Bai, Y. (2019). Rumen-derived Lipopolysaccharide Induced Ruminal Epithelium Barrier Damage in Goats Fed a High-Concentrate Diet. *Microb. Pathogenesis* 131, 81–86. doi:10.1016/j.micpath.2019.02.007
- Zumsande, M., and Gross, T. (2010). Bifurcations and Chaos in the MAPK Signaling cascade. *J. Theor. Biol.* 265 (3), 481–491. doi:10.1016/j.jtbi.2010.04.025
- Zusso, M., Lunardi, V., Franceschini, D., Pagetta, A., Lo, R., Stifani, S., et al. (2019). Ciprofloxacin and Levofloxacin Attenuate Microglia Inflammatory Response via TLR4/NF-kB Pathway. *J. Neuroinflammation* 16 (1), 1–12. doi:10.1186/s12974-019-1538-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Guo, Gao, Ding, He, Ma and Bu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Skeletal Muscle Expression of Actinin-3 (ACTN3) in Relation to Feed Efficiency Phenotype of F₂ *Bos indicus* - *Bos taurus* Steers

Robert N. Vaughn¹, Kelli J. Kochan¹, Aline K. Torres¹, Min Du², David G. Riley¹, Clare A. Gill¹, Andy D. Herring¹, James O. Sanders¹ and Penny K. Riggs^{1*}

¹Department of Animal Science, Texas A&M University, College Station, TX, United States, ²Department of Animal Sciences, Washington State University, Pullman, WA, United States

OPEN ACCESS

Edited by:

Marcos De Donato,
Instituto de Tecnología y Educación
Superior de Monterrey (ITESM),
Mexico

Reviewed by:

Pablo Fonseca,
University of Guelph, Canada
Marcio Duarte,
University of Guelph, Canada

*Correspondence:

Penny K. Riggs
riggs@tamu.edu

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 15 October 2021

Accepted: 10 January 2022

Published: 03 February 2022

Citation:

Vaughn RN, Kochan KJ, Torres AK,
Du M, Riley DG, Gill CA, Herring AD,
Sanders JO and Riggs PK (2022)
Skeletal Muscle Expression of Actinin-
3 (ACTN3) in Relation to Feed
Efficiency Phenotype of F₂ *Bos indicus*
- *Bos taurus* Steers.
Front. Genet. 13:796038.
doi: 10.3389/fgene.2022.796038

In this study, actinin-3 (ACTN3) gene expression was investigated in relation to the feed efficiency phenotype in *Bos indicus* - *Bos taurus* crossbred steers. A measure of relative feed efficiency based on residual feed intake relative to predictions from the NRC beef cattle model was analyzed by the use of a mixed linear model that included sire and family nested within sire as fixed effects and age, animal type, sex, condition, and breed as random effects for 173 F₂ Nellore-Angus steers. Based on these residual intake observations, individuals were ranked from most efficient to least efficient. Skeletal muscle samples were analyzed from 54 steers in three groups of 18 (high efficiency, low efficiency, and a statistically average group). ACTN3, which encodes a muscle-specific structural protein, was previously identified as a candidate gene from a microarray analysis of RNA extracted from muscle samples obtained from a subset of steers from each of these three efficiency groups. The expression of ACTN3 was evaluated by quantitative reverse transcriptase PCR analysis. The expression of ACTN3 in skeletal muscle was 1.6-fold greater in the inefficient steer group than in the efficient group ($p = 0.007$). In addition to expression measurements, blocks of SNP haplotypes were assessed for breed or parent of origin effects. A maternal effect was observed for ACTN3 inheritance, indicating that a maternal *B. indicus* block conferred improved residual feed efficiency relative to the *B. taurus* copy ($p = 0.03$). A SNP haplotype analysis was also conducted for m-calpain (CAPN2) and fibronectin 1 (FN1), and a significant breed effect was observed for both genes, with *B. indicus* and *B. taurus* alleles each conferring favorable efficiency when inherited maternally ($p = 0.03$ and $p = 0.04$). Because the ACTN3 structural protein is specific to fast-twitch (type II) muscle fibers and not present in slow-twitch muscle fibers (type I), muscle samples used for expression analysis were also assayed for fiber type ratio (type II/type I). Inefficient animals had a fast fiber type ratio 1.8-fold greater than the efficient animals ($p = 0.027$). Because these fiber-types exhibit different metabolic profiles, we hypothesize that animals with a greater proportion of fast-twitch muscle fibers are also less feed efficient.

Keywords: bovine, feed efficiency, fiber type, metabolism, gene network, SNP, ACTN3, sustainability

INTRODUCTION

Efficient use of resources is regarded as a critical area of emphasis for livestock production (Kenny et al., 2018; Brito et al., 2021). Feed costs can contribute to 70% of total livestock production costs (Becker 2008). Environmental costs associated with beef production have become increasingly important to the consumer. Although the beef industry made substantial strides in reducing the environmental footprint of cattle production over several decades (Capper 2011), feed efficiency remains a target today in beef producers' efforts toward sustainability. In considering nutrient security for humans, utilizing new and emerging technologies to better understand the complex physiological mechanisms of production traits will be key for enabling even greater efficiency of food animal production (Riggs et al., 2017; Riggs et al., 2018). Increased access to nutrient-dense animal source foods requires greater production volume if global population growth reaches 9.8 billion people in the coming decades (FAO, 2018). Additionally, a systematic review of literature that examined life cycle assessment in cattle (published between 2000 and 2016) contained climate change as a category, with nearly a third of those papers focusing on only that topic (McClelland et al., 2018). Understanding feed efficiency in the context of breeds adapted to specific environmental conditions—particularly hot, dry conditions found in much of the world's grazing lands—has important economic and environmental implications.

Brunes et al. (2021) noted that feed efficiency traits in beef cattle are highly complex and affected by numerous biological and polygenic mechanisms. In their GWAS analysis of feed efficiency from 4,329 Nellore cattle, multiple genomic regions were identified that harbor sequence variation associated with greater than 0.5% of the additive genetic variance for feed efficiency in the population (Brunes et al., 2021). Genes within the identified regions contribute to the metabolism of protein, lipids, and numerous metabolic, energetic, and behavioral pathways. A study of dairy cattle in New Zealand (Puillet et al., 2021) examined gene-by-environment interactions. As might be expected, this work demonstrated that the relationships between feed efficiency and other production traits are complex and must be carefully balanced when one considers overall measures of lifetime production efficiencies.

In the present study, we focused on aspects of feed efficiency influenced by skeletal muscle by specifically examining *ACTN3* within a genetic mapping herd of beef cattle that utilized *Bos indicus*–*Bos taurus* cattle suited to the subtropical environment of Texas. Resting energy consumption contributes to feed efficiency as a major component of an animal's overall energy requirements. Skeletal muscle has lower energy requirements by weight compared to other tissue types. However, because skeletal muscle mass contributes as much as half of the live animal's weight, a great deal of an animal's total caloric intake may be utilized for the growth and maintenance of this tissue. Approximately 20% of an animal's overall energy consumption is consumed by the skeletal muscle (Ortigues 1991), and this percentage can increase drastically during stress, activity, or changes in temperature.

We evaluated *ACTN3* sequence haplotypes and gene expression in F₂ Nellore-Angus steers that had been evaluated for efficiency. For the study we describe here, we specifically wanted to investigate whether different *ACTN3* alleles in this population corresponded to feed efficiency rates. This gene was identified as one candidate contributing to a feed efficiency phenotype from a pilot gene expression network analysis, and we suspected that animals differed in muscle turnover rate. We hypothesized that we could identify genetic variation in *ACTN3* responsible for differences in muscle physiology and homeostasis that are important for overall feed efficiency. In pigs, a relationship between muscle turnover and residual feed intake was previously described (Cruzen et al., 2013). In beef cattle, the negative correlation of muscle proteolysis rate and feed efficiency has been noted (McDonagh et al., 2001). Other investigators have evaluated skeletal muscle gene expression in beef cattle feed efficiency studies. For example, Lam et al., 2020 developed an RNA sequencing pipeline from liver and muscle transcriptomes (Tizioto et al., 2016) to identify single nucleotide polymorphisms (SNP) associated with feed efficiency in Nellore steers classified with high or low residual feed intake (RFI). In this study, the authors suggested relationships between feed efficiency and both immune and metabolic function in livestock. Regulation of the *RAC1* gene was associated with less efficient cattle. Interestingly, the genetic pathways that include *RAC1* and *ACTN3* may have related functional roles associated with calcium metabolism-based regulation of immune function (Calender et al., 2020).

Also contributing to energetic requirements, different skeletal muscle fiber types have different energy requirements. Type II, or fast-twitch fibers, have greater energy demands than type I, or slow-twitch fibers. A study in horses identified *ACTN3* as a potential element associated with muscle performance (Ropka-Molik et al., 2019). Welch et al., 2013 found a positive correlation between relative feed intake (RFI) and type II muscle fibers. This finding is consistent with the idea that the physiological skeletal muscle phenotype of an animal contributes to its overall efficiency phenotype. Because *ACTN3* is expressed only in type II fibers, this study also examined fiber type profiles in muscle samples.

MATERIALS AND METHODS

Feed Efficiency Phenotype

For this study, feed and carcass data, obtained from 173 Nellore-Angus F₂ steers in 13 full-sibling families produced by embryo transfer from the Texas A&M McGregor Genomics research population located in central Texas (latitude: 31.3865, longitude: −97.4105), were utilized, as described by Amen (2007). Briefly, calves were produced in the spring and fall calving seasons, and this study used calves born from 2003 to 2005. After weaning (approximately 230 days of age), the animals were grass-fed for approximately 130 days until they reached 11–13 months of age. Steers were moved to a Calan Broadbent Feeding System (American Calan, Northwood, NH, USA) to measure individual feed intake. Over 28 days, the steers were

TABLE 1 | Ration formulation used for steers in study.

Ingredient	%
Ground milo	20.00
Ground corn	31.25
Cottonseed meal	9.00
Cottonseed hulls	25.00
Molasses	10.00
Premix ^a	3.00
Ammonium chloride	0.25
R-1500 ^b	1.50

Ingredients are represented as a percent on an as-fed basis.

^aComposition of premix: ground limestone, 60%; trace mineralized salt, 16.7% (NaCl, 98%; Zn, 0.35%; Mn, 0.28%; Fe, 0.175%; Cu, 0.035%; I, 0.007%; Co, 0.007%); monocalcium phosphate, 13%; potassium chloride 6.7%; vitamin premix, 3.3% (vitamin A, 2,200,000 IU/kg; vitamin D, 1,100,000 IU/kg; vitamin E, 2,200 IU/kg); zinc oxide, 0.33%.

^bR-1500 contains 1.65 g monensin sodium (RumensinTM) per kg.

adjusted to the finishing diet (Table 1). Feeding was ad libitum, and uneaten feed was removed and measured every 7 days. Animals were weighed every 28 days for ~140 days at the same time of day and in the same order of pens at each weigh day to equalize gut fill effects across time, as much as possible.

Using the NRC (2000) model application, the daily feed intake required to achieve observed ADG was predicted. Model inputs included feeding period mid-weight (used as the current reference weight for model predictions) and final weight at slaughter for each animal. Standardized inputs were used for animal type (growing/finishing), age (14 months), sex (steer), BCS (5), breed (Nellore-Angus two-way cross), management (ionophore), diet (Table 1), and quality grade (select). Environmental factors in the model were set to be thermoneutral. The model predicted intake was subtracted from the observed dry matter intake (DMI), and the difference is defined as residual feed intake based on the NRC model (RFI_{NRC}), such that those animals that consumed less than predicted (thus, were more efficient) had negative RFI_{NRC} (Liu et al., 2000). Mixed procedures of SAS were then used to analyze RFI_{NRC} with fixed factors of sire and family nested within sire (Amen 2007). This method of residual prediction parallels more typical calculations of RFI, in which individual residuals represent a deviation from a modeled mean intake at an observed rate of gain. However, the standard method is generally restricted to use within a single contemporary group fed simultaneously or with the use of the contemporary group as a model effect. Importantly, in this case, a standard model (the NRC) was utilized to construct the modeled mean intake rather than a regression on observed data, thus enabling animals in multiple contemporary groups to be evaluated against a common model.

Sample Selection

For gene expression analysis, 36 animals were identified at the tails of the efficiency distribution based on RFI_{NRC} as described above. A total of 18 animals were classified as most “efficient” for this population and had negative RFI_{NRC} residuals indicating that they had consumed less feed than would be expected based on the model. A total of 18 animals were classified as most “inefficient,”

TABLE 2 | Simple means (±STD err) for RFI_{NRC} residuals by efficiency groups.

Item	Efficient			Average			Inefficient		
<i>n</i>	18			18			18		
RFI _{NRC} residuals	−2.3	±	0.1 ^a	0.0	±	0.0 ^b	2.3	±	0.1 ^c

^aMeans within a row with different superscripts differ significantly ($p < 0.01$).

^bMeans within a row with different superscripts differ significantly ($p < 0.01$).

^cMeans within a row with different superscripts differ significantly ($p < 0.01$).

TABLE 3 | Birth year season (BYS) distribution among efficiency groups.

Group	<i>N</i>	F03	F04	F05	S03	S04	S05
Efficient	18	0	1	0	1	2	14
Average	18	3	4	4	1	4	2
Inefficient	18	5	4	6	1	2	0

Season is designated as Fall (F) or Spring (S).

TABLE 4 | Frequency table of family distribution across BYS contemporary groups from animals evaluated for feed efficiency in a previous study (Amen, 2007).

Family ID ^a	Contemporary group						Total
	S2003	F2003	S2004	F2004	S2005	F2005	
70	1	5	4 (E 1)	2 (I 1)	1 (E 1)	4	17
71	2	2	5	4	2 (E 2)	0	15
72	4 (E 1)	0	5 (I 1)	0	2 (E 2)	7	18
73	2	3	0	0	0	0	5
74	4	0	0	0	0	0	4
75	5 (I 1)	0	0	2	4 (E 1)	0	11
76	2	3	0	0	0	0	5
77	1	5 (I 5)	1 (I 1)	1	9 (E 3)	0	17
80	0	7	3 (E 1)	16 (I 1)	0	1	27
81	0	1	11	3	5 (E 3)	5 (I 3)	25
82	0	0	0	0	0	6	6
83	0	0	3	2	4 (E 2), (I 1)	2 (I 2)	11
84	0	0	0	1	7 (E 1)	4 (I 1)	12
Total	21	26	32	31	34	29	173

Season is designated as Fall (F) or Spring (S). Distribution of animals identified in the group of 18 most efficient (E-n) or 18 most inefficient (I-n) evaluated in the current study is identified in parentheses.

^aSire ID and respective families 297J–70, 71; 432H–72, 73; 437J–74, 75, 81, 82, 83; 551G–76, 77, 80, 84. Although not all families are represented in all contemporary groups, sires are much more evenly distributed across these groups. All sires had at least three steers per contemporary group except for 432H in Fall 2004 ($n = 0$) and Spring 2005 ($n = 2$) and 437J in Fall 2003 ($n = 1$).

with positive RFI_{NRC} residuals indicating they had consumed more feed than would be expected based on the model. Muscle samples from these 36 animals were used for subsequent expression analysis. In addition, a statistically average group of 18 animals with an RFI_{NRC} residual clustered around zero was included for comparison purposes. Thus, a total of 54 animals from the middle, and both tails of the residual distribution were analyzed for gene expression. Means for these groups are presented in Table 2. The distribution of these groups across

birth year seasons (BYS) is shown in **Table 3**. In addition, the frequency distribution of animals from the Amen (2007) study is presented in **Table 4**, with the set of most and least efficient animals shown in parentheses. Amen (2007) noted that imbalance exists among families across BYS. While this imbalance may complicate efforts to separate family effects from year-season effects, the approach provides a means to examine the combined overall impact of gene X environmental interactions on resultant phenotypes to understand the influence of gene function on phenotype despite a range of regional climate conditions.

Tissue Collection and Extraction of RNA

Steers were harvested at 18 months of age at the Rosenthal Meat Center at Texas A&M University in College Station, TX using humane harvesting procedures, as described by Savell and Smith (2000). Animals were restricted from feed for approximately 12 h before harvest but had continual access to water. Animals were immobilized using a captive bolt stunning mechanism and further processed using standard industry procedures. Approximately 1 g of muscle tissue from the *Longissimus cervicis* (in the neck region of the carcass) was collected before electrical stimulation (ES; less than 1 h after exsanguination) of the carcass. The muscle sample was flash-frozen in liquid nitrogen to prevent mRNA degradation. Samples were stored at -80°C until RNA was extracted.

Total RNA was extracted from a portion (~100–200 mg) of the frozen whole muscle tissue samples (*L. cervicis*) from each of the 54 animals with TRI Reagent® (Molecular Research Center, Cincinnati, OH, USA) and 1-bromo-3-chloropropane (Molecular Research Center). Next, RNA was precipitated with isopropanol (Sigma Aldrich, St. Louis, MO, USA), washed with 70% ethanol (Sigma Aldrich), and reconstituted in 50 μl nuclease-free water (Invitrogen, Carlsbad, CA, USA). The quality of the RNA was assessed via an Agilent 2100 series Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) according to the manufacturer's protocol. Samples with an RNA integrity number (RIN) >8.0 and appropriate electropherogram image were treated with DNase (Invitrogen) and column-purified via the RNeasy Mini kit (Qiagen, Valencia, CA, USA). Samples were quantified by spectrophotometry (NanoDrop 1000), and identical quantities of total RNA for each sample were utilized in downstream applications. RNA extracts were stored at -80°C prior to expression analyses.

Microarray Procedures

Microarray analysis (described in Vaughn, 2013; Riggs and Vaughn, 2015) was conducted prior to this study as a pilot study with the Agilent 44k bovine array (*B. taurus* oligo microarray V2 Agilent 4x44k GPL11649) according to manufacturer's recommendations. Twenty-four RNA samples (200 ng each from eight most efficient and eight least efficient animals) were labeled for two-color microarray gene expression analysis (Quick Amp Labeling Kit, Agilent Technologies) according to the manufacturer's protocol. Following hybridization and scanning, normalization and quality control were performed using embedded quantile

normalization functions in the Genespring GX v11.0.2 software. Following quality filtering, the Mann–Whitney unpaired test was utilized. A non-parametric test was used to avoid relying on the *a priori* assumption of a normal distribution of gene expression within these populations selected for extremes. In this experimental design, samples fail the assumption of independence, making false discovery analyses invalid. A cut-off of $p \leq 0.05$ and a fold-difference of 1.4-fold or greater were set as thresholds to generate lists of probes for subsequent pathways analysis. Pathway analysis was performed using DAVID Bioinformatics Resources 6.7 (<http://david.ncifcrf.gov/>). Gene Ontology analysis was also performed within the DAVID software using the same data set used for pathway analysis. The GO-fat category was used with the default ease setting of 0.1 and a minimum count of 2. From these pathways analyses, as well as a separate and independent array and pathway experiments (not described here), muscle metabolic pathways we formed a hypothesis related to genes involved in skeletal muscle turnover. As a result, *ACTN3* was identified as a target for further investigation by qRT-PCR and haplotype analysis. Array data are deposited in the Gene Expression Omnibus, GEO accession number GSE56705 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56705>.

Real Time Quantitative Reverse Transcriptase Polymerase Chain Reaction

Quantitative RT-PCR analysis was conducted on the full set of 54 samples. Synthesis of cDNA from all RNA samples with the high-capacity cDNA Reverse Transcription Kit (Invitrogen) was accomplished with a starting quantity of 2 μg mRNA per 40 μl reaction. Oligo(dT)20 primers (Integrated DNA Technologies, Coralville, IA, USA), 5 μM final concentration, were used for reverse transcription. cDNA was diluted 1:2 in yeast tRNA (25 ng/ μl ; Invitrogen). Samples were amplified in a total volume of 20 μl containing 1X SYBR GreenER™ qPCR SuperMix (Invitrogen), 300 nM primers and 2 μl template cDNA. Real-time quantitative RT-PCR (qRT-PCR) was performed at 95°C for 10 min followed by 40 cycles of 15 s at 95°C and 60 s at 60°C , in a 7900HT thermal cycler (Applied Biosystems).

Primers for genes validated by qRT-PCR were designed within Oligo 6 Primer Analysis Software v6.71 (Molecular Biology Insights, Cascade, CO, USA). The analyzed genes included actinin-2 (*ACTN2*), actinin-3 (*ACTN3*), adipose differentiation-related protein (*ADFP*, also known as perilipin 2 (*PLIN2*)), ATP synthase H⁺ transporting, mitochondrial F1 complex, beta polypeptide (*ATP5B*), hexokinase 2 (*HK2*), and lactate dehydrogenase B (*LDHB*). Ribosomal protein S20 (*RPS20*) was utilized as a reference gene. Genes and primers used are described in **Table 5**. Primer pairs were evaluated by BLAST sequence similarity search (Altschul et al., 1990). Primer pairs were selected, which did not cross-amplify across species or between different mRNA transcripts. Additionally, primers were selected to span an exon junction to prevent genomic amplification.

TABLE 5 | Complete list of primer pairs used for qRT-PCR assays and their function in muscle.

Gene symbol	Description	Sequence ^a
<i>ACTN2</i>	Actinin, alpha 2	5'-GGTCTTTGACAACAAGCA-3' 5'-TGATGGTTCTGGCGATA-3'
<i>ACTN3</i>	Actinin, alpha 3	5'-CGGGAGACAAGAACTACATCA-3' 5'-CGTAGAGGGCACTGGAGAA-3'
<i>ATP5B</i>	ATP synthase, H+ transporting Mitochondrial F1 complex Beta polypeptide	5'-CCCATCAAAACCAAGCAA-3' 5'-TCAACACTCATCTCCACGAA-3'
<i>CAPN1</i>	Calpain 1, (m/l) large subunit	5'-GACCATAGGCTTCGCTGTCT-3' 5'-AGGTTGATGAAGTCTCGGA-3'
<i>CAPN2</i>	Calpain 2, (m/l) large subunit	5'-CGACTGGAGACACTGTTTCAGGA-3' 5'-CTTCAGGCAGATTGGTTATCACTT-3'
<i>CAST</i>	Calpastatin	5'-GCTGTCGTCTCTGAAGTGGTT-3' 5'-GGCATCGTCAAGTCTTTGTTGT-3'
<i>HKII</i>	Hexokinase 2	5'-TCAACACTCATCTCCACGAA-3' 5'-CACCACAGCAACCACATC-3'
<i>LDHB</i>	Lactate dehydrogenase B	5'-CAGAAATGGGAACAGACAA-3' 5'-GACTTCATAGGCACTCTCAAC-3'
<i>MYH1</i>	Myosin, heavy chain 1, skeletal muscle, adult	5'-TGAGGAAGCGGAGGAACAAT-3' 5'-TGGGACTCGGCAATGTCA-3'
<i>MYH2</i>	Heavy chain 2, skeletal muscle, adult	5'-CAATGACCTGACAACCCAGA-3' 5'-CCTTGACAACCTGAGACACCAGA-3'
<i>RPS20</i>	Ribosomal protein S20	5'-ACCAGCCGCAACGTGAA-3' 5'-CCTTCGCGCCTCTGATCA-3'

^aRPS20 primer sequences provided by Kochan et al. (2009).

To quantify qRT-PCR results, amplification data were analyzed *via* Sequence Detection System v2.4 software (Applied Biosystems, Carlsbad, CA, USA). Expression was normalized to RPS20 as a reference gene (De Jonge et al., 2007), and relative expression was quantified as described by Livak and Schmittgen (2001). A threshold value of 0.20 was used to determine the C_t value. In summary, raw C_t values were normalized to RPS20 based on internal expression stability between groups (Vandesompele et al., 2002). The normalized value was subtracted from the raw C_t for each sample (ΔC_t). From the ΔC_t , the average value of the efficient group was subtracted ($\Delta\Delta C_t$). The $\Delta\Delta C_t$ value was linearized by conversion to the inverse negative \log_2 (RQ). The efficient group in this study thus has a median expression value of 1.0. It should be noted that these are arbitrary units (AU) of expression and that no direct comparison between different genes in total expression levels can be reliably made. All values are relative and applicable directly only as a within-group comparison of relative expression.

Muscle Fiber Type Classification

Fiber type analysis was determined by gel electrophoresis. *L. cervicis* muscle samples (0.1 g) were homogenized in 500 μ l buffer consisting of 250 mM sucrose, 100 mM KCl, 5 mM EDTA, and 20 mM Tris-HCl pH 6.8. The homogenate was filtered through nylon cloth to remove debris and centrifuged at $10,000 \times g$ for 10 min. The pellet obtained was resuspended in 500 μ l washing buffer (200 mM KCl, 5 mM EDTA, 0.5% Triton X-100, and 20 mM Tris-HCl pH 6.8). The suspension was centrifuged at $10,000 \times g$ for 10 min. The pellet containing purified myofibrillar proteins was resuspended in 200 μ l water and 300 μ l of standard 2X sample loading buffer, boiled for 5 min, and then centrifuged at $12,000 \times g$ for 5 min. The resultant supernatant was used for electrophoresis.

TABLE 6 | SNP haplotype block location information.

Gene	Chromosome	Coordinate (UMD 3.1)		Number of SNPs in 1 Mb region
<i>CAPN1</i>	29	44063463	44113492	24
<i>CAPN2</i>	16	27781671	27840011	24
<i>CAPN3</i>	10	37829007	37885645	24
<i>CAST</i>	7	98444979	98581253	18
<i>ACTN2</i>	28	9403203	9450920	16
<i>ACTN3</i>	29	45230630	45242406	16
<i>MYH1/2</i>	19	30110728	30165109	18
<i>FN1</i>	2	103881402	103950562	12

Locations according to Bos taurus UMD 3.1 genome assembly.

Stacking gels consisted of 4% acrylamide (acrylamide: bis = 50:1) and 5% (v/v) glycerol, 70 mM Tris-HCl pH 6.7, 0.4% (w/v) SDS, 4 mM EDTA, 0.1% (w/v) APS, and 0.01% (v/v) TEMED. The separation gel contained 5% (w/v) glycerol, acrylamide: bis (50:1) at a concentration ranging from 5 to 20%, 200 mM Tris pH 8.8, 4 mM EDTA, 0.4% (w/v) SDS, 0.01% (v/v) TEMED, and 0.1% (w/v) ammonium persulfate. The upper running buffer consisted of 0.1 M Tris-HCl pH 8.8, 0.1% (w/v) SDS, 150 mM glycine, and 10 mM mercaptoethanol, and the lower running buffer consisted of 50 mM Tris-HCl pH 8.8, 0.01% (w/v) SDS, and 75 mM glycine. Gels were run at 8°C in a Hoefer SE 600 (Hoefer Scientific, San Francisco, CA, USA) unit, at constant 200 V for 24 h (Bamman et al., 1999). After electrophoresis, gels were stained with Coomassie blue and scanned with a densitometer to determine the amount of each myosin isoform and the percentage of type I and type II muscle fibers was reported (Underwood et al., 2007).

Haplotype Analysis

All individuals ($n = 776$) from the first three generations of the experimental population were previously genotyped with the Illumina BovSNP50 v1 assay (Illumina, Inc., San Diego, CA, USA), and data were available for use in this study. The haplotype analysis from these data allowed us to trace the inheritance of Nellore or Angus blocks of SNPs and identify the parent of origin for the 173 Nellore-Angus F_2 steers for which RFI_{NRC} data and tissue samples were collected and used in this study. After pruning genotypes to remove animals with poor completion rate (<0.9), SNP with low minor allele frequency (<0.05), SNP with poor completion rates (<0.9), and those SNP that deviated from Hardy–Weinberg equilibrium ($p < 0.0001$), 39,890 genotypes per individual remained. To determine whether breed of origin or parent of origin played a role in the efficiency phenotype, SNP genotypes spanning 1 Mb intervals flanking several genes of interest (Table 6) based on expression analyses were extracted using PLINK v1.07 (Purcell et al., 2007) and formatted for phase v2.1.1 software (Stephens et al., 2001). Haplotypes were established using 100 iterations of phase v2.1.1, with a thinning interval of 2 and a burn-in of 10. Resultant phased haplotypes were ordered by generation, and breed (Nellore or Angus) and parent of origin were manually tracked through the pedigree to assign the source of each haplotype block in the F_2 generation. Because gene expression analyses of μ -calpain (CAPN1), m-calpain (CAPN2), calpastatin (CAST), myosin heavy chain 1 (MYH1), and myosin heavy chain 2 (MYH2) genes were also available from another study and are relevant to pathways associated with muscle growth and proteolysis (Vaughn 2013), their relationship with feed efficiency was also examined. The ACTN2 gene was selected for analysis because it is expressed in all skeletal muscle fiber types and has conserved structural and functional similarity to ACTN3. The genes CAPN1, CAPN2, and CAST were examined because of their roles in muscle proteolysis (Goll et al., 2003), and MYH1 and MYH2 were selected for their function in muscle fibers (Wang et al., 2012).

Statistical Analysis

SPSS 16.0 software (IBM, Armonk, NY) was used for all statistical analyses. To test for significance between efficiency groups, an independent samples t -test was used. Correlation analysis used a bivariate two-tailed Pearson's correlation. All qRT-PCR expression was normalized to the RPS20 reference gene (De Jonge et al., 2007). Expression data are reported relative to the feed efficient group, which was set to a value of 1.0. The inefficient and average groups are represented as fold difference relative to the efficient group. One sample was removed from all analyses because expression values measured for the reference gene, RPS20, were not consistent with other samples.

RESULTS

Gene Expression Analysis

Following an initial pilot microarray (GEO accession number GSE56705) and network pathway analysis, the ACTN3 gene was

TABLE 7 | Relative mRNA expression of selected genes by qRT-PCR analysis.

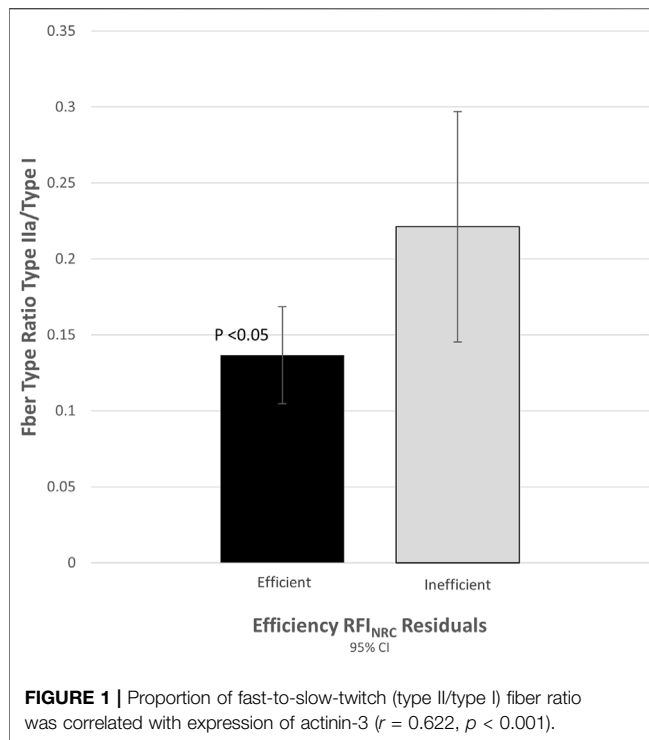
Gene	Ratio average/efficient			Ratio inefficient/efficient		
N		18			18	
ACTN3	1.6	\pm	0.06 ^a	1.6	\pm	0.05 ^a
ACTN2	1.2	\pm	0.050	1.0	\pm	0.050
COX3	1.7	\pm	0.05 ^a	1.4	\pm	0.06 ^b
COX7C	1.1	\pm	0.030	0.9	\pm	0.040
HKII	0.9	\pm	0.040	0.8	\pm	0.050
LDHB	0.6	\pm	0.03 ^a	0.9	\pm	0.04 ^b
PRDX3	0.9	\pm	0.030	1.0	\pm	0.040

Relative expression for the average and inefficient groups is calculated and presented as a fold ratio compared to the efficient group. Expression was normalized to RPS20, and expression value is presented as the geometric mean ratio of arbitrary expression units relative to expression in the efficient group. Each group (average, efficient, inefficient) reflects samples from 18 animals.

^aMeans with superscripts differ from the efficient group ($p < 0.05$).

^bMeans with different superscripts differ from each other ($p < 0.05$).

selected for further investigation. From the initial hypothesis generating microarray experiment, 58 genes were expressed significantly differently between efficiency groups with a fold difference between groups of 1.5-fold or greater (Supplementary Table S1). The ACTN3 gene was expressed as 2.5-fold greater in the inefficient group of steers but fell short of significance ($p = 0.051$). However, because different alleles of this gene have been associated with differential athletic performance in humans, and ACTN3 genotype is thought to contribute to variation in muscle phenotype (North et al., 2003; Yang et al., 2003), ACTN3 was selected for further investigation. Despite missing the initial statistical cut-off, ACTN3 also remained of interest because it is a member of skeletal muscle networks determined to be relevant. Following these early investigations, this study was developed to examine ACTN3 as a candidate gene contributing to a feed efficiency phenotype. Genes used for qRT-PCR analysis are described in Table 5. Additional genes that have some relationship to ACTN3 and the skeletal muscle pathways in which it participates were included in qRT-PCR expression and haplotype analyses. These genes were selected prior to expression and haplotype analysis as negative control genes or as genes expected to function similarly in given pathways. We thought that, by examining haplotypes according to parent of origin, we may be able to differentiate favorable *Bos indicus* and *Bos taurus* alleles that would be of use for selection. The ACTN2 gene was selected for analysis because it is expressed in all skeletal muscle fiber types and has conserved structural and functional similarity to ACTN3. Although their functions are not identical, evidence exists that ACTN2 can largely compensate for the absence of ACTN3 (Mills et al., 2001; Lek et al., 2010). The previously evaluated reference genes were tested for stability, and RPS20 was utilized as a control reference gene. Because of our hypothesis that ACTN3 was associated with feed efficiency and might reflect differences in overall metabolic activity, we also selected a set of genes that are known to play a role in other metabolic processes within muscle tissue to examine at the same time. For example, ADFP was selected because of its role in lipid metabolism and storage (Hiller et al., 2012). Genes that place a role in muscle metabolism and glucose homeostasis (ATP5B, HKII, and LDHB) were also examined because these are genes that play a



role in muscle metabolism (Rajesh et al., 2011; Egan and Zierath 2013; Li et al., 2013). The genes *CAPN1*, *CAPN2*, and *CAST* were examined because of their roles in muscle proteolysis (Goll et al., 2003), and *MYH1* and *MYH2* were selected for their function in muscle fibers (Wang et al., 2012).

By qRT-PCR, expression of *ACTN3* was 1.6-fold greater ($p = 0.009$) in the average and inefficient groups than in the efficient group (Table 7). *ACTN2* expression did not vary significantly between any of the groups, nor did it correlate significantly with *ACTN3* expression or fiber type ratio (as expected). Of the other genes assayed, *LDHB* expression was significantly lower in the average group but not the inefficient group, relative to the efficient steers. The mRNA quantity of the remaining genes examined (Table 7) did not differ significantly between groups.

Post-transcriptional modifications, other regulatory mechanisms, and differences in the timing of expression can make the correlation of mRNA expression with protein expression difficult (Greenbaum et al., 2002). To verify that the observed difference in *ACTN3* gene expression translated to actual differences in muscle protein expression, 12 samples (portions of the same samples used for qRT-PCR) from each tail of the distribution ($n = 24$) were assayed for fiber type ratios based on gel separation of muscle fiber type specific isoforms. The ratio of fast-twitch to slow-twitch muscle fiber (type II/type I) was 1.8-fold greater ($p = 0.027$) in the inefficient group compared to the efficient group (Figure 1).

Haplotype Analysis

Haplotype block analysis was conducted for *ACTN2*, *ACTN3*, *CAPN1*, *CAPN2*, *CAPN3*, *CAST*, *FN1*, and *MYH1/2*. Three of the haplotype blocks were associated with significant differences in

efficiency between Angus and Nellore in the larger population studied. The *CAPN2* and *ACTN3* Nellore haplotype blocks were associated with a superior efficiency when inherited maternally ($p = 0.03$). The *FN1* Angus haplotype block was associated with a superior efficiency when inherited maternally ($p = 0.04$). Neither the *CAST* nor the *CAPN1* haplotypes were associated with any improvement in efficiency (Table 8).

Season of test feeding period may have influenced efficiency. A general linear model was used to produce the least square means using RFI_{NRC} residuals as the dependent variable with birth year season (BYS) as the fixed factor described in Amen (2007). Six BYS groups were included in this study: F03 ($n = 26$), F04 ($n = 31$), F05 ($n = 29$), S03 ($n = 21$), S04 ($n = 32$), and S05 ($n = 34$), with a total of 87 spring-born animals weaned in the fall and 85 weaned in the spring. Steers weaned in the spring were more efficient than those weaned in the fall; linear contrast of fall RFI_{NRC} residuals means minus spring RFI_{NRC} residuals means was $1.09 + 0.15$ ($p < 0.0001$). One BYS level, Spring of 2005, had a lower mean RFI than the other BYS groups ($p < 0.0001$). Within the subset selected for expression analysis, 14 of the 18 samples in the efficient group were from that BYS level (Table 3). The distribution of most and least efficient animals examined is also described in Table 4.

ACTN3 expression was higher in fall weaned animals than spring weaned animals by 2.1-fold ($p < 0.001$). No significant differences between seasons were noted in any other genes assayed. Additionally, a difference ($p < 0.001$) in fiber type by the season of weaning was observed. Those calves weaned in the fall ($n = 11$) had a fast-to-slow-twitch (type II/type I) fiber type ratio 1.6-fold greater than those weaned in the spring ($n = 13$). These ratios were 0.23 and 0.14, respectively.

DISCUSSION

Multiple biological and developmental processes, driven by complex genetic networks and response to environmental conditions, each contribute to phenotypic measures of efficiency and the overall relationships of inputs (e.g., feed intake) to outputs (e.g., meat and other products). The identification of specific genetic variation contributes to better understanding and insight into mechanisms of growth and energy utilization. Such knowledge can drive further gains in productivity, reducing waste and environmental impacts. Increased feed efficiency can reduce production costs (Ho et al., 2013) and environmental footprint, both of which are important in a world with an expanding population to feed and finite resources for the production of necessary animal source foods. Waste concerns, from feed waste to methane production, have also gained emphasis in recent years (Brito et al., 2020). Simultaneously, consumers are demanding environmentally conscious options, while producers strive to maintain fiscal viability.

In the present study, a difference in *ACTN3* expression between groups was confirmed by qRT-PCR analysis in skeletal muscle samples from the high and low feed efficiency groups of steers. Expression of *ACTN2* did not differ between

TABLE 8 | RFI residuals based on haplotype block analysis.

		Haplotype block								Paternal				Maternal			
		NN		NA		AN		AA		NN/NA		AN/AA		NN/AN		NA/AA	
CAST	RFI res	-0.7	± 0.5	0.5	± 0.4	-0.3	± 0.4	0.5	± 0.4	0.0	± 0.3	0.1	± 0.3	-0.5 ^a	± 0.3	0.5 ^b	± 0.3
	<i>n</i>	46		43		42		43		89		85		88		86	
CAPN1	RFI res	0.5	± 0.4	-0.8	± 0.4	0.2	± 0.3	0.2	± 0.5	-0.1	± 0.3	0.2	± 0.3	0.4	± 0.3	-0.3	± 0.3
	<i>n</i>	23		39		37		75		62		112		60		114	
CAPN2	RFI res	-0.6	± 0.6	0.4	± 0.4	-0.3	± 0.4	0.2	± 0.4	0.0	± 0.3	0.0	± 0.3	-0.4	± 0.3	0.3	± 0.3
	<i>n</i>	41		52		40		41		93		81		81		93	
FN1	RFI res	-0.7	± 0.5	0.5	± 0.4	-0.3	± 0.4	0.4	± 0.5	0.0	± 0.3	0.1	± 0.3	-0.5 ^a	± 0.3	0.5 ^b	± 0.3
	<i>n</i>	59		52		32		31		111		63		91		83	
	RFI res	0.4	± 0.3	-0.5	± 0.5	0.6	± 0.5	-0.4	± 0.5	0.0	± 0.3	0.1	± 0.3	0.5 ^a	± 0.3	-0.4 ^b	± 0.3

NN, NA, AN, AA refer to breed haplotype. N refers to a haplotype inherited from Nellore, A from Angus. Male is listed first and female second. For example, NA refers to a homozygous animal that has inherited the Nellore haplotype paternally and the Angus haplotype maternally.

^aMeans within a row with different superscripts differ significantly ($p < 0.05$).

^bMeans within a row with different superscripts differ significantly ($p < 0.05$).

groups. Relative to the efficient group, the inefficient group overexpressed *ACTN3* 1.6-fold. The fiber type ratio measured by fast-twitch (type II)/slow-twitch (type I) differed between groups, with a 1.8-fold increase in the inefficient group relative to the efficient group. Season of weaning may have influenced the results since steers weaned in the fall also had a greater fast-twitch to slow-twitch fiber type ratio and greater expression of *ACTN3* compared with those weaned during spring. However, the model predicted RFI_{NRC} does consider birth year season. Feed intake and efficiency have been previously shown to be affected by seasonal effects (Mujibi et al., 2010).

Although not all families are represented in all contemporary groups for which feed efficiency data were collected (Table 4), sires are much more evenly distributed across these groups. All sires had at least three steers per contemporary group except for 432H in Fall 2004 ($n = 0$) and Spring 2005 ($n = 2$) and 437J in Fall 2003 ($n = 1$). Because RFI_{NRC} is a population-based calculation rather than a cohort-based calculation, its use may produce unequal numbers of efficient/inefficient observations per contemporary group. Although only 1 or 0 animals are evaluated from a single contemporary group, this should not bias the results compared to if all animals were included in the analyses.

Our use of RFI_{NRC} enabled animals to be compared across BYS groups. However, as evidenced by the imbalance of animals deemed “most efficient” across BYS, we are likely observing the impact of G X E interaction on the efficiency phenotype. However, despite a potential environmental influence that also affected efficiency, we identified the most efficient animals across all BYS. We are interested in understanding genes and gene networks that contribute to multigenic traits such as feed efficiency in the context of variably subtropical conditions that are typically experienced in beef cattle-producing regions such as Texas.

As *ACTN3* is expressed only in type II fibers, the fiber type data are consistent with our expression data. The inefficient group also had a larger standard error than the efficient group. A possible explanation for this variability in expression could be that an

animal may be inefficient, or just simply average, due to a wide array of combined genotypic, environmental, and other factors. In contrast, an efficient animal might be expected to possess only specific genotype combinations that, in conjunction with certain environmental conditions, return an efficient phenotype. Fiber type ratio is also just one variable among many that could possibly reduce overall efficiency.

A role for *ACTN3* variation in growth and metabolism is supported by studies in other species, including humans. A relationship between *ACTN3* expression and metabolic rates has been reported in rodents (MacArthur et al., 2007; MacArthur et al., 2008; Ogura et al., 2009) and humans (North and Beggs 1996; North et al., 2003; Yang et al., 2003). Additionally, the relationship between the *ACTN3* R577X (loss of function) polymorphism and elite athletic performance has been described in humans (Roth et al., 2008; Ahmetov et al., 2010; Fiuza-Luces et al., 2011). Interestingly, loss of function has been shown to be associated with the reduced cross-sectional area and thigh muscle volume in older women (Min et al., 2016; Kiuchi et al., 2021).

Among all steers with records, across all seasons in the study ($n = 173$), from which a subset was used for gene expression assays, parent and breed of origin of the *ACTN3* haplotype block had a significant impact on efficiency as measured by RFI residuals. The Nellore (*Bos indicus*) haplotype block was associated with greater efficiency when inherited maternally (RFI residual -0.47 for maternal Nellore inheritance compared to 0.49) regardless of the breed of origin of the paternal haplotype block. If efficiency is represented in terms of feed efficiency during a 180-day feedlot finishing period in kg/d, at current US prices of \$0.352/kg, this difference will result in about \$63 savings per animal for the most efficient animals. Additionally, the Nellore haplotype blocks of *CAPN2* and *FN1* were associated with differences in efficiency when inherited maternally. For *CAPN2*, a maternally inherited Nellore haplotype was linked to improved efficiency (RFI residual -0.50 compared to 0.50). The *FN1* Angus haplotype block was associated with improved efficiency relative to the Nellore (-0.40 compared to 0.50). No paternal haplotype blocks were implicated to play a role in this trait. These results suggest that parent of origin

effects or other epigenetic effects may play an important role in the efficiency phenotype and should be investigated further to optimize management practices. Reciprocal differences in birthweight between *B. indicus* X *B. taurus* and *B. taurus* X *B. indicus* offspring are well known (Dillon et al., 2015). Epigenetic differences in regulation of skeletal muscle growth have also been described, notably in callipyge sheep, where certain transcripts detected in skeletal muscle are transcribed from only a single parental allele (Bidwell et al., 2004). Although feed efficiency data are available only for the F₂ steers utilized in this study and not the parent/grandparent generations nor sibling heifers, these findings are relevant for beef cattle production strategies in subtropical regions. It will be necessary to understand the impact of these same regions on heifers, replacement cows, and future sires. Additionally, these data begin to contribute to the understanding of parent-of-origin effects on the inheritance of chromosomal regions—particularly important for regions where *Bos taurus*–*Bos indicus* cross cattle are utilized.

Regarding fiber types, fast- and slow-twitch muscle fibers rely primarily on different metabolic pathways. Fast-twitch fibers (type II) rely on the anaerobic degradation of glucose for energy (glycolysis). In contrast, slow fibers (type I) utilize the aerobic citric acid cycle as a primary energy source. Due to the differences in metabolic pathways utilized, a shift in one direction can lead to overall differences in energy utilization during the lifespan of the animal. A reduction in *ACTN3* can result in differences in glycogen phosphorylase activity in muscle and changes in calcium metabolism (MacArthur et al., 2007; Quinlan et al., 2009; Quinlan et al., 2010). Mice entirely deficient in *Actn3* show an increase in expression of enzymes relating to the glycolytic pathway and a decrease in expression of enzymes of the aerobic cellular respiration pathway (MacArthur et al., 2008). Aerobic cellular respiration in total produces 38 molecules of ATP per molecule of glucose input compared to only two molecules of ATP produced by glycolysis per molecule of glucose, making aerobic metabolism 19 times more efficient than glycolysis. Therefore, any shift towards one over the other may affect the overall efficiency of energy metabolism. A greater abundance of *ACTN3* and Type II fibers may contribute mechanistically to seasonal differences.

Cattle facing insufficient nutrition degrade fast-twitch muscle fibers initially to preserve slow-twitch muscle fibers (Lehnert et al., 2006), suggesting a preference for higher efficiency muscle under periods of nutritional stress. To our knowledge, a role for *ACTN3* expression specifically influencing bovine metabolic efficiency has not been previously described. However, Nolte et al., 2019 examined biological networks to identify key long non-coding RNAs (lncRNA) associated with bovine metabolic efficiency. Of three key lncRNAs, one lncRNA with connectivity to low metabolic efficiency (MSTRG.10337) and expressed in muscle tissue was correlated with calcium signaling and other skeletal muscle genes, including *ACTN3*. It is unclear what role *CAPN2* and *FN1* might play in affecting efficiency, but these results are also consistent with the Nolte study of a potential role for calcium and cytoskeletal signaling. Because both *CAPN2* and *FN1* are associated with muscle turnover and growth, it is possible they are linked to differences in the rates of these physiological functions, which would, in turn, alter the

metabolic rate of the animal. Interestingly, Karisa et al. (2013) demonstrated the association of *CAST* genes with efficiency. The animals in their study were crossbred *Bos taurus* breeds reared in Canada. Potentially, the favorable *Bos indicus* *ACNT3*, *CAPN2*, and *FN1* haplotypes reflect the more favorable suitability of these breeds for subtropical environments.

In this experiment, we demonstrated that, in skeletal muscle of Nellore-Angus F₂ steers, samples from animals classified as feed efficient (relative to RFI_{NRC}) contained fewer *ACTN3* transcripts in comparison to the inefficient group. Similarly, muscle samples from the efficient group possessed fewer type II muscle fibers than the inefficient group. Maternal inheritance of the Nellore *Bos indicus* *ACTN3* haplotype block conferred the greatest improvement in efficiency. Two additional genes involved in skeletal muscle turnover, *CAPN2* and *FN1*, were also associated with improved efficiency when inherited as maternal *Bos indicus* haplotype blocks. Collectively these data demonstrate that expression of *ACTN3*, the gene network in which it resides, and its regulation, may represent a novel target for improving feed efficiency in cattle, particularly in subtropical environments. While *ACTN3* may reflect just one element of a trait, these data, in connection with other studies, help provide a base for a better understanding of complex biological mechanisms and their interaction with fluctuating climatic conditions.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repository and accession number(s) can be found at: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56705>, GSE56705.

ETHICS STATEMENT

All procedures involving animals were approved by the Texas A&M AgriLife Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

PR and RV designed the study. CG, DR, AH, and JS provided animal genotype and phenotype data and designed the genetic mapping herd structure. RV, KK, and AT prepared samples and conducted gene expression analyses. MD conducted fiber type analysis. RV conducted data analysis. PR, CG, and DR designed and supervised experiments and data analysis. RV and PR wrote the manuscript with input from all authors.

FUNDING

Financial support for this project was provided, in part, by the Beef Competitiveness Exceptional Item funding from the Texas legislature. Genotypes utilized for analyses were produced with support from The Beef Checkoff and National Research Initiative

Competitive Grant no. 2008-35205-18767 from the USDA National Institute of Food and Agriculture Animal Genome Program. RV was partially supported by the Whole Systems Genomics Initiative for Human, Animal, and Environmental Wellbeing at Texas A&M University.

ACKNOWLEDGMENTS

The authors gratefully acknowledge numerous undergraduate and graduate students, staff, and faculty colleagues at Texas

A&M who contributed to sample and data collection for this research project. This work was included as a portion of the doctoral dissertation of RV, in partial fulfillment of the degree requirements.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.796038/full#supplementary-material>

REFERENCES

- Ahmetov, II, Druzhevskaya, A. M., Astratenkova, I. V., Popov, D. V., Vinogradova, O. L., and Rogozkin, V. A. (2010). The ACTN3 R577X Polymorphism in Russian Endurance Athletes. *Br. J. Sports Med.* 44, 649–652. doi:10.1136/bjsm.2008.051540
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/s0022-2836(05)80360-2
- Amen, T. S. (2007). Feed Efficiency, Carcass and Temperament Traits in F₂ Nellore-Angus Steers. Doctoral Dissertation. College Station (TX): Texas A&M University.
- Bamman, M. M., Clarke, M. S. F., Talmadge, R. J., and Feeback, D. L. (1999). Enhanced Protein Electrophoresis Technique for Separating Human Skeletal Muscle Myosin Heavy Chain Isoforms. *Electrophoresis* 20, 466–468. doi:10.1002/(sici)1522-2683(19990301)20:3<466aid-elps466>3.0.co;2-7
- Becker, G. S. 2008. Livestock Feed Costs: Concerns and Options. CRS Report for Congress. Available at: <http://congressionalresearch.com/RS22908/document.php?studyDLivestockCFeedCCostsCConcernsCandCOptions> (Accessed September 17, 2008)
- Bidwell, C. A., Kramer, L. N., Perkins, A. C., Hadfield, T. S., Moody, D. E., and Cockett, N. E. (2004). Expression of PEG11 and PEG11AS Transcripts in Normal and Callipyge Sheep. *BMC Biol.* 2, 17. doi:10.1186/1741-7007-2-17
- Brito, L. F., Bedere, N., Douhard, F., Oliveira, H. R., Arnal, M., Peñagaricano, F., et al. (2021). Genetic Selection of High-Yielding Dairy Cattle Toward Sustainable Farming Systems in a Rapidly Changing World. *Animal* 15, 100292. doi:10.1016/j.animal.2021.100292
- Brito, L. F., Oliveira, H. R., Houlahan, K., Fonseca, P. A. S., Lam, S., Butty, A. M., et al. (2020). Genetic Mechanisms Underlying Feed Utilization and Implementation of Genomic Selection for Improved Feed Efficiency in Dairy Cattle. *Can. J. Anim. Sci.* 100, 587–604. doi:10.1139/cjas-2019-0193
- Brunes, L. C., Baldi, F., Lopes, F. B., Lôbo, R. B., Espigolan, R., Costa, M. F. O., et al. (2021). Weighted Single-step Genome-wide Association Study and Pathway Analyses for Feed Efficiency Traits in Nellore Cattle. *J. Anim. Breed. Genet.* 138, 23–44. doi:10.1111/jbg.12496
- Calender, A., Weichhart, T., Valeyre, D., and Pacheco, Y. (2020). Current Insights in Genetics of Sarcoidosis: Functional and Clinical Impacts. *Jcm* 9, 2633. doi:10.3390/jcm9082633
- Capper, J. L. (2011). The Environmental Impact of Beef Production in the United States: 1977 Compared with 2007. *J. Anim. Sci.* 89, 4249–4261. doi:10.2527/jas.2010-3784
- Cruzen, S. M., Harris, A. J., Hollinger, K., Punt, R. M., Grubbs, J. K., Selsby, J. T., et al. (2013). Evidence of Decreased Muscle Protein Turnover in Gilts Selected for Low Residual Feed Intake. *J. Anim. Sci.* 91, 4007–4016. doi:10.2527/jas.2013-6413
- De Jonge, H. J. M., Fehrmann, R. S. N., de Bont, E. S. J. M., Hofstra, R. M. W., Gerbens, F., Kamps, W. A., et al. (2007). Evidence Based Selection of Housekeeping Genes. *PLoS One* 2, e898. doi:10.1371/journal.pone.0000898
- Dillon, J. A., Riley, D. G., Herring, A. D., Sanders, J. O., and Thallman, R. M. (2015). Genetic Effects on Birth Weight in Reciprocal Brahman-Simmental Crossbred Calves. *J. Anim. Sci.* 93, 553–561. doi:10.2527/jas.2014-8525
- Egan, B., and Zierath, J. R. (2013). Exercise Metabolism and the Molecular Regulation of Skeletal Muscle Adaptation. *Cel Metab.* 17, 162–184. doi:10.1016/j.cmet.2012.12.012
- FAO (2018). “Shaping the Future of Livestock,” in 10th Global Forum for Food and Agriculture (GFFA), Berlin, January 18–20, (Food and Agriculture Organization of the United Nations). Available at: <https://www.fao.org/publications/card/en/c/18384EN/>.
- Fiuzza-Luces, C., Ruiz, J. R., Rodríguez-Romo, G., Santiago, C., Gómez-Gallego, F., Yvert, T., et al. (2011). Are ‘Endurance’ Alleles ‘Survival’ Alleles? Insights from the ACTN3 R577X Polymorphism. *PLoS One* 6, e17558. doi:10.1371/journal.pone.0017558
- Goll, D. E., Thompson, V. F., Li, H., Wei, W., and Cong, J. (2003). The Calpain System. *Physiol. Rev.* 83, 731–801. doi:10.1152/physrev.00029.2002
- Greenbaum, D., Jansen, R., and Gerstein, M. (2002). Analysis of mRNA Expression and Protein Abundance Data: An Approach for the Comparison of the Enrichment of Features in the Cellular Population of Proteins and Transcripts. *Bioinformatics* 18, 585–596. doi:10.1093/bioinformatics/18.4.585
- Hiller, B., Hocquette, J.-F., Cassar-Malek, I., Nuernberg, G., and Nuernberg, K. (2012). Dietary N-3 PUFA Affect Lipid Metabolism and Tissue Function-Related Genes in Bovine Muscle. *Br. J. Nutr.* 108, 858–863. doi:10.1017/s0007114511006179
- Ho, C. K. M., Malcolm, B., and Doyle, P. T. (2013). Potential Impacts of Negative Associative Effects between Concentrate Supplements, Pasture and Conserved Forage for Milk Production and Dairy Farm Profit. *Anim. Prod. Sci.* 53, 437–452. doi:10.1071/an12140
- Karisa, B. K., Thomson, J., Wang, Z., Stothard, P., Moore, S. S., and Plastow, G. S. (2013). Candidate Genes and Single Nucleotide Polymorphisms Associated with Variation in Residual Feed Intake in Beef Cattle. *J. Anim. Sci.* 91, 3502–3513. doi:10.2527/jas.2012-6170
- Kenny, D. A., Fitzsimons, C., Waters, S. M., and McGee, M. (2018). Invited Review: Improving Feed Efficiency of Beef Cattle - The Current State of the Art and Future Challenges. *Animal* 12, 1815–1826. doi:10.1017/s1751731118000976
- Kiuchi, Y., Makizako, H., Nakai, Y., Taniguchi, Y., Tomioka, K., Sato, N., et al. (2021). Associations of Alpha-Actinin-3 Genotype with Thigh Muscle Volume and Physical Performance in Older Adults with Sarcopenia or Pre-sarcopenia. *Exp. Gerontol.* 154, 111525. doi:10.1016/j.exger.2021.111525
- Kochan, K. J., Vaughn, R. N., Amen, T. S., Abbey, C. A., Sanders, J. O., Lunt, D. K., et al. (2009). Expression of Mitochondrial Respiratory Complex Genes in Liver Tissue of Cattle with Different Feed Efficiency Phenotypes. *Proc. Assoc. Advmt Anim. Breed. Genet.* 18, 175–178.
- Lam, S., Zeidan, J., Miglior, F., Suárez-Vega, A., Gómez-Redondo, I., Fonseca, P. A. S., et al. (2020). Development and Comparison of RNA-Sequencing Pipelines for More Accurate SNP Identification: Practical Example of Functional SNP Detection Associated with Feed Efficiency in Nellore Beef Cattle. *BMC Genomics* 21, 703. doi:10.1186/s12864-020-07107-7
- Lehnert, S. A., Byrne, K. A., Reverter, A., Nattrass, G. S., Greenwood, P. L., Wang, Y. H., et al. (2006). Gene Expression Profiling of Bovine Skeletal Muscle in Response to and During Recovery from Chronic and Severe Undernutrition. *J. Anim. Sci.* 84, 3239–3250. doi:10.2527/jas.2006-192
- Lek, M., Quinlan, K. G. R., and North, K. N. (2010). The Evolution of Skeletal Muscle Performance: Gene Duplication and Divergence of Human Sarcomeric β -Actinins. *Bioessays* 32, 17–25. doi:10.1002/bies.200900110
- Li, A., Mo, D., Zhao, X., Jiang, W., Cong, P., He, Z., et al. (2013). Comparison of the *Longissimus* Muscle Proteome Between Obese and Lean Pigs at 180 Days. *Mamm. Genome* 24, 72–79. doi:10.1007/s00335-012-9440-0

- Liu, M. F., Goonewardene, L. A., Bailey, D. R. C., Basarab, J. A., Kemp, R. A., Arthur, P. F., et al. (2000). A Study on the Variation of Feed Efficiency in Station Tested Beef Bulls. *Can. J. Anim. Sci.* 80, 435–441. doi:10.4141/a99-030
- MacArthur, D. G., Seto, J. T., Chan, S., Quinlan, K. G. R., Raftery, J. M., Turner, N., et al. (2008). An Actn3 Knockout Mouse Provides Mechanistic Insights into the Association between α -actinin-3 Deficiency and Human Athletic Performance. *Hum. Mol. Genet.* 17, 1076–1086. doi:10.1093/hmg/ddm380
- MacArthur, D. G., Seto, J. T., Raftery, J. M., Quinlan, K. G., Huttley, G. A., Hook, J. W., et al. (2007). Loss of ACTN3 Gene Function Alters Mouse Muscle Metabolism and Shows Evidence of Positive Selection in Humans. *Nat. Genet.* 39, 1261–1265. doi:10.1038/ng2122
- McClelland, S. C., Arndt, C., Gordon, D. R., and Thoma, G. (2018). Type and Number of Environmental Impact Categories Used in Livestock Life Cycle Assessment: A Systematic Review. *Livestock Sci.* 209, 39–45. doi:10.1016/j.livsci.2018.01.008
- McDonagh, M. B., Herd, R. M., Richardson, E. C., Oddy, V. H., Archer, J. A., and Arthur, P. F. (2001). Meat Quality and the Calpain System of Feedlot Steers Following a Single Generation of Divergent Selection for Residual Feed Intake. *Aust. J. Exp. Agric.* 41, 1013–1021. doi:10.1071/ea00024
- Mills, M., Yang, N., Weinberger, R., Vander Woude, D. L., Beggs, A. H., Easteal, S., et al. (2001). Differential Expression of the Actin-Binding Proteins, Alpha-Actinin-2 and -3, in Different Species: Implications for the Evolution of Functional Redundancy. *Hum. Mol. Genet.* 10, 1335–1346. doi:10.1093/hmg/10.13.1335
- Min, S.-K., Lim, S.-T., and Kim, C.-S. (2016). Association of ACTN3 Polymorphisms with BMD, and Physical Fitness of Elderly Women. *J. Phys. Ther. Sci.* 28, 2731–2736. doi:10.1589/jpts.28.2731
- Mujibi, F. D. N., Moore, S. S., Nkrumah, D. J., Wang, Z., and Basarab, J. A. (2010). Season of Testing and its Effect on Feed Intake and Efficiency in Growing Beef Cattle. *J. Anim. Sci.* 88, 3789–3799. doi:10.2527/jas.2009-2407
- Nolte, W., Weikard, R., Brunner, R. M., Albrecht, E., Hammon, H. M., Reverter, A., et al. (2019). Biological Network Approach for the Identification of Regulatory Long Non-coding RNAs Associated with Metabolic Efficiency in Cattle. *Front. Genet.* 10, 1130. doi:10.3389/fgenet.2019.01130
- North, K. N., Yang, N., Macarthur, D., Gulbin, J., and Easteal, S. (2003). “A Common Polymorphism in the Skeletal Muscle Gene ACTN3 Influences Human Athletic Performance,” in Meeting 8th International Congress of the World-Muscle-Society, Sneged, Hungary, September 3–6, 2003 (Sponsor World Muscle Soc).
- North, K. N., and Beggs, A. H. (1996). Deficiency of a Skeletal Muscle Isoform of α -actinin (α -Actinin-3) in Merosin-Positive Congenital Muscular Dystrophy. *Neuromuscul. Disord.* 6, 229–235. doi:10.1016/0960-8966(96)00361-6
- NRC (2000). *Nutrient Requirements of Beef Cattle*. 7th Revised Edition. Washington, DC: The National Academies Press.
- Ogura, Y., Naito, H., Kakigi, R., Akema, T., Sugiura, T., Katamoto, S., et al. (2009). Different Adaptations of Alpha-Actinin Isoforms to Exercise Training in Rat Skeletal Muscles. *Acta Physiol.* 196, 341–349. doi:10.1111/j.1748-1716.2008.01945.x
- Ortigue, I. (1991). Adaptation du métabolisme énergétique des ruminants à la sous-alimentation. Quantification au niveau de l'animal entier et de tissus corporels. *Reprod. Nutr. Dev.* 31, 593–616. doi:10.1051/rnd:19910601
- Puillet, L., Ducrocq, V., Friggens, N. C., and Amer, P. R. (2021). Exploring Underlying Drivers of Genotype by Environment Interactions in Feed Efficiency Traits for Dairy Cattle with a Mechanistic Model Involving Energy Acquisition and Allocation. *J. Dairy Sci.* 104, 5805–5816. doi:10.3168/jds.2020-19610
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795
- Quinlan, K. G. R., Seto, J. T., Turner, N., Floetenmeyer, M., Macarthur, D. G., Raftery, J. M., et al. (2009). G.O.4 α -Actinin-3 Regulates Muscle Glycogen Phosphorylase: A Potential Mechanism for the Metabolic Consequences of the Common Human Null Allele of ACTN3. *Neuromuscul. Disord.* 19, 545–546. doi:10.1016/j.nmd.2009.06.011
- Quinlan, K. G. R., Seto, J. T., Turner, N., Vandebrouck, A., Floetenmeyer, M., Macarthur, D. G., et al. (2010). α -Actinin-3 Deficiency Results in Reduced Glycogen Phosphorylase Activity and Altered Calcium Handling in Skeletal Muscle. *Hum. Mol. Genet.* 19, 1335–1346. doi:10.1093/hmg/ddq010
- Rajesh, R. V., Jang, E. J., Choi, I., Heo, K.-N., Yoon, D., Kim, T.-H., et al. (2011). Proteomic Analysis of Bovine Muscle Satellite Cells During Myogenic Differentiation. *Asian Australas. J. Anim. Sci.* 24, 1288–1302. doi:10.5713/ajas.2011.10344
- Riggs, P. K., Tedeschi, L. O., Turner, N. D., Braga-Neto, U., and Jayaraman, A. (2017). The Role of 'omics Technologies for Livestock Sustainability. *Archivos Latinoamericanos de Produccion Anim.* 25, 149–155.
- Riggs, P. K., and Vaughn, R. N. (2015). *DNA Markers for Feed Efficiency in Cattle*. United States Letters Patent, Application No. 14/724,696.
- Riggs, P. K., Fields, M. J., and Cross, H. R. (2018). Food and Nutrient Security for a Growing Population. *Anim. Front.* 8, 3–4. doi:10.1093/af/vfy014
- Ropka-Molik, K., Stefaniuk-Szmukier, M., Musiał, A. D., Piórkowska, K., and Szmatoła, T. (2019). Sequence Analysis and Expression Profiling of the Equine ACTN3 Gene During Exercise in Arabian Horses. *Gene* 685, 149–155. doi:10.1016/j.gene.2018.10.079
- Roth, S. M., Walsh, S., Liu, D., Metter, E. J., Ferrucci, L., and Hurley, B. F. (2008). The ACTN3 R577X Nonsense Allele Is Under-represented in Elite-Level Strength Athletes. *Eur. J. Hum. Genet.* 16, 391–394. doi:10.1038/sj.ejhg.5201964
- Savell, J. W., and Smith, G. C. (2000). *Meat Science Laboratory Manual*. 7th Edn. Boston, MA: American Press.
- Stephens, M., Smith, N. J., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *Am. J. Hum. Genet.* 68, 978–989. doi:10.1086/319501
- Tizioto, P. C., Coutinho, L. L., Oliveira, P. S. N., Cesar, A. S. M., Diniz, W. J. S., Lima, A. O., et al. (2016). Gene Expression Differences in Longissimus Muscle of Nelore Steers Genetically Divergent for Residual Feed Intake. *Sci. Rep.* 6, 39493. doi:10.1038/srep39493
- Underwood, K. R., Tong, J., Zhu, M. J., Shen, Q. W., Means, W. J., Ford, S. P., et al. (2007). Relationship Between Kinase Phosphorylation, Muscle Fiber Typing, and Glycogen Accumulation in Longissimus Muscle of Beef Cattle with High and Low Intramuscular Fat. *J. Agric. Food Chem.* 55, 9698–9703. doi:10.1021/jf071573z
- Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes. *Genome Biol.* 3, RESEARCH0034. doi:10.1186/gb-2002-3-7-research0034
- Vaughn, R. N. (2013). Genetic Mechanisms that Contribute to Differences in Beef Tenderness Following Electrical Stimulation. Doctoral Dissertation. College Station (TX): Texas A&M University.
- Wang, M., Yu, H., Kim, Y. S., Bidwell, C. A., and Kuang, S. (2012). Myostatin Facilitates Slow and Inhibits Fast Myosin Heavy Chain Expression During Myogenic Differentiation. *Biochem. Biophysical Res. Commun.* 426, 83–88. doi:10.1016/j.bbrc.2012.08.040
- Welch, C. M., Thornton, K. J., Murdoch, G. K., Chapalamadugu, K. C., Schneider, C. S., Ahola, J. K., et al. (2013). An Examination of the Association of Serum IGF-I Concentration, Potential Candidate Genes, and Fiber Type Composition with Variation in Residual Feed Intake in Progeny of Red Angus Sires Divergent for Maintenance Energy EPD1. *J. Anim. Sci.* 91, 5626–5636. doi:10.2527/jas.2013-6609
- Yang, N., MacArthur, D. G., Gulbin, J. P., Hahn, A. G., Beggs, A. H., Easteal, S., et al. (2003). ACTN3 Genotype Is Associated with Human Elite Athletic Performance. *Am. J. Hum. Genet.* 73, 627–631. doi:10.1086/377590

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Vaughn, Kochan, Torres, Du, Riley, Gill, Herring, Sanders and Riggs. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome Wide Scan to Identify Potential Genomic Regions Associated With Milk Protein and Minerals in Vrindavani Cattle

Akansha Singh¹, Amit Kumar^{1*}, Cedric Gondro², A. K. Pandey¹, Triveni Dutt³ and B. P. Mishra⁴

¹ Animal Genetics Division, Indian Council of Agricultural Research-Indian Veterinary Research Institute, Bareilly, India,

² Department of Animal Science, Michigan State University, East Lansing, MI, United States, ³ Livestock Production and

Management, Indian Council of Agricultural Research-Indian Veterinary Research Institute, Bareilly, India, ⁴ Division of Animal Biotechnology, Indian Council of Agricultural Research-Indian Veterinary Research Institute, Bareilly, India

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

George R. Wiggins,
Council on Dairy Cattle Breeding,
United States

Tara G. McDanel,
U.S. Meat Animal Research Center
(USDA), United States

*Correspondence:

Amit Kumar
vetamitchandan07@gmail.com

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Veterinary Science

Received: 18 August 2021

Accepted: 11 February 2022

Published: 10 March 2022

Citation:

Singh A, Kumar A, Gondro C,
Pandey AK, Dutt T and Mishra BP
(2022) Genome Wide Scan to Identify
Potential Genomic Regions
Associated With Milk Protein and
Minerals in Vrindavani Cattle.
Front. Vet. Sci. 9:760364.
doi: 10.3389/fvets.2022.760364

In this study, genome-wide association study (GWAS) was conducted for identifying significantly associated genomic regions/SNPs with milk protein and minerals in the 96 taurine-indicine crossbred (*Vrindavani*) cows using 50K SNP Chip. After quality control, a total of 41,427 SNPs were retained and were further analyzed using a single-SNP additive linear model. Lactation stage, parity, test day milk yield and proportion of exotic inheritance were included as fixed effects in GWAS model. Across all traits, 13 genome-wide significant ($p < 1.20 \times 10^{-6}$) and 49 suggestive significant ($p < 2.41 \times 10^{-5}$) SNPs were identified which were located on 18 different autosomes. The strongest association for protein percentage, calcium (Ca), phosphorus (P), copper (Cu), zinc (Zn), and iron (Fe) were found on BTA 18, 7, 2, 3, 14, and 2, respectively. No significant SNP was detected for manganese (Mn). Several significant SNPs identified were within or close proximity to *CDH13*, *BHLHE40*, *EDIL3*, *HAPLN1*, *INHBB*, *USP24*, *ZFAT*, and *IKZF2* gene, respectively. Enrichment analysis of the identified candidate genes elucidated biological processes, cellular components, and molecular functions involved in metal ion binding, ion transportation, transmembrane protein, and signaling pathways. This study provided a groundwork to characterize the molecular mechanism for the phenotypic variation in milk protein percentage and minerals in crossbred cattle. Further work is required on a larger sample size with fine mapping of identified QTL to validate potential candidate regions.

Keywords: crossbred cattle, genome-wide association studies, milk minerals, milk protein, Vrindavani

INTRODUCTION

Milk and dairy products are considered a nutrient-rich diet, consisting of fat, protein, lactose along with several essential micronutrients including vitamins and minerals. Minerals representing a small fraction (~9 g/L) of bovine milk, are important for nutritional and technological properties of milk (1).

The minerals in milk are present either as an inorganic ion (soluble phase) or form colloidal complexes with organic matter such as proteins, carbohydrates, and ligands including amino acids and citrates (2, 3).

Minerals play an important role in several physiological and biological activities. For instance, calcium, phosphorus, and magnesium are involved in the development of connective tissues such as bones, muscles, cartilages, and teeth in humans (4). Calcium has been reported to play an important role in reducing cholesterol absorption, and in regulating blood pressure in humans (5). The zinc, manganese, iron, and copper are important components of several enzymes and play role in the immune system (6). These elements also serve as catalysts for many biochemical processes such as muscle contraction, nervous transmission, and nutrient absorption (7).

The concentration of milk minerals influences the stabilization of casein micelle in milk. Calcium, phosphorus, and magnesium are an important components of casein micelle and positively regulates the coagulation properties of the milk (8). The milk minerals are complex traits, influenced by both genetic and non-genetic factors, including nutrition, lactation stage, parity, season, and breed (9). Studies have shown that milk minerals have low to moderate heritability for Cu, Zn, Fe, and Mn (10); and moderate to high heritability for Ca, P, and Zn (1). This provides the possibility to alter the concentration of bovine milk minerals through selective breeding.

The genome-wide association studies may improve our understanding of the genetic architecture and underlying molecular mechanism for variation in protein and minerals content in bovine milk. Studies have revealed about several potential candidate genes associated with milk protein percentage in different cattle breeds (11, 12). However, for milk minerals only a limited number of genome-wide association studies has been reported (1, 3). Thus, the aim of this study was to identify genomic regions associated with milk protein percentage (%) and milk minerals including Ca, P, Cu, Zn, Fe, and Mn in Vrindavani cattle, a tropically adapted composite crossbred breed of dairy cattle of India.

MATERIALS AND METHODS

Animals and Phenotypes

Morning milk samples (50 ml) were collected from 96 crossbred Vrindavani cattle of the ICAR-Indian Veterinary Research Institute, Bareilly. Vrindavani is a composite breed developed at Indian Veterinary Research Institute by crossing indicine (Hariana) with three taurine breeds (Holstein, Brown Swiss, and Jersey). Details of population structure and synthetic breed information of the Vrindavani population were described in a prior study by (13). The cows were kept in a loose housing system with free-stall dairy barn, fed *ad libitum* with mixed ration, and milked in a free-stall barn using automated milking systems. The cows were in milk (24–403 days) and from parity 1–8.

The protein percentage in milk was determined using a LactoScan milk analyzer (10). The milk minerals (Ca, P, Cu, Zn, Fe, and Mn) were estimated using Flame Atomic

Absorption Spectrometer in technical triplicate samples. For the mineralization of milk samples, 5 ml of milk were dried in silica crucible using a hot-air oven. The dried samples were ashed using a muffle furnace. The ashed sample was mixed with 15 ml 1:1 diluted HCl and heated to dissolve the acid-soluble portion of total ash. The soluble portion was filtered and diluted to 100 ml using double distilled water. The digested samples were analyzed, using a flame atomic absorption spectrophotometer (14).

Genotyping and Quality Control

Blood samples were collected from individual cows, with the approval from the Institutional Animal Ethics Committee (IAEC) on ICAR-Indian Veterinary Research Institute, Bareilly. Genomic DNA was isolated using Qiagen DNeasy Blood Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The quality and quantity of DNA were evaluated using NanoDrop spectrophotometer, agarose gel electrophoresis and Qubit fluorometer. The extracted genomic DNA were genotyped using Illumina BovineSNP50 Bead Chip platform (Illumina Inc., San Diego, CA) consisting of 53,218 SNPs across the genome. The quality control of the SNP genotypes was performed using PLINK v 1.9 (15). SNPs with call rate < 90%, minor allele frequency (MAF) < 0.05 and significantly deviating from the Hardy-Weinberg equilibrium were excluded from the analysis. A total of 41,427 SNPs mapping autosomes were retained for further downstream analysis.

Statistical Analyses

Prior to GWAS, the admixture profile and the level of exotic inheritance of individual animal was estimated using ADMIXTURE V.13.0 (16) using a reference panel of purebred taurine (Holstein-Friesian, Brown Swiss, and Jersey) and indicine (Hariana) cattle downloaded from WIDDE ([http://widde.toulouse.inra.fr/widde/widde/main.do?module\\$=cattle](http://widde.toulouse.inra.fr/widde/widde/main.do?module$=cattle)) and KRISHI (<https://krishi.icar.gov.in/jspui/handle/123456789/31167>) web portals. The fixed effect of lactation stage, parity, test day milk yield, and the proportion of exotic inheritance (percentage of Holstein-Friesian, Jersey and Brown Swiss, estimated by admixture analysis) were tested with *lm* function in R. A single-SNP linear regression of phenotype on genotype was fitted for GWAS analysis using snpStats package in R following Gondro (17). The genotypes were tested for additive effect using model.

$$y = X\beta + Z\alpha + e$$

Where *y* is the vector of phenotypic observations; *X* is an incidence matrix relating the phenotypes to the fixed effects including lactation stage, parity, test day milk yield, and the level of exotic inheritance computed by admixture analysis; β is the vector of fixed effects; *Z* is incidence matrix of genotypes (0 for the first homozygote AA; 1 for the heterozygote AB or BA; 2 for the second homozygote BB) of the fitted SNP; α is the vector of effects of the regression coefficient for SNPs, and *e* is the vector of residual effects with a normal distribution $N \sim (0, I\sigma_e^2)$, where σ_e^2 is the residual variance.

The Bonferroni correction using $0.05/N$, where “N” is the number of SNPs, was applied to the genome-wide significance threshold (18). The SNP effects were declared significant at a genome-wide level of $P = 1.20 \times 10^{-06}$ ($0.05/41,427$). Since Bonferroni correction was stringent, a suggestive significance threshold of $P = 2.41 \times 10^{-05}$ ($1/41,427$) was calculated (19). Association between individual SNPs and each trait are shown in the Manhattan plot using R.

The Ensembl database UMDv3.1 (https://www.ensembl.org/Bos_taurus/Info/Annotation) was used to search for genes flanking a region of 500 kb from the genome-wide significant and suggestive SNPs using BEDtools (20). Candidate genes close or containing significant SNPs were analyzed for functional enrichment using DAVID 6.8 (21). The significantly enriched pathways were identified based on enriched scores.

RESULTS AND DISCUSSION

The descriptive statistics for protein percentage and 6 individual minerals (Ca, P, Cu, Mn, Zn, and Fe) in the milk of Vrindavani cattle are presented in **Table 1**. A total of 9 (2 genome-wide and 7 suggestive) significant SNPs for protein percentage; and 53 (11 genome-wide and 42 suggestive) significant SNPs for six different mineral traits were identified.

Protein Percentage

For protein percentage, two genome-wide significant SNPs ($p < 1.20 \times 10^{-06}$) and 7 suggestive significant SNPs ($p < 2.41 \times 10^{-05}$) were detected on BTA 3, BTA 6, BTA18, and BTA22 (**Figure 1**; **Table 2**). Of these, four significant SNPs found on BTA18 (7.63–99.15Mb) includes Beta-carotene oxygenase 1 (*BCO1*), C-Maf inducing protein (*CMIP*), N-terminal EF-hand calcium binding protein 2 (*NECAB2*), Cadherin 13 (*CDH13*), oxidative stress induced growth inhibitor 1 (*OSGIN1*) and Solute carrier family 38 member 8 (*SLC38A8*) genes.

Minerals

Among minerals, a total of 4, 17, 10, 7, and 15 significant SNPs were detected for Ca, P, Cu, Zn, and Fe, respectively. However, no significant SNP was detected for Mn (**Figures 2A–F**; **Table 3**).

A total three genome-wide significant and 1 suggestive significant SNPs on BTA7 (86.04–89.65Mb) were found to be

associated with Ca. The region contains several genes including Hyaluronan and proteoglycan link protein 1 (*HAPLN1*), Versican (*VCAN*), EGF like repeats, and discoidin domains 3 (*EDIL3*). A total of three genome-wide and 14 suggestive significant SNPs were found to be associated with P. The most significantly ($p = 6.89 \times 10^{-09}$) associated SNP (Hapmap45428-BTA-47987) was located 0.10 Mb downstream from Inhibin subunit beta B (*INHBB*) gene and 0.81 Mb downstream from Cilia and flagella associated protein 221 (*CFAP221*) gene on BTA2. Total 1 genome-wide and 9 suggestive significant SNPs were observed to be associated with Cu. The most significantly ($p = 1.68 \times 10^{-07}$) associated SNP (ARS-BFGL-NGS-1062) with Cu was identified at 0.63 Mb downstream from the 24-dehydrocholesterol reductase (*DHCR 24*) gene on BTA3. The SNP (BTA-59703-no-rs) significantly associated with Cu ($p = 4.73 \times 10^{-06}$) on BTA25 was located within Calcium voltage-gated channel auxiliary subunit gamma 3 (*CACNG3*) gene. A total of 2 genome-wide and 5 suggestive significant SNP were associated with Zn. Out of these 7 SNPs, 5 significant SNPs were located within 5.29 Mb (4.17–9.47 Mb) region of BTA14. The most significant ($p = 3.50 \times 10^{-07}$) SNP (ARS-BFGL-NGS-76248) associated with Zn was located 0.09 Mb upstream from the ST3 beta-galactoside alpha-1 (*ST3GAL1*) gene. A total of 2 genome-wide and 15 suggestive significant SNPs were significantly associated with Fe. Total six significant SNPs were identified within 4.10 Mb region (65.60–72.17 Mb) on BTA2. The most significant ($p = 3.26 \times 10^{-07}$) SNP (ARS-BFGL-NGS-96204) associated with Fe was located 0.45 Mb away from *NCKAP5* gene on BTA2. The most significantly ($p = 1.02 \times 10^{-04}$) associated SNP (Hapmap58269-rs29018185) with Mn was located on BTA 19, although it was below the cut-off threshold level of significance.

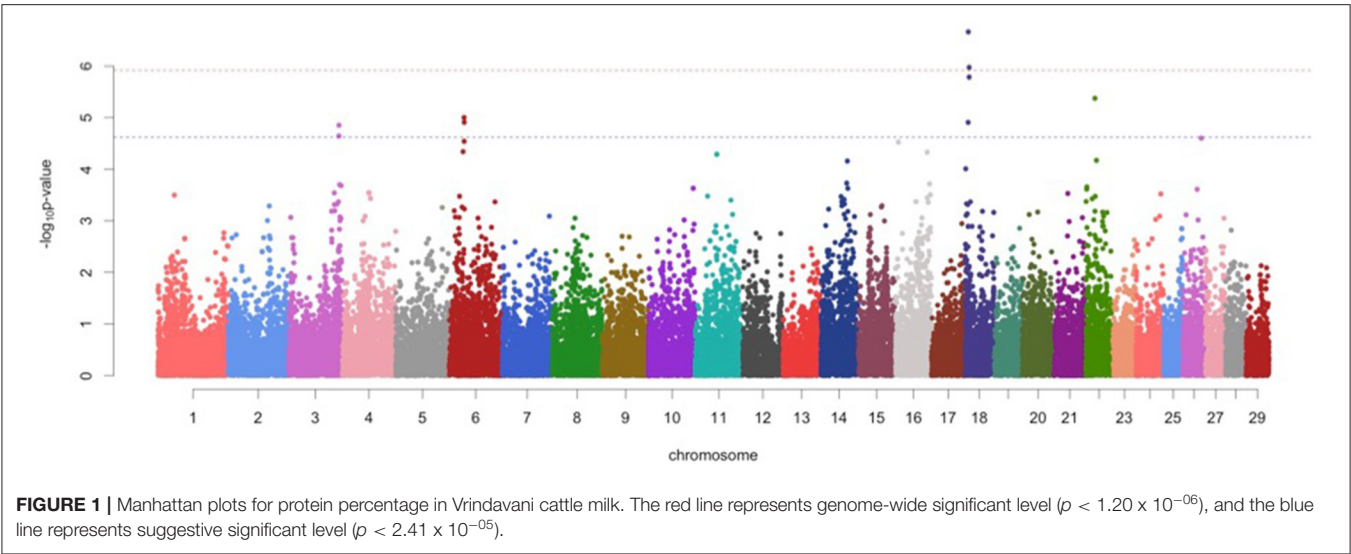
The functional enrichment of genes close to significant SNPs was involved in ion transport (GO:0005248), metal ion binding (GO:0005509), integral and transmembrane protein (GO:0086010), and signaling pathway (**Supplementary Table S1**).

In the present study, several significant SNPs found to be associated with protein percentage and individual mineral concentration in the milk of Vrindavani cattle. To our knowledge, this is the first study to identify genomic regions associated with milk mineral content in composite Vrindavani cattle. Detection of several previously reported genes and genomic regions associated with different milk composition traits indicates their potential role in regulating the concentration of minerals and protein percentage in bovine milk.

For protein percentage, the significant 1.9 Mb region (8.03–9.94 Mb) on BTA 18 was partly overlapping with previously reported QTLs associated with milk coagulation (22). The milk coagulation is directly influenced by the casein composition, which indicates that this region may have the potential to influence the protein percentage in milk (23). This region includes the *CDH 13* (Cadherin-13) gene which is expressed in mammary tissues and associated with the protein content of the milk (24).

TABLE 1 | Descriptive statistics for protein percentage (%) and mineral (Ca, P, Cu, Mn, Zn, and Fe) in the Vrindavani milk.

Traits	Mean	SD	Min	Max	CV
Protein percentage(%)	2.814	0.193	2.480	4.170	6.859
Ca(mg/l)	1,471	589	681	2,960	40.015
P(mg/l)	1,192	319	602	2,360	26.762
Zn(mg/l)	3.694	1.526	1.130	7.773	41.310
Cu(mg/l)	0.892	0.709	0.130	2.213	79.442
Fe (mg/l)	8.925	4.423	3.180	22.400	49.560
Mn(mg/l)	0.794	0.350	0.180	1.930	44.121



On BTA2, 18 significant SNPs were identified for P and Fe content. Two significant SNPs associated with P and Fe were 0.10 and 0.35 Mb away from the inhibin subunit beta B (*INHBB*) gene on BTA2. In humans, *INHBB* gene plays a major role in regulating calcium and phosphorus during bone formation (25). More specifically, *INHBB* facilitates interaction with the beta glycan, which stimulates extracellular matrix mineralisation (26). The SNP associated with Fe is located 0.35 Mb downstream from the IKAROS family zinc finger 2 (*IKZF2*) is involved in metal ion binding (27) and host immune response (28).

On BTA3, three significant SNPs were associated with milk Cu content. Buitenhuis et al. (1) identified several significant SNPs for Cu on BTA3 (91.0-91.3 Mb) in Danish Holstein milk. However, these SNPs do not overlap with the significant SNPs for Cu identified in our study. The most significant SNP associated with Cu is located 0.79 Mb upstream from the Ubiquitin Specific Peptidase 24 (*USP24*) gene. The *USP24* gene belongs to the cysteine proteases family and plays role in protein deubiquitination of the proteins which influences the stability of the casein micelle in the milk (29). However, the role of this gene in the regulation of milk Cu concentration is not known.

On BTA7, four significant SNPs associated with Ca, were located close to *HAPLN1*, *VCAN*, and *EDIL3* genes. These genes are referred to as calcium-binding proteins (GO:0050850, positive regulation of calcium-mediated signaling) and are involved in tissue calcification and bone mineralization in vertebrates (30). In chickens, the EGF-like repeats and discoidin domains 3 (*EDIL3*) gene bind with the calcium ion to guide vesicular transportation of minerals during eggshell calcification (31). Even though its role in milk calcium is not known, *EDIL3* could be considered as a candidate gene, which warrants fine mapping.

The SNP named Hapmap3 1987-BTC-062044 was significantly ($p = 3.50 \times 10^{-07}$) associated with Zn, is located at 0.12 Mb away from the Zinc finger and AT-hook domain containing (ZFAT) on

TABLE 2 | Genome-wide and suggestive significant SNPs for protein percentage in Vrindavani milk.

SNP	BTA	Position	P-value	Nearest candidate gene
ARS-BFGL-NGS-73708	18	8037156	2.16×10^{-07}	BCO1, CMIP
ARS-BFGL-NGS-85875	18	9948949	1.06×10^{-06}	NECAB2, CDH13
ARS-BFGL-NGS-39549	18	9915057	1.62×10^{-06}	OSGIN1, SLC38A8
Hapmap24079-BTA-136416	22	21000933	4.25×10^{-06}	BHLHE40
Hapmap33287-BTC-032371	6	33026144	1.00×10^{-05}	-
Hapmap33121-BTC-033043	6	33441527	1.24×10^{-05}	-
Hapmap52589-rs29020496	18	7630875	1.24×10^{-05}	CENPN, CMC2
ARS-BFGL-NGS-1337	3	113666410	1.41×10^{-05}	USP 40, ATG16L1
BTB-00158861	3	113701688	2.28×10^{-05}	MROH2A, UGT1A1

BTA14, which is involved in the transcription of immune-related genes.

CONCLUSION

Our study identified several candidate genes associated with milk protein percentage and milk minerals, that are involved in ion transport, signaling pathways *via* integral protein, transmembrane membranes, zinc-fingers, and metal ion binding. The strongest association for protein percentage was identified on BTA18. Among studied minerals, the strongest association for Ca, P, Cu, Zn, and Fe were found on BTA 7, 2, 3, 14, and 2, respectively. These roles of genes and genomic regions suggested

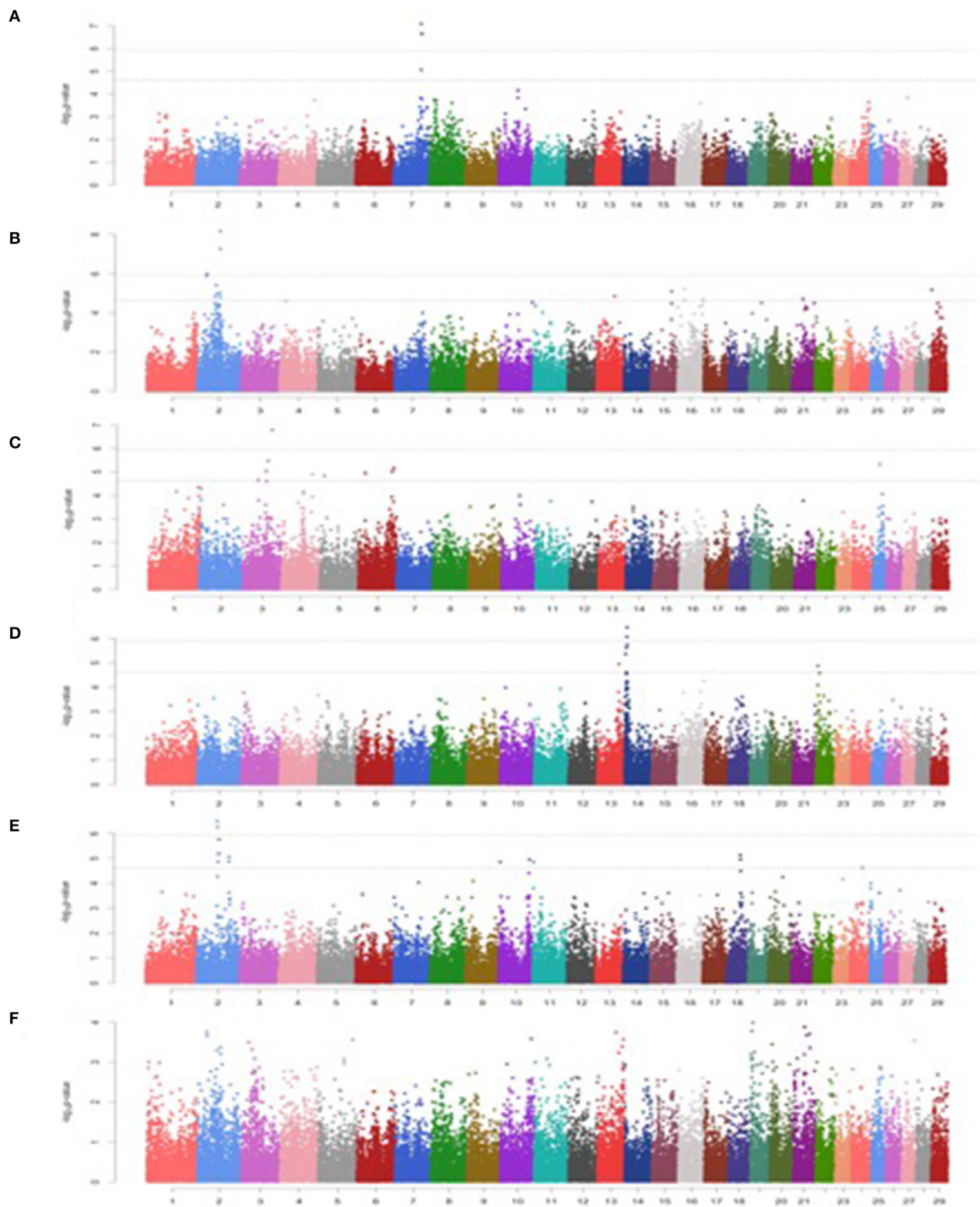


FIGURE 2 | Manhattan plots for minerals in Vrindavani cattle milk. **(A)** Ca, **(B)** P, **(C)** Cu, **(D)** Zn, **(E)** Fe, **(F)** Mn. The red line represents genome-wide significant level ($p < 1.20 \times 10^{-6}$), and the blue line represents suggestive significant level ($p < 2.41 \times 10^{-5}$).

TABLE 3 | Genome-wide and suggestive significant SNPs for minerals (Ca, P, Cu, Zn, Fe, and Mn) in Vrindavani milk.

Minerals	SNP	BTA	Position	P-value	Nearest candidate gene
Ca	BTB-01363189	7	86130978	8.20×10^{-08}	HAPLN1, VCAN
Ca	BTB-00005259	7	86801082	2.29×10^{-07}	EDIL3
Ca	BTB-01259879	7	89658972	2.29×10^{-07}	RASA1
Ca	BTB-00326076	7	86044835	8.72×10^{-06}	XRCC4
P	Hapmap45428-BTA-47987	2	72214554	6.89×10^{-09}	INHBB,CFAP221
P	ARS-BFGL-NGS-85372	2	72369610	5.47×10^{-08}	CFAP221
P	Hapmap39006-BTA-116188	2	30481073	1.03×10^{-06}	SCN9A,SCN1A, SCN7A, TTC21B
P	BTB-01987495	2	29215002	1.27×10^{-06}	XIRP2
P	BTB-01902301	2	59461527	3.91×10^{-06}	HNMT
P	UA-IFASA-4860	16	20472340	6.17×10^{-06}	ESRRG
P	ARS-BFGL-NGS-20615	29	5085168	6.67×10^{-06}	NAALAD2
P	ARS-BFGL-NGS-105277	15	65530435	8.04×10^{-06}	NAT10
P	ARS-BFGL-NGS-20993	2	70957217	9.92×10^{-06}	EN1
P	ARS-BFGL-NGS-109852	2	61611083	1.08×10^{-05}	CXCR4
P	ARS-BFGL-NGS-104967	13	55642499	1.44×10^{-05}	CDH4
P	Hapmap38699-BTA-81626	2	55831708	1.55×10^{-05}	-
P	ARS-BFGL-NGS-58619	2	71779928	1.60×10^{-05}	CFAP221
P	ARS-BFGL-NGS-25171	21	34486610	2.01×10^{-05}	ARID3B
P	Hapmap60572-rs29010980	16	78157760	2.15×10^{-05}	CRB1
P	Hapmap49377-BTA-91839	16	21134492	2.28×10^{-05}	ESRRG
P	ARS-BFGL-NGS-82451	16	81016831	2.37×10^{-05}	CACNA1S
Cu	ARS-BFGL-NGS-1062	3	92082212	1.68×10^{-07}	DHCR24, USP 24
Cu	BTB-01155381	3	79282515	3.40×10^{-06}	INSL5
Cu	BTA-59703-no-rs	25	22197501	4.73×10^{-06}	CACNG3
Cu	BTB-00280101	6	110508154	7.04×10^{-06}	CLNK
Cu	BTA-107777-no-rs	3	73921609	9.19×10^{-06}	-
Cu	ARS-BFGL-NGS-100916	6	105095758	9.61×10^{-06}	PPP2R2C
Cu	BTB-01951920	6	21332376	1.12×10^{-05}	PPA2
Cu	ARS-BFGL-NGS-116643	4	96258114	1.25×10^{-05}	PLXNA4
Cu	Hapmap32870-BTA-162083	5	14189730	1.46×10^{-05}	SLC6A15
Cu	ARS-BFGL-NGS-54036	3	48734443	2.24×10^{-05}	ALG14,ABCD3
Zn	ARS-BFGL-NGS-76248	14	8879810	3.50×10^{-07}	ST3GAL1
Zn	Hapmap31987-BTC-062044	14	8425401	8.59×10^{-07}	ZFAT
Zn	UA-IFASA-7842	14	9472109	1.81×10^{-06}	PHF20L1,KCNQ3
Zn	ARS-BFGL-BAC-10375	14	6616434	2.42×10^{-06}	-
Zn	Hapmap26527-BTC-005059	14	4176618	4.31×10^{-06}	AGO2, DENND3
Zn	ARS-BFGL-NGS-92308	13	67081069	1.13×10^{-05}	MANBAL
Zn	ARS-BFGL-NGS-94105	22	7691228	1.34×10^{-05}	FBXL2
Fe	ARS-BFGL-NGS-96204	2	65600025	3.26×10^{-07}	NCKAP5
Fe	ARS-BFGL-NGS-32707	2	67078207	5.59×10^{-07}	-
Fe	BTA-11592-rs29017351	2	69569936	1.75×10^{-06}	CFAP221
Fe	ARS-BFGL-NGS-101411	2	72163562	1.75×10^{-06}	INHBB
Fe	ARS-BFGL-NGS-22157	2	71906695	6.42×10^{-06}	TMEM177
Fe	BTB-01536946	2	68429456	6.70×10^{-06}	DPP10
Fe	ARS-BFGL-NGS-29914	18	42706451	7.38×10^{-06}	DPY19L3
Fe	Hapmap40319-BTA-19899	2	102634206	8.86×10^{-06}	VWC2L
Fe	ARS-BFGL-NGS-5074	18	43588701	1.10×10^{-05}	RHPN2, FAAP24, LRP3
Fe	Hapmap43494-BTA-122491	10	94570405	1.12×10^{-05}	-
Fe	ARS-BFGL-NGS-60714	2	102060979	1.33×10^{-05}	IKZF2
Fe	Hapmap49519-BTA-19205	2	68058489	1.39×10^{-05}	DPP10
Fe	ARS-BFGL-NGS-113875	11	3106988	1.41×10^{-05}	ACTR1B, ZAP70
Fe	ARS-BFGL-NGS-40198	10	2617295	1.41×10^{-05}	YTHDC2
Fe	Hapmap57118-rs29009938	24	39365195	2.35×10^{-05}	EPB41L3

that milk mineral concentration is probably regulated by transportation and homeostasis of ions. These identified variants are a step forward to characterize the molecular mechanism affecting milk minerals in Vrindavani cattle. However, additional validation of detected variants and their association with milk minerals is required on large sample size.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: Figshare, <https://figshare.com/articles/dataset/Genotypes/12808343/2>.

ETHICS STATEMENT

The animal study was reviewed and approved by Institute Animal Ethics Committee of Indian Veterinary Research Institute, Bareilly.

REFERENCES

- Buitenhuis B, Poulsen NA, Larsen LB, Sehested J. Estimation of genetic parameters and detection of quantitative trait loci for minerals in Danish Holstein and Danish Jersey milk. *BMC Genet.* (2015) 16: 1–8. doi: 10.1186/s12863-015-0209-9
- Vegarud GE, Langsrud T, Svenning C. Mineral-binding milk proteins and peptides; occurrence, biochemical and technological characteristics. *British J Nutr.* (2000) 84(S1):91–8. doi: 10.1017/S0007114500002300
- Sanchez MP, Rocha D, Charles M, Boussaha M, Hozé C, Brochard M, et al. Sequence-based GWAS and post-GWAS analyses reveal a key role of SLC37A1, ANKH, and regulatory regions on bovine milk mineral content. *Sci Rep.* (2021) 11:1–15. doi: 10.1038/s41598-021-87078-1
- Cashman KD. Milk minerals (including trace elements) and bone health. *Int Dairy J.* (2006) 16:1389–98. doi: 10.1016/j.idairyj.2006.06.017
- Górska-Warsewicz H, Rejman K, Laskowski W, Czacotko M. Milk and dairy products and their nutritional contribution to the average polish diet. *Nutrients.* (2019) 11:1771. doi: 10.3390/nu11081771
- Diniz WJDS, Coutinho LL, Tizioto PC, Cesar ASM, Gromboni CF, Nogueira ARA, et al. Iron content affects lipogenic gene expression in the muscle of Nelore beef cattle. *PLoS ONE.* (2016) 11:e0161160. doi: 10.1371/journal.pone.0161160
- Toffanin V, De Marchi M, Lopez-Villalobos N, Cassandro M. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. *Int Dairy J.* (2015) 41:68–73. doi: 10.1016/j.idairyj.2014.10.002
- Malacarne M, Franceschi P, Formaggioni P, Sandri S, Mariani P, Summer A. Influence of micellar calcium and phosphorus on rennet coagulation properties of cows milk. *J Dairy Res.* (2014) 81:129–36. doi: 10.1017/S0022029913000630
- Manuelian CL, Penasa M, Visentin G, Zidi A, Cassandro M, De Marchi M. Mineral composition of cow milk from multibreed herds. *Anim Sci J.* (2018) 89:1622–7. doi: 10.1111/asj.13095
- Patel J, Singh A, Chaudhary R, Kumar A, Jadhav SE, Naskar S, et al. Factors affecting milk minerals and constituents in indigenous vis-à-vis crossbred cattle and buffaloes. *Ind J Anim Sci.* (2018) 88:71–7.
- Zhou C, Li C, Cai W, Liu S, Yin H, Shi S, et al. Genome-wide association study for milk protein composition traits in a Chinese

AUTHOR CONTRIBUTIONS

AK conceived and designed the experiments. AS and AK performed the experiments, analyzed the data, and wrote the paper. CG, AP, and TD contributed reagents/materials/analysis tools. All authors contributed to the article and approved the submitted version.

FUNDING

This project work was supported by the CAAST-ACLH project of NAHEP. AS was recipient of fellowship from ICMR fellowship during her PhD programme.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fvets.2022.760364/full#supplementary-material>

- Holstein population using a single-step approach. *Front Genet.* (2019) 10:72. doi: 10.3389/fgene.2019.00072
- van den Berg I, Xiang R, Jenko J, Pausch H, Boussaha M, Schrooten C, et al. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. *Genet Sel Evol.* (2020) 52:1–16. doi: 10.1186/s12711-020-00556-4
- Singh A, Mehrotra A, Gondro C, da Silva Romero AR, Pandey AK, Karthikeyan A, et al. Signatures of selection in composite vrindavani cattle of India. *Front Genet.* (2020) 11:589496. doi: 10.3389/fgene.2020.589496
- Trancoso I, Roseiro LB, Martins AP, Trancoso MA. Validation and quality assurance applied to goat milk chemical composition: minerals and trace elements measurements. *Dairy Sci Technol.* (2009) 89:241–56. doi: 10.1051/dst/2009013
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* (2007) 81:559–75. doi: 10.1086/519795
- Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* (2009) 19:1655–64. doi: 10.1101/gr.094052.109
- Gondro C. Genome wide association studies. In: *Primer to Analysis of Genomic Data Using R*. Cham: Springer (2015). p. 73–103.
- Han B, Kang HM, Eskin E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* (2009) 5:e1000456. doi: 10.1371/journal.pgen.1000456
- Li C, Sun D, Zhang S, Wang S, Wu X, Zhang Q, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS ONE.* (2014) 9:e96186. doi: 10.1371/journal.pone.0096186
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* (2010) 26:841–2. doi: 10.1093/bioinformatics/btq033
- Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* (2009) 37:1–13. doi: 10.1093/nar/gkn923
- Duchemin SI, Glantz M, de Koning DJ, Paulsson M, Fikse WF. Identification of QTL on chromosome 18 associated with non-coagulating milk in Swedish Red cows. *Front Genet.* (2016) 7:57. doi: 10.3389/fgene.2016.00057
- Frederiksen PD, Andersen KK, Hammershøj M, Poulsen HD, Sørensen J, Bakman M, et al. Composition and effect of blending of noncoagulating,

- poorly coagulating, and well-coagulating bovine milk from individual Danish Holstein cows. *J Dairy Sci.* (2011) 94:4787–99. doi: 10.3168/jds.2011-4343
24. Glantz M, Gustavsson F, Bertelsen HP, Stålhammar H, Lindmark-Månsson H, Paulsson M, et al. Bovine chromosomal regions affecting rheological traits in acid-induced skim milk gels. *J Dairy Sci.* (2015) 98:1273–85. doi: 10.3168/jds.2014-8137
 25. Farr JN, Roforth MM, Fujita K, Nicks KM, Cunningham JM, Atkinson EJ, et al. Effects of age and estrogen on skeletal gene expression in humans as assessed by RNA sequencing. *PLoS ONE.* (2015) 10:e0138347. doi: 10.1371/journal.pone.0138347
 26. Eijken M, Swagemakers S, Koedam M, Steenbergen C, Derkx P, Uitterlinden AG, et al. The activin A-follistatin system: Potent regulator of human extracellular matrix mineralization. *FASEB J.* (2007) 21:2949–60. doi: 10.1096/fj.07-8080com
 27. Cizkova M, Cizeron-Clairac G, Vacher S, Susini A, Andrieu C, Lidereau R, et al. Gene expression profiling reveals new aspects of PIK3CA mutation in ERalpha-positive breast cancer: major implication of the Wnt signaling pathway. *PLoS ONE.* (2010) 5:e15647. doi: 10.1371/journal.pone.0015647
 28. Reverter A, Ballester M, Alexandre PA, Mármol-Sánchez E, Dalmau A, Quintanilla R, et al. A gene co-association network regulating gut microbial communities in a Duroc pig population. *Microbiome.* (2021) 9:1–16. doi: 10.1186/s40168-020-00994-8
 29. Wickramasinghe S, Rincon G, Islas-Trejo A, Medrano JF. Transcriptional profiling of bovine milk using RNA sequencing. *BMC Genomics.* (2012) 13:1–14. doi: 10.1186/1471-2164-13-45
 30. Oh SH, Kim JW, Kim Y, Lee MN, Kook MS, Choi EY, et al. The extracellular matrix protein Edil3 stimulates osteoblast differentiation through the integrin $\alpha 5 \beta 1$ /ERK/Runx2 pathway. *PLoS ONE.* (2017) 12:e0188749. doi: 10.1371/journal.pone.0188749
 31. Gautron J, Stapane L, Le Roy N, Nys Y, Rodriguez-Navarro AB, Hincke MT. Avian eggshell biomineralization: an update on its structure, mineralogy and protein tool kit. *BMC Mol. Cell Biol.* (2021) 22:1–17. doi: 10.1186/s12860-021-00350-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Singh, Kumar, Gondro, Pandey, Dutt and Mishra. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Whole Genome Sequencing Analysis to Identify Candidate Genes Associated With the rib eye Muscle Area in Hu Sheep

Yuan Zhao¹, Xiaoxue Zhang¹, Fadi Li^{2,3}, Deyin Zhang¹, Yukun Zhang¹, Xiaolong Li¹, Qizhi Song⁴, Bubo Zhou¹, Liming Zhao¹, Jianghui Wang¹, Dan Xu¹, Jiangbo Cheng¹, Wenxin Li¹, Changchun Lin¹, Xiaobin Yang¹, Xiwen Zeng¹ and Weimin Wang^{1*}

¹College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China, ²The State Key Laboratory of Grassland Agro-ecosystems, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, China, ³Engineering Laboratory of Sheep Breeding and Reproduction Biotechnology in Gansu Province, Minqin, China, ⁴Linze County Animal Disease Prevention and Control Center of Gansu Province, Linze, China

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Francesco Tiezzi,
University of Florence, Italy
Pablo Fonseca,
University of Guelph, Canada

*Correspondence:

Weimin Wang
wangwm@gsau.edu.cn

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 29 November 2021

Accepted: 14 February 2022

Published: 14 March 2022

Citation:

Zhao Y, Zhang X, Li F, Zhang D,
Zhang Y, Li X, Song Q, Zhou B, Zhao L,
Wang J, Xu D, Cheng J, Li W, Lin C,
Yang X, Zeng X and Wang W (2022)
Whole Genome Sequencing Analysis
to Identify Candidate Genes
Associated With the rib eye Muscle
Area in Hu Sheep.
Front. Genet. 13:824742.
doi: 10.3389/fgene.2022.824742

In sheep meat production, the rib eye area is an important index to evaluate carcass traits. However, conventional breeding programs have led to slow genetic progression in rib eye muscle area. Operationalizing molecular marker assisted breeding is an optimized breeding method that might improve this situation. Therefore, the present study used whole genome sequencing data to excavate candidate genes associated with the rib eye muscle. Male Hu lambs ($n = 776$) with pedigrees and 274 lambs with no pedigree were included. The genetic parameters of the rib eye area were estimated using a mixed linear mixed model. The rib eye area showed medium heritability (0.32 ± 0.13). Whole-genome sequencing of 40 large rib eye sheep [17.97 ± 1.14 , (cm^2)] and 40 small rib eye sheep [7.89 ± 0.79 , (cm^2)] was performed. Case-control genome-wide association studies and the fixation index identified candidate rib eye-associated genes. Seven single nucleotide polymorphisms (SNPs) in six genes (*ALS2*, *ST6GAL2*, *LOC105611989*, *PLXNA4*, *DPP6*, and *COL12A1*) were identified as candidates. The study population was expanded to 1050 lambs to perform KASPar genotyping on five SNPs, which demonstrated that SNPs in *LOC105611989*, *DPP6*, and *COL12A1* correlated significantly with the rib eye area, which could be used as genetic markers for molecular breeding of the rib eye area. The results provided genetic parameters estimated on the rib eye area and information for breeding based on carcass traits in Hu sheep.

Keywords: whole genome sequencing, case-control GWAS, genetic parameters, association analysis, sheep breeding

INTRODUCTION

Improvement of carcass traits is very important for farmers and the sheep industry, because the commercial value of a sheep carcass is determined by a range of different aspects, notably the rib eye area. Studies suggest that the rib eye area is significantly associated with the carcass lean meat yield (Anderson et al., 2015). Among many carcass traits, the first to be assessed using ultrasonic scanning was the rib eye area, which revealed the importance attached by consumers and farmers to this trait (Arias et al., 2007). Therefore, adding the rib eye area to breeding programs is an important tool to

improve sheep carcass values. Hu sheep predominate among Chinese mutton sheep breeds, but show huge variation in their rib eye area because of the selection pressure resulting from breeding methods and geographical differences. The ability to accurately predict the size of the rib eye area early in the life of Hu sheep is extremely valuable for producers to meet the requirements of their target market, and for genetic selection.

In the last 20 years, the heritability of sheep carcass traits has been estimated in many sheep breeds from different regions and countries using different models and approaches (Kefelegn et al., 1999; Greeff et al., 2015; Mortimer et al., 2017; Blasco et al., 2019; Savoia et al., 2019). Several studies have quantified the heritability of certain measured or scored meat quality traits in sheep breeds in different farming and market systems, which demonstrated the existence of genetic variability among these traits, which could be exploited for genetic improvement. However, because the techniques used to obtain phenotypes at the population level are expensive and laborious, it is very rare for measurements of the rib eye area to be used directly in the selection of specialized breeds, e.g., in cattle (Savoia et al., 2019). Carcass merit traits are expressed at the later stages of animal production and are mostly assessed at slaughter, which sacrifices potential breeding stock, although real-time ultrasound imaging technologies can be used to measure the rib eye area (Meirelles et al., 2011; Li et al., 2018).

Reducing the generation interval while maintaining a good level of selection accuracy would improve breeding efficiency (Kefelegn et al., 1999). Early selection is a practical and effective method, which is mainly carried out using genomic selection and molecular marker-assisted breeding (Dekkers, 2003; Tong and Sun, 2015). Identification of the quantitative trait major genes of the rib eye area is necessary. Mutation sites identified using Whole genome sequencing (WGS) represent the ideal pan DNA markers for genetic analyses, because theoretically, they contain all causative polymorphisms. In cattle breeding, genome-wide association studies (GWAs) on production, meat quality, and lactation traits have been reported widely and have been applied in practical production (Lin et al., 2020; Mukiibi, 2020). However, in sheep, only a few GWAs for body size, body weight, and wool traits have been reported (Kominakis and Hager-Theodorides, 2017; Palombo et al., 2020; Tao et al., 2020), and most research used single nucleotide polymorphism (SNP) chips, not the whole genome sequence.

Designing effective breeding programs based on accurate estimates of genetic parameters and molecular markers is one way to solve the problem of slow genetic progression of the rib eye area of sheep. In the present study, the genetic parameters of the rib eye area were estimated using different models, and some genetic correlations between the rib eye area and economic traits were estimated. WGS of High and Low rib eye area groups from a reference population of Hu sheep was performed. The aim was to estimate the heritability of the rib eye area. In addition, we aimed to establish the feasibility of adding the rib eye area trait to breeding goal, and to identify candidate genes that could be used in molecular marker assisted breeding, which will increase the breeding value of Hu sheep.

MATERIALS AND METHODS

Animal Management and Data Collection

In the present study, a total of 1050 male Hu lambs, born between 2018 and 2019, were included. Among them, 776 individuals possessed a complete pedigree record. The full pedigree of the experiment population comprised 1500 individuals, including 70 male parents and 600 female parents, which were from three National Core Breeding Farms of Sheep and Goats (NCBFSG) (Gansu Zhongtian Sheep Industry Co. Ltd., Gansu Zhongsheng Huamei Sheep Industry Co. Ltd., and Gansu Pukang Sheep Industry Co. Ltd.) and a large scale Hu sheep farm (Gansu Sanyang Sheep Industry Co. Ltd.). The population with pedigree information could be used to estimate genetic parameters. Each lamb was weaned at 56 days old and transferred to Minqin Defu Agricultural Technology Co., Ltd. (Performance Measurement Centre, PMC; Gansu Province, China). There were two batches in 2018 (May to September and September to December, respectively) and two batches in 2019 (May to September and September to December, respectively).

From 64 days old, all male Hu lambs were raised indoors in individual 0.8×1 m pens until the lambs were 180 days old, and each column of pens had a separate trough and drinking bowl for the sheep to feed freely. Before the study, a 14-day transition period was implemented, during which the proportion of pellets in the diet was gradually increased by 7.1% per day, until the diet became totally granulated. The formula, raw materials, and manufacturing plant of the feed were consistent. The experiment began with lambs at 80 days of age. They were weighed at 80 days to obtain their initial weight (BW80) and raised for up to 180 days. The body weight of each lamb at 180 days old was measured as the final weight (BW180). The feed intake of the lambs during the experimental period was recorded every 20 days. The feed intake for 80–180 days was calculated from five records. Then, we calculated the average daily gain (ADG) and average daily feed intake (ADFI). In addition, individual feed efficiency [residual feed intake (RFI) and feed conversion ratio (FCR)] were estimated. RFI was the residual of the multiple linear regression of ADFI on ADG and the medium metabolic weight [MBW = $(0.5 \times (BW80 + BW180))^{0.75}$] (Tortereau et al., 2020). The FCR was calculated as the ratio of ADFI and ADG.

At the 180 days old, blood was collected and the lambs were slaughtered in accordance with established national standards by Ministry of Agriculture and Rural Affairs of the People's Republic of China (NY/T 3471-2019). The blood was stored at -20°C until DNA isolation. Twenty-four hours after slaughter, the rib eye muscle area was covered with oleic acid paper from between the fifth and sixth thoracic vertebrae, and used to sketch out the muscle area in a one-to-one ratio. These sketches were scanned using an EPSON Scanner (V19, IDN, Epson, Nagano, Japan) and analyzed using ImageJ (1.8.0; NIH, Bethesda, MD, USA) from which we obtained the rib eye area [REA (cm^2)]. Thickness (cm) of the backfat (BF) above the REA was measured using calipers. We then selected 5% of the lambs ($776 \times 0.05 \approx 40$ individuals)

with extreme REA phenotypes and categorized them into high and low groups, respectively.

DNA Isolation and Sequencing

Genomic DNA of every lamb was isolated using an EasyPure Blood Genomic DNA Kit (TransGen Biotech, Beijing, China), according to the manufacturer's product description. The quality and integrity of the DNA was measured using the A260/280 ratio and by agarose gel electrophoresis. Then, 1.5 µg of high quality genomic DNA from the high and low groups (80 individuals) were used to generate 80 paired-end sequencing libraries with an inset size of 500 bp. The libraries were sequenced as 150 bp paired end reads using the Illumina NovaSeq PE 150 platform (Illumina, San Diego, CA, USA).

Sequence Alignment and Single Nucleotide Polymorphism Calling

Quality control was performed on the raw data to improve subsequent analyses using FastQC (Version 0.11.1). Reads with joint sequences and sequences for which the number of undetected bases in single-end sequencing exceeded 10% of the total length of the sequence were eliminated. Sequences with low-quality bases with a mass value of $Q \leq 5$ in the single-end sequence were removed. The clean reads were mapped onto the *Ovis aries* reference genome (Oar_rambouillet_v1.0) using BWA (Burrows-Wheeler Aligner) (Version 0.7.8) software with default parameters (Li and Durbin, 2009). To reduce mismatches generated by PCR amplification before sequencing, duplicated reads were removed using Genome Analysis Toolkit (GATK, version 3.4.0) and GATK was used to generate variant call format (VCF) files (McKenna et al., 2010). SNPs were filtered using Vcftools (version 0.1.14) software and processed as follows: SNPs with call rates <75%, minor allele frequencies (MAF) < 0.05, and minor coverage depth <3. Thus, 80 individuals with 9,875,037 SNPs were kept for further genomic analysis (Danecek et al., 2011).

Statistical Analysis

Estimation of (co)variance Components and Genetic Parameters

To accurately estimate (co)variance components, we constructed an animal mixed model with univariate and bivariate analysis between each pair of feature pairs, using the restricted maximum likelihood (REML) method implemented in the ASReml software (Gilmour et al., 2009). First, the random effects included in the analyses were the direct additive genetic effect of the animal, maternal genetic effects, and random errors. The fixed effects tested and their levels were: Litter size at individual birth (four levels), fattening season (two levels), and the farm before weaning (four levels). All the fixed effects that significantly affected the trait (p -value < 0.05), as determined using the ASReml Wald program, were incorporated into the (co)variance estimation. Three models that accounted for the genetic effects and bivariate genetic (co)covariance were fitted and were as follows:

$$Y = X\beta + Z_a a + e \quad (1)$$

$$Y = X\beta + Z_a a + Z_m m_a + e \quad (\text{cov}(a, m_a) = 0) \quad (2)$$

$$G = \begin{bmatrix} A\sigma^2 & 0 \\ 0 & A\sigma_m^2 \end{bmatrix}$$

Where Y is the phenotype measurement vector of a trait, while β , a , m_a , and e are vectors of fixed, additive direct animal genetic effects, maternal genetic effects, and residual effects, respectively, with association matrices X , Z_a , and Z_m . G is the variance structure of the model; A is the additive relationship matrix based on the pedigree; and σ^2 and σ_m^2 are the additive direct and maternal genetic. Residuals were assumed to follow a normal distribution, $e \sim N(0, R \otimes I)$.

The bivariate model was as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad (3)$$

$$G = \begin{bmatrix} \sigma_{a_1}^2 & r_g \sigma_{a_1} \sigma_{a_2} \\ r_g \sigma_{a_1} \sigma_{a_2} & \sigma_{a_2}^2 \end{bmatrix} \otimes A$$

Where y_1 and y_2 are the phenotypic measurement vectors of traits 1 and 2, b_1 and b_2 are the fixed effect vectors previously described; a_1 and a_2 are the vectors of the animal random genetic effects; and e_1 and e_2 are the vectors of random residuals. X_1 and X_2 are design matrices of the fixed effects and Z_1 and Z_2 are design matrices relating traits to random animal genetic effects. σ_{a_1} and σ_{a_2} represent the additive genetic variances of traits 1 and 2, and r_g represents the genetic correlation between them. We used Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Loglik to judge the merits of the model through ASReml software.

Candidate Gene Screening

Three methods (fixation index (FST), Fisher's exact test, and the Chi-squared test) were used to screen mutation sites with significant differences in their allele frequency between the High and Low groups. FST quantified the allele frequency differences between the High and Low groups using Vcftools software (version 0.1.14) (Danecek et al., 2011). We used FST because Weir and Cockerham proposed that genetic structures (or genetic polymorphisms) cause the degree of population differentiation and formulated the FST value (0–1) to represent the degree of allele frequency differences (Cockerham, 1984; Danecek et al., 2011). In the present study, mutation sites with the top 20 FST values were identified as putative selection sites. The genomic data of 80 individuals were subjected to principal components analysis (PCA) [using PLINK (version 1.9)] and permutational multivariate analysis of variance (PERMANOVA) (using Vegan R packages) to verify whether there is population stratification. We constructed a genome-wide case/control design within the High and Low groups using Fisher's exact test and the Chi-squared test to calculate the p -values using PLINK (version 1.9) (Purcell et al., 2007). In addition, the first five principal components were added as covariates for the association analysis using Fisher's test. We used the false discovery rate (FDR) method to set the p -value threshold. The genome-wide threshold value was calculated according to an

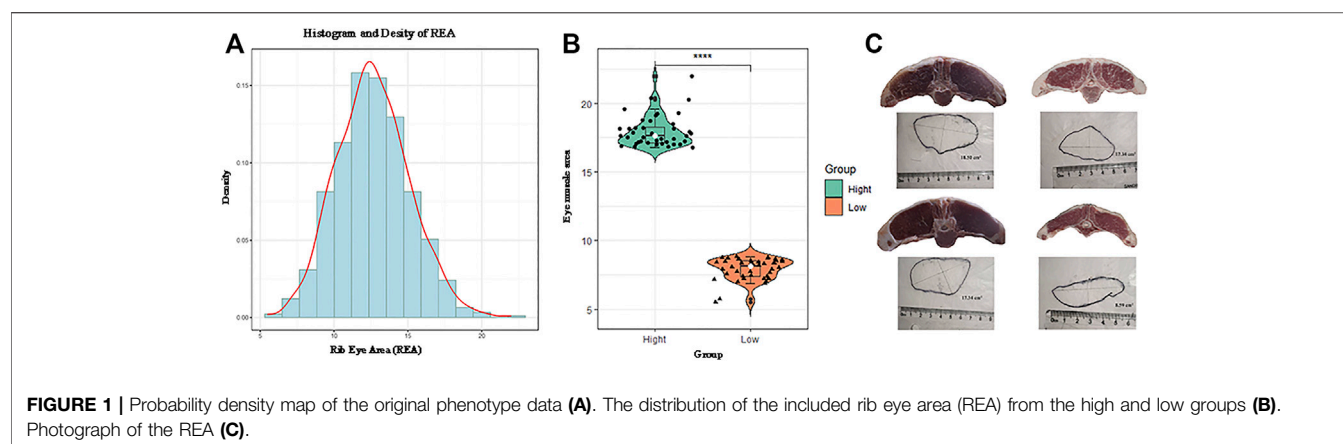


FIGURE 1 | Probability density map of the original phenotype data (A). The distribution of the included rib eye area (REA) from the high and low groups (B). Photograph of the REA (C).

TABLE 1 | Estimates of heritability (h^2) and variance components.

	log	σ_a^2	σ_e^2	σ_m^2	h^2	h_m	AIC	BIC
Mod1	-1357.78	1.42 ± 0.62	3.04 ± 0.58		0.32 ± 0.13		2719.568	2729.525
Mod2	-1356.14	1.27 ± 0.83	2.96 ± 0.61	0.25 ± 0.43	0.28 ± 0.13	0.06 ± 0.05	2718.275	2733.210

Note: Mod1 $Y = X\beta + Z_a a + e$; Mod2 $Y = X\beta + Z_a a + Z_m m_a + e$; σ_a^2 additive effect variance component; σ_e^2 residual variance component; σ_m^2 Maternal genetic effect variance component; h^2 heritability; h_m Maternal heritability; AIC: Akaike information criterion; BIC: Bayesian information criterion.

FDR of 0.05. The p.adjust function of R (version 4.0.3) was used to calculate the FDR. (Supplementary Tables S1, S2).

Genotyping and Association

We selected SNPs that met the requirements in all three screening methods (FST, Fisher's exact test, and the Chi-squared test). PCR extension primers for the SNPs were designed using the genomic DNA sequences. DNA samples from 1050 lambs with phenotypic records were genotyped using KAspar genotyping technology (LGC Genomics, Hoddesdon, UK) (Smith and Maughan, 2015) to further screen the SNPs. The general linear model analysis in the SPSS 22.0 software (IBM Corp., Armonk, NY, USA) was used to analyze the association between the genotypes and the REA (Zhang et al., 2019). The linear model with the fixed and covariate effects was as follows:

$$y_{ijkl} = \mu + \text{Genotype}_i + \text{Batch}_j + \text{Season}_k + 180\text{BW} + \varepsilon_{ijkl}$$

In this model, y_{ijkl} is the vector of the phenotypic information; μ is the population mean; Genotype_i is the i^{th} genotype; Batch_j is the effect of the j^{th} batch; Season_k is the effect of the k^{th} season; 180BW was a covariate added to the model; ε_{ijkl} is the residual corresponding to the trait observation value with $\varepsilon \sim N(0, \sigma^2)$. All effects met the statistically significant criterion ($p < 0.05$).

RESULTS

Phenotype and Genetic Parameters

In the present study, 776 animals with their REA phenotypes were included to estimate the genetic parameters. The average

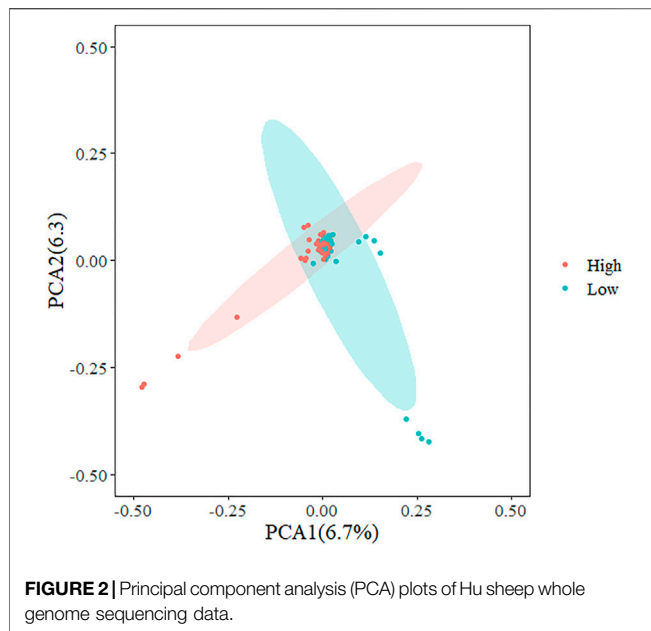
TABLE 2 | Estimates of genetic (rg) and phenotypic (rp) correlations for rib eye area, carcass, and important economic traits in Male Hu sheep.

Character	REA	
	rg	rp
BW80	0.58 ± 0.18	0.32 ± 0.03
BW180	0.80 ± 0.19	0.42 ± 0.03
ADFI80-180	0.55 ± 0.24	0.37 ± 0.03
ADG80-180	0.91 ± 0.16	0.34 ± 0.03
FCR80-180	-0.05 ± 0.08	-0.02 ± 0.03
RFI80-180	-0.01 ± 0.10	0 ± 0.03
BF	0.06 ± 0.33	0.11 ± 0.03

Note: rg: genetic correlation; rp: phenotypic correlation; REA = rib eye area (cm^2); BW80 = weight at 80 days old (kg); BW180 = weight at 180 days old (kg); ADFI = average daily feed intake (kg); ADG80-180 = average daily gain (kg); FCR80-180 = Feed conversion ratio; RFI80-180 = residual feed intake (kg); BF = backfat (cm).

REA was $12.64 \pm 2.46 \text{ cm}^2$. The probability density of the phenotype data is shown in Figure 1A. The REA of the high (mean = $17.97 \pm 1.14 \text{ cm}^2$) and low (mean = $7.89 \pm 0.79 \text{ cm}^2$) groups screened from the large population are shown in Figure 1B.

Heritability and variance components for the REA are shown in Table 1. The REA heritability estimated using different models showed little difference (0.32 and 0.28). The litter common environmental effect accounted for the majority of the REA. The REA trait showed moderate heritability. Maternal heritability was significantly greater than zero for the REA trait in Hu sheep; however, the influence of the maternal genetic effect on the heritability and the additive variance component estimation of the REA trait was minimal. We calculated the parameters (AIC and BIC) of the two evaluation models (Table 1).



Estimates of genetic and phenotype correlations between the REA and important economic traits (body weight, feed intake, feed efficiency, and backfat, respectively) are shown in **Table 2**. The correlations were high for the REA and production traits, with high genetic correlations between BW80, BW180, ADFI80–180, ADG80–180, and REA (0.58, 0.80, 0.55, and 0.91, respectively). Positive moderate phenotype correlations were estimated between the REA and production traits, at 0.32, 0.42, 0.37, and 0.34, respectively. Genetic correlations between feed efficiency traits and the REA were negligible (−0.05 and −0.01). Low correlations were estimated between feed efficiency and the REA at the phenotypic level (−0.02 and 0). BF and the REA were genetically correlated with a large standard error, while there was a positive phenotypic correlation between the BF and the REA.

Candidate Genes and SNPs

To identify variation in the REA in sheep genomes, we identified the candidate genes associated with an increased REA in Hu sheep. We performed WGS of Hu lambs in the high and low groups. We obtained a total of 700 Gb of raw data, with an average depth of 5-fold for each individual. The average number of reads was 58785931 and the average % GC was 43.6. (**Supplementary Table S3**). The effective sequencing rate was 99.42%, which suggested that the data had high quality and could be used for further in-depth analysis.

In the present study, three methods were used to identify SNPs and genes associated with the REA in Hu sheep. **Figure 2** shows that population stratification did not appear. The top two principal components accounted for a small proportion of the variance (6.7 and 6.3%). In addition, we performed permutational multivariate analysis of variance (PERMANOVA) using vegan

packages, which produced a p -value of 0.0014. This verified that there were no subgroups in the population ($p < 0.05$). To detect strongly selected signals, we searched the sheep genome for single SNPs with an increased genetic distance (FST). We set a threshold to select the top 20 SNPs and we used the PLINK software for PCA. The first five principal components were included in Fisher's exact test. According to the p -value of every SNP, eleven significant SNPs (**Supplementary Table S1**) were identified for the REA at the secondary identified signal ($p < 4.82 \times 10^{-8}$). Then, we used the Chi-squared test to detect seven significant SNPs (**Supplementary Table S2**) for the REA at the secondary identified signal ($p < 3.77 \times 10^{-8}$). Manhattan maps of the GWAS are shown in **Figure 3**.

We found that the SNPs selected using the three methods were decreasing and coincident (Fst > Fisher's exact > Chi-squared). Subsequently, seven target SNPs were selected that overlapped among the three methods and were annotated to the closest gene in the Oar_rambouillet_v1.0 genome. Six genes containing these seven SNPs were defined as candidate genes (**Table 3**).

Association Analysis of the Candidate Genes With the REA

To determine whether the selected candidate genes had a significant effect in the REA, in the enlarged experimental population ($n = 1050$), five SNPs (*ST6GAL2* (encoding ST6 beta-galactoside alpha-2,6-sialyltransferase 2) SNP g.65927208 T > C, *LOC105611989* SNP g.65927208A > T, *DPP6* (encoding dipeptidyl peptidase like 6) SNP g.126636893A > G, *COL12A1* (encoding collagen type XII alpha 1 chain) SNP g.2261361 T > A, and *COL12A1* SNP g.2261369 T > A) were subjected to genotyping, which generated three genotypes (**Figure 4**). The result of association analysis indicated that *LOC105611989* SNP g.65927208A > T, *DPP6* g.126636893 SNP A > G, *COL12A1* SNP g.2261361 T > A, and *COL12A1* SNP g.2261369 T > A were significantly associated with the REA ($p < 0.05$); however, *ST6GAL2* SNP g.65927208 T > C did not have a significant impact on the REA. For *LOC105611989*, the male Hu sheep with the TT genotype had largest REA and those with the AA genotype had the smallest REA among the three genotypes ($p < 0.05$). The REA of male Hu sheep carrying the AA and AG genotypes of the *DPP6* SNP g.126636893A > G was increased compared with that in male Hu sheep carrying the GG genotype ($p < 0.05$). The effect of the *COL12A1* g.2261361 T > A and *COL12A1* g.2261369 T > A SNPs was significant on the REA. For *COL12A1* SNP g.2261361 T > A, male Hu sheep with the TT genotype had a larger REA than those with the AA genotype. By contrast, the REA in the experimental population with the TT genotype of the *COL12A1* g.2261369 T > A SNP was significantly lower compared with that in the sheep with the AT genotype.

LOC105611989 SNP g.65927208A > T, *COL12A1* SNP g.2261361 T > A, and *COL12A1* SNP g.2261369 T > A were in Hardy–Weinberg equilibrium ($p > 0.05$). However, *ST6GAL2* SNP g.65927208 T > C and *DPP6* SNP g.126636893A > G did not conform to the Hardy–Weinberg equilibrium ($p < 0.05$).

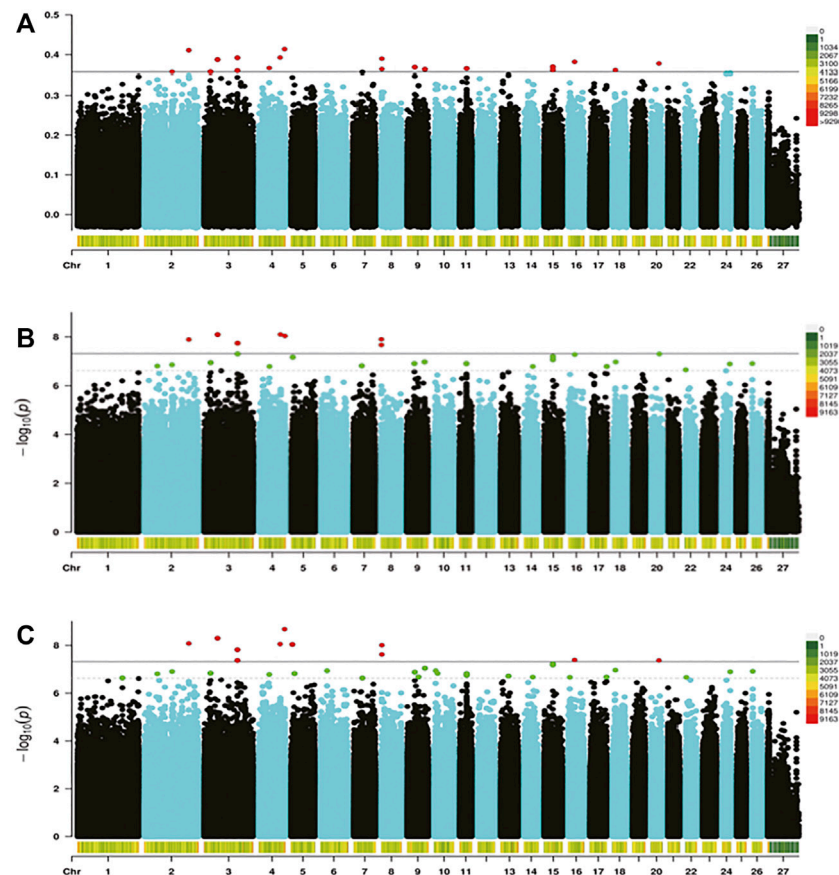


FIGURE 3 | Genome-wide distribution of *Fst* (A), Fisher's exact test (B), and the chi-squared test (C). The horizontal black line in the figure shows the threshold of methods at the secondary identified signal [$-\log(4.82 \times 10^{-8})$ and $-\log(3.77 \times 10^{-8})$, respectively].

DISCUSSION

Genetic improvement has played an important role in productivity gains in animal farming. The REA is a valuable economic trait that could affect production traits in animals (Savoia et al., 2019). The REA has been reported in many studies that mainly investigated nutrition (Silva et al., 2019; Jiao et al., 2020; Redoy et al., 2020); however, there have been relatively few studies reporting the genetic mechanism of the REA, and it was not until 2013 that a GWAS for the REA in sheep was reported (Zhang et al., 2013), although GWAS studies had been used earlier to study the REA in other livestock.

In our study, we estimated the genetic parameters of the REA in Hu breeding, and candidate gene mining was carried out. The heritability of the REA-based genetic parameters with different models was moderate. Roden et al. reported lower estimates of for rib eye muscle depth and width (0.11 and 0.08) in Scottish Blackface breed than those in the present study (Roden et al., 2003). In our population, maternal genetics was not an important influence on the REA. Previous studies on different cattle breeds (Nellore, Red Angus, and Piedmontese) reported h^2 values for the REA ranging from 0.21 to 0.26 (Boldt et al., 2018; Savoia et al., 2019). In 2007, rib eye muscle depth heritability of Kivircik lambs

was estimated using Bayesian inference as 0.23 (Cemal et al., 2016). Our results of the heritability of the REA are also similar to those of a previous report (0.23) (Pollott and Greeff, 2004).

In addition, we found interesting correlations between the REA and certain important economic traits. Positive genetic correlations were estimated between the REA and production traits (0.58 for BW80, 0.80 for BW180, 0.55 for ADFI, and 0.91 for ADG). As expected, the size of an individual has a decisive impact on the size of the REA. Mortimer et al. reported high genetic correlations (from 0.37 to 0.70) between the REA and body weight at different stages (Mortimer et al., 2018). Phenotype correlation analysis supported the results of the genetic correlation analysis (Table 2).

We identified negligible genetic correlations between REA and feed efficiency traits (-0.05 and -0.01), and ADFI and BW correlated positively with the REA, suggesting that animals with a larger REA grow faster and consume more feed, so it is expected that their feed efficiency would be the same. A negative genetic correlation between the measured muscle depth and feed efficiency traits (-0.3 ± 0.15) with RFI and -0.15 ± 0.18 with FCR has been observed in other breeds (Tortereau et al., 2020). This correlation suggests that animals with a larger REA are expected to have excellent feed efficiency. In 2004, the genetic correlation

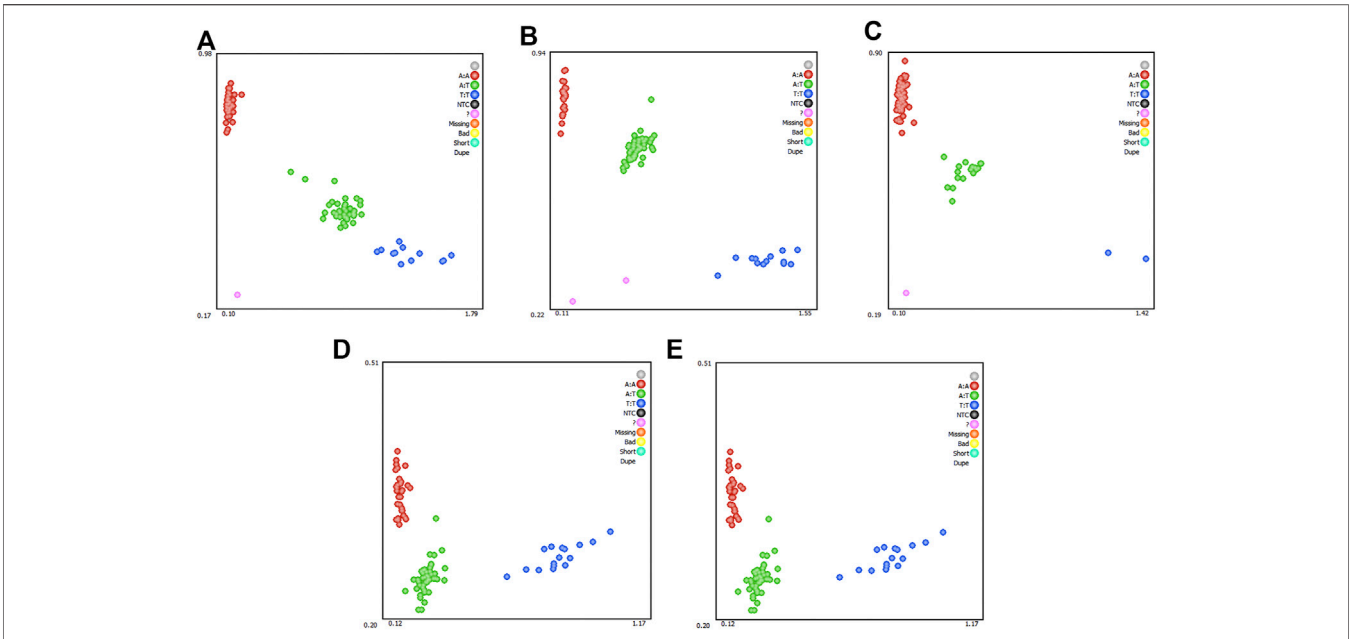


FIGURE 4 | Genotyping of *ST6GAL2* (A), *LOC105611989* (B), *DPP6* (C), and *COL12A1* (D and E) single nucleotide polymorphisms (SNPs). Note: Red, green, and blue represents three genotypes, respectively; while the pink dots indicate genotyping failure.

TABLE 3 | The significant SNPs for the rib eye area in Hu sheep.

Chr	Position	Gene	Fst-value	Fisher's-p-Value	Chi-squared-p-Value	REF	ALT	Post type
chr2	218145714	<i>ALS2</i>	0.41	8.38e-09	1.28e-08	T	C	downstream
chr3	65927208	<i>ST6GAL2</i>	0.39	5.04e-09	8.02e-09	T	C	downstream
chr3	163720238	<i>LOC105611989</i>	0.39	1.52e-08	1.81e-08	A	T	downstream
chr4	104628299	<i>PLXNA4</i>	0.39	8.81e-09	7.92e-09	TC	T	intronic
chr4	126636893	<i>DPP6</i>	0.41	2.10e-09	9.13e-09	A	G	intronic
chr8	2261361	<i>COL12A1</i>	0.39	9.38e-09	1.26e-08	T	A	intronic
chr8	2261369	<i>COL12A1</i>	0.36	2.39e-08	2.14e-08	T	A	intronic

Gene: Gene obtained by annotation of significant SNP.

TABLE 4 | Association between the REA and different genotypes of the *ST6GAL2*, *LOC105611989*, *DPP6*, and *COL12A1* genes.

Gene/Loci	Genotype	N	p-value	REA	HWE (p-value)
<i>ST6GAL2</i>	T:T	93	0.40	12.42 ± 2.18	<0.001
chr3:65927208	T:C	356		12.68 ± 2.33	
T > C	C:C	587		12.38 ± 2.37	
<i>LOC105611989</i>	A:A	145	0.0004	11.78 ± 2.43 ^c	0.792
chr3:163720238	A; T	479		12.40 ± 2.35 ^b	
A > T	T:T	415		12.82 ± 2.27 ^a	
<i>DPP6</i>	A:A	25	0.019	12.89 ± 2.30 ^a	0.007
chr4:126636893	A:G	198		12.67 ± 2.27 ^a	
A > G	G:G	727		12.27 ± 2.29 ^b	
<i>COL12A1</i>	T:T	222	<0.01	12.50 ± 4.46 ^a	0.094
chr8:2261361	A:T	449		12.44 ± 2.22 ^{ab}	
T > A	A:A	279		12.13 ± 2.21 ^b	
<i>COL12A1</i>	T:T	267	0.19	12.12 ± 2.22 ^b	0.248
chr8:2261369	A:T	460		12.47 ± 2.25 ^a	
T > A	A:A	226		12.43 ± 2.22 ^{ab}	

p-value: Significance of genotype as fixed effect in linearity model (p < 0.05); REA: The mean value of rib eye muscle area in three genotypes; The letters (a,b,c) represents the mean value with different superscripts differ significantly (p < 0.05); HWE(p-value): The Significance of Hardy-Weinberg equilibrium.

between fat depth and the REA was estimated as 0.05 (Pollott and Greeff, 2004). In our study, negative genetic relationship between the REA and BF was estimated (0.06). However, the genetic correlation between the REA and BF had a large standard error. Fat serves as an energy store in animals. A previous study found evidence that the genetic correlation between fat reserves and production traits can change across environments (Pollott and Greeff, 2004). Based on the results of the genetic parameters, we suggest that the REA trait should be incorporated into breeding of Hu sheep.

The REA is commonly used to reflect the muscular development of the carcass. To further investigate the mechanisms affecting the REA, we identified genes that regulate the REA in the Hu sheep genome using three methods. SNPs were identified in six genes: *ALS2* (encoding Alsin Rho guanine nucleotide exchange factor 2) SNP g.218145714 T > C, *ST6GAL2* SNP g.65927208 T > C, *LOC105611989* SNP g.163720238 A > T, *PLXNA4* (encoding plexin A4) SNP g.104628299 TC > T, *DPP6* SNP g.126636893 A > G, *COL12A1* SNP g.2261361 T > A, and *COL12A1* SNP g.2261369 T > A.

The mutations in the *ALS2* gene may cause Familial amyotrophic lateral sclerosis type 2, which is a juvenile autosomal recessive motor neuron disease. The *ALS2* gene product, ALS2/alsin, forms a homophilic oligomer and acts as a guanine nucleotide-exchange factor (GEF) for the small GTPase Rab5. This oligomerization is crucial for both Rab5 activation and ALS2-mediated endosome fusion and maturation in cells (Asako et al., 2003; Sato et al., 2018). *ALS2* might act as a modulator in neuronal differentiation and development through regulation of membrane dynamics (Otomo et al., 2008). Actin and myoglobin is involved in guiding regulation of membrane dynamics (Poukkula et al., 2011). Macro pinocytosis is the cytoplasmic membrane folding mediated by actin that can non-selectively encapsulate a large number of extracellular nutrients and liquid macromolecules, and transport them to the lysosome for degradation after the cell is stimulated by nutrient factors or phorbol ester. Macro pinocytosis alters the nutritional physiology of muscle cells. The development and nutritional metabolism of muscle cells are inseparable from this process. The mutation of *ALS2* has led to a certain degree of muscle atrophy in the longissimus dorsi muscle in Hu sheep. This caused the REA to become smaller. The association analysis of the larger group also supported this view.

Studies have reported that *PLXNA4* and Semaphorin3A show a high degree of interaction (Quintá et al., 2014; Poltavski et al., 2019; Limoni et al., 2021). *PLXNA4* is used as a star gene in the Axon guidance - *Homo sapiens* KEGG category. *PLXNA4* acts directly upstream of Rac, p21-activated kinase (PAK), and RhoD functions. Rho GTPases are known to regulate actin dynamics (Ridley, 2006). In motile cells, during the process of mitosis, the activities of GTPases of different Rho families are separated in intracellular space (Pal et al., 2020). A previous study found that RhoD could stimulate actin polymerization and affect plasma membrane protrusion and/or vesicular traffic (Ridley, 2006). Pal et al. (2020) reported that manipulating Rac and Arp2/3 activity resulted in polar body defects during mitosis and meiosis in sea urchin embryos and sea star oocytes, and suggested that Rac and

Arp2/3-mediated actin networks might directly antagonize Rho signaling. PAK is an effector downstream target of Rho-GTPases Rac1 and Cdc42, which can activate its downstream target cofilin via LIM kinase-1, and subsequently supports cell migration and invasion through the polymerization of actin filaments (Mierke et al., 2020). The Rac/PAK/GC/cGMP signaling pathway allows the action of RAC and PAK to affect actin simultaneously (Guo et al., 2010). Actin forms part of the muscle cytoskeleton. If there is a lack of actin production, the development and mitosis of the longissimus dorsi muscle will be affected, resulting in muscular dysplasia and thus reducing the rib eye area muscle.

DPP6 is an auxiliary subunit of the Kv4 family of voltage-gated K (+) channels, which is known to enhance channel surface expression and potentially accelerate their kinetics. Many studies suggest that *DPP6* is consistently and strongly associated with susceptibility to amyotrophic lateral sclerosis (ALS) in different human populations of European ancestry (Del Bo et al., 2008; Es et al., 2008; Lin et al., 2014). This suggests that *DPP6* acts as a susceptibility gene for muscle atrophy symptoms in humans. It is possible that *DPP6* has the same effect on muscle atrophy in animals.

After fitting multiple factors in the association analysis in a large population, we found that the *COL12A* gene is related to the size of the REA of Hu sheep. *COL12A* is mainly involved in the degradation of the extracellular matrix pathway and collagen chain trimerization. In 2014, a myopathy/anhydrotic ectodermal dysplasia (EDA)-like disease caused by heterozygous *COL12A1* variants was reported (Punetha et al., 2017). Mice with *Col12A1* knockout showed muscle weakness with decreased grip strength, combined with bone fragility, short stature, and kyphoscoliosis (Yaquán et al., 2014). Delbaere et al. (2020) believed that mixed myopathy was caused by defects in collagen XII and VI, and by variant-specific alterations in the extracellular matrix resulting from *COL12A1* mutation. We believe that this disease also occurs in Hu sheep.

In addition, we performed genotyping and association analysis on the seven SNPs in the enlarged experimental population. Five loci were typed successfully and three loci were significantly associated with the REA. These three SNPs could be used for marker-assisted selection. Therefore, we hypothesized that these candidate genes might affect muscle development in Hu sheep. We cannot perform whole genome sequencing in a large Hu sheep population because of the cost; however, further studies are needed to confirm the biological effects of the identified SNPs in the rib eye area.

According to the association analysis of candidate genes, we found that in male Hu sheep *ST6GAL2* SNP g. 65927208 T > C has no association with the REA. The results showed that the SNP has skewness in this population, which led to a lack of association with the REA. Although this research verified the association between candidate SNPs and the REA in a large population, verification at the cell and protein levels is still needed.

CONCLUSION

The results obtained in the present study showed that the rib eye area was heritable and there is genetic variability, which could

theoretically be exploited for the genetic improvement of sheep. We identified *LOC105611989* SNP g.65927208A > T, *DPP6* SNP g.126636893A > G, *COL12A1* SNP g.2261361 T > A, and *COL12A1* SNP g.2261369 T > A, whose protein products have metabolic functions in muscle cells, as candidate genetic markers of the rib eye area. Our results could be applied to marker assisted breeding of sheep and are expected to improve for the slow genetic progress of the rib eye area in sheep. Moreover, this study provides important information for estimating the heritability parameters of the rib eye area and for molecular breeding of carcass traits in Hu sheep.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA779188>.

ETHICS STATEMENT

The animal study was reviewed and approved by The Institutional Animal Care and Ethics Committee of Gansu Agriculture University.

REFERENCES

- Anderson, F., Pannier, L., Pethick, D. W., and Gardner, G. E. (2015). Intramuscular Fat in Lamb Muscle and the Impact of Selection for Improved Carcass Lean Meat Yield. *Animal* 9 (6), 1081–1090. doi:10.1017/S1751731114002900
- Arias, P., Pini, A., Sanguinetti, G., Sprechmann, P., Cancela, P., Fernandez, A., et al. (2007). Ultrasound Image Segmentation with Shape Priors: Application to Automatic Cattle Rib-Eye Area Estimation. *IEEE Trans. Image Process.* 16 (6), 1637–1645. doi:10.1109/TIP.2007.896604
- Asako, O., Shinji, H., Takeya, O., Hikaru, M., Ryota, K., Hitoshi, N., et al. (2003). ALS2, a Novel Guanine Nucleotide Exchange Factor for the Small GTPase Rab5, Is Implicated in Endosomal Dynamics. *Hum. Mol. Genet.* 12 (14), 1671. doi:10.1021/es001545w
- Blasco, M., Campo, M. M., Balado, J., and Sañudo, C. (2019). Effect of Texel Crossbreeding on Productive Traits, Carcass and Meat Quality of Segureña Lambs. *J. Sci. Food Agric.* 99 (7), 3335–3342. doi:10.1002/jsfa.9549
- Boldt, R. J., Speidel, S. E., Thomas, M. G., and Enns, R. M. (2018). Genetic Parameters for Fertility and Production Traits in Red Angus Cattle. *J. Anim. Sci.* 96 (10), 4100–4111. doi:10.1093/jas/sky294
- Cemal, I., Karaman, E., Firat, M. Z., Yilmaz, O., Ata, N., and Karaca, O. (2016). Bayesian Inference of Genetic Parameters for Ultrasound Scanning Traits of Kivircik Lambs. *Animal* 11 (03), 375–381. doi:10.1017/S1751731116001774
- Cockerham, B. S. W. C. (1984). Estimating F-Statistics for Analysis of Population Structure. *Evolution* 38 (6), 1359–1370. doi:10.2307/2408641
- da Silva, J. R. C., de Carvalho, F. F. R., de Andrade Fereira, M., de Souza, E. J. O., Maciel, M. I. S., Barreto, L. M. G., et al. (2019). Carcass Characteristics and Meat Quality of Sheep Fed Alfalfa hay to Replace Bermuda Grass hay. *Trop. Anim. Health Prod.* 51 (8), 2455–2463. doi:10.1007/s11250-019-01962-7
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330
- Dekkers, J. C. (2003). Commercial Application of Marker- and Gene-Assisted Selection in Livestock: Strategies and Lessons. *J. Anim. Sci.* 82 (E-Suppl), E313–E328. doi:10.2527/2004.8213_supplE313x

AUTHOR CONTRIBUTIONS

YZ and WW conceived the study. FL, DZ, XL, QS, LZ, and YZ contributed to construction of the reference population and growth performance measurement. JW, DX, JC, WL, and CL contributed extracting DNA. BZ, XY, and XZ contributed to the calculation of the rib eye area. YZ, XZ, YKZ, XL, and WW analyzed the data. YZ wrote the paper. YZ and WW revised the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 31760651), the National Joint Research on Improved Breeds of Livestock and Poultry (grant no.19210365), and the Earmarked Fund for China Agriculture Research System (grant no. CARS-38).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.824742/full#supplementary-material>

- Del Bo, R., Ghezzi, S., Corti, S., Santoro, D., Prelle, A., Mancuso, M., et al. (2008). DPP6 Gene Variability Confers Increased Risk of Developing Sporadic Amyotrophic Lateral Sclerosis in Italian Patients. *J. Neurol. Neurosurg. Psychiatry* 79 (9), 1085. doi:10.1136/jnnp.2008.149146
- Delbaere, S., Dhooze, T., Syx, D., Petit, F., Goemans, N., Destrée, A., et al. (2020). Novel defects in collagen XII and VI expand the mixed myopathy/Ehlers–Danlos syndrome spectrum and lead to variant-specific alterations in the extracellular matrix. *Genet. Med.* 22 (1), 112–123. doi:10.1038/s41436-019-0599-6
- Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2009). *ASReml User Guide Release 3.0*. Hemel Hempstead, United Kingdom: VSN International Ltd.
- Greeff, J. C., Safari, E., Fogarty, N. M., Hopkins, D. L., Brien, F. D., Atkins, K. D., et al. (2015). Genetic Parameters for Carcass and Meat Quality Traits and Their Relationships to Liveweight and Wool Production in Hogget Merino Rams. *J. Anim. Breed. Genet.* 125 (3), 205–215. doi:10.1111/j.1439-0388.2007.00711.x
- Guo, D., Zhang, J. J., and Huang, X. Y. (2010). A New Rac/PAK/GC/cGMP Signaling Pathway. *Mol. Cell Biochem* 334 (1–2), 99–103. doi:10.1007/s11010-009-0327-7
- Jiao, J., Wang, T., Zhou, J., Degen, A. A., Gou, N., Li, S., et al. (2020). Carcass Parameters and Meat Quality of Tibetan Sheep and Small-tailed Han Sheep Consuming Diets of Low-protein Content and Different Energy Yields. *J. Anim. Physiol. Anim. Nutr.* 104 (4), 1010–1023. doi:10.1111/jpn.13298
- Keefeleg, K., Suess, R., Mielenz, N., Schuele, L., and Lengerken, G. V. (1999). Genetic and Phenotypic Parameter Estimates of Growth and Carcass Value Traits in Sheep. *J. Agric. Res.* 35 (3), 287–298.
- Kominakis, A., Hager-Theodorides, A. L., Zoidis, E., Saridakis, A., Antonakos, G., and Tsiamis, G. (2017). Combined GWAS and 'guilt by Association'-Based Prioritization Analysis Identifies Functional Candidate Genes for Body Size in Sheep. *Genet. Sel. Evol.* 49 (1), 41. doi:10.1186/s12711-017-0316-3
- Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324
- Li, Q., Wang, Y., Tan, L., Leng, J., Lu, Q., Tian, S., et al. (2018). Effects of Age on slaughter Performance and Meat Quality of Binlangiang Male buffalo. *Saudi J. Biol. Sci.* 25 (2), 248–252. doi:10.1016/j.sjbs.2017.10.001

- Limoni, G., Murthy, S., Jabaudon, D., Dayer, A., and Niquille, M. (2021). PlexinA4-Semaphorin3A-mediated Crosstalk between Main Cortical Interneuron Classes Is Required for Superficial Interneuron Lamination. *Cell Rep.* 34 (4), 108644. doi:10.1016/j.celrep.2020.108644
- Lin, L., Long, L. K., Hatch, M. M., and Hoffman, D. A. (2014). DPP6 Domains Responsible for its Localization and Function. *J. Biol. Chem.* 289 (46), 32153–32165. doi:10.1074/jbc.M114.578070
- Lin, S., Wan, Z., Zhang, J., Xu, L., Han, B., and Sun, D. (2020). Genome-Wide Association Studies for the Concentration of Albumin in Colostrum and Serum in Chinese Holstein. *Animals* 10 (12), 2211. doi:10.3390/ani10122211
- Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110
- Meirelles, S. L., Gouveia, G. V., Gasparin, G., Alencar, M. M., Gouveia, J. J. S., and Regitano, L. C. A. (2011). Candidate Gene Region for Control of Rib Eye Area in Canchim Beef Cattle. *Genet. Mol. Res.* 10 (2), 1220–1226. doi:10.4238/vol10-2gmr1175
- Mierke, C. T., Puder, S., Aermes, C., Fischer, T., and Kunschmann, T. (2020). Effect of PAK Inhibition on Cell Mechanics Depends on Rac1. *Front. Cell Dev. Biol.* 8, 13. doi:10.3389/fcell.2020.00013
- Mortimer, S. I., Fogarty, N. M., van der Werf, J. H. J., Brown, D. J., Swan, A. A., Jacob, R. H., et al. (2018). Genetic Correlations between Meat Quality Traits and Growth and Carcass Traits in Merino Sheep1. *J. Anim. Sci.* 96 (9), 3582–3598. doi:10.1093/jas/sky232
- Mortimer, S. I., Hatcher, S., Fogarty, N. M., van der Werf, J. H. J., Brown, D. J., Swan, A. A., et al. (2017). Genetic Correlations between Wool Traits and Carcass Traits in Merino Sheep1. *J. Anim. Sci.* 95 (6), 2385–2398. doi:10.2527/jas2017.162810.2527/jas.2017.1385
- Mukiibi, R. (2020). Genetic Architecture of Quantitative Traits in Beef Cattle Revealed by Genome Wide Association Studies of Imputed Whole Genome Sequence Variants: II: Carcass merit Traits. *BMC Genomics* 21 (1), 38. doi:10.1186/s12864-019-6273-1
- Otomo, A., Kunita, R., Suzuki-Utsunomiya, K., Mizumura, H., Onoe, K., Osuga, H., et al. (2008). ALS2/alsin Deficiency in Neurons Leads to Mild Defects in Macropinocytosis and Axonal Growth. *Biochem. Biophysical Res. Commun.* 370 (1), 87–92. doi:10.1016/j.bbrc.2008.01.177
- Pal, D., Ellis, A., Sepúlveda-Ramírez, S. P., Salgado, T., Terrazas, I., Reyes, G., et al. (2020). Rac and Arp2/3-Nucleated Actin Networks Antagonize Rho during Mitotic and Meiotic Cleavages. *Front. Cell Dev. Biol.* 8, 591141. doi:10.3389/fcell.2020.591141
- Palombo, V., Gaspa, G., Conte, G., Pilla, F., Macciotta, N., Mele, M., et al. (2020). Combined Multivariate Factor Analysis and GWAS for Milk Fatty Acids Trait in Comisana Sheep Breed. *Anim. Genet.* 51 (4), 630–631. doi:10.1111/age.12948
- Pollott, G. E., and Greeff, J. C. (2004). Genotype × Environment Interactions and Genetic Parameters for Fecal Egg Count and Production Traits of Merino Sheep1. *J. Anim. Sci.* 82 (10), 2840–2851. doi:10.2527/2004.82102840x
- Poltavski, D. M., Colombier, P., Hu, J., Duron, A., Black, B. L., and Makita, T. (2019). Venous Endothelin Modulates Responsiveness of Cardiac Sympathetic Axons to Arterial Semaphorin. *Elife* 8, e42528. doi:10.7554/eLife.42528
- Poukkula, M., Kremneva, E., Serlachius, M., and Lappalainen, P. (2011). Actin-depolymerizing Factor Homology Domain: a Conserved Fold Performing Diverse Roles in Cytoskeletal Dynamics. *Cytoskeleton* 68 (9), 471–490. doi:10.1002/cm.20530
- Punetha, J., Kesari, A., Hoffman, E. P., Gos, M., Kamińska, A., Kostera-Pruszczyk, A., et al. (2017). NovelCol12A1variant Expands the Clinical Picture of Congenital Myopathies with Extracellular Matrix Defects. *Muscle Nerve* 55 (2), 277–281. doi:10.1002/mus.25232
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795
- Quintá, H. R., Pasquini, J. M., Rabinovich, G. A., and Pasquini, L. A. (2014). Glycan-dependent Binding of Galectin-1 to Neuropilin-1 Promotes Axonal Regeneration after Spinal Cord Injury. *Cell Death Differ* 21 (6), 941–955. doi:10.1038/cdd.2014.14
- Redoy, M. R. A., Shuvo, A. A. S., Cheng, L., and Al-Mamun, M. (2020). Effect of Herbal Supplementation on Growth, Immunity, Rumen Histology, Serum Antioxidants and Meat Quality of Sheep. *Animal* 14 (11), 2433–2441. doi:10.1017/S1751731120001196
- Ridley, A. J. (2006). Rho GTPases and Actin Dynamics in Membrane Protrusions and Vesicle Trafficking. *Trends Cell Biol.* 16 (10), 522–529. doi:10.1016/j.tcb.2006.08.006
- Roden, J. A., Merrell, B. G., Murray, W. A., and Haresign, W. (2003). Genetic Analysis of Live Weight and Ultrasonic Fat and Muscle Traits in a hill Sheep Flock Undergoing Breed Improvement Utilizing an Embryo Transfer Programme. *Animalence* 76 (03), 367–373. doi:10.1046/j.1365-2052.2003.01012.x10.1017/s1357729800058598
- Sato, K., Otomo, A., Ueda, M. T., Hiratsuka, Y., Suzuki-Utsunomiya, K., Sugiyama, J., et al. (2018). Altered Oligomeric States in Pathogenic ALS2 Variants Associated with Juvenile Motor Neuron Diseases Cause Loss of ALS2-Mediated Endosomal Function. *J. Biol. Chem.* 293 (44), 17135–17153. doi:10.1074/jbc.RA118.003849
- Savoia, S., Albera, A., Brugiapaglia, A., Di Stasio, L., Cecchinato, A., and Bittante, G. (2019). Heritability and Genetic Correlations of Carcass and Meat Quality Traits in Piemontese Young Bulls. *Meat Sci.* 156, 111–117. doi:10.1016/j.meatsci.2019.05.024
- Smith, S. M., and Maughan, P. J. (2015). SNP Genotyping Using KASPar Assays. *Methods Mol. Biol.* 1245, 243–256. doi:10.1007/978-1-4939-1966-6_18
- Tao, L., He, X. Y., Pan, L. X., Wang, J. W., Gan, S. Q., and Chu, M. X. (2020). Genome-wide Association Study of Body Weight and Conformation Traits in Neonatal Sheep. *Anim. Genet.* 51 (2), 336–340. doi:10.1111/age.12904
- Tong, J., and Sun, X. (2015). Genetic and Genomic Analyses for Economically Important Traits and Their Applications in Molecular Breeding of Cultured Fish. *Sci. China Life Sci.* 58 (2), 178–186. doi:10.1007/s11427-015-4804-9
- Tortoreau, F., Marie-Etancelin, C., Weisbecker, J.-L., Marcon, D., Bouvier, F., Moreno-Romieux, C., et al. (2020). Genetic Parameters for Feed Efficiency in Romane Rams and Responses to Single-Generation Selection. *Animal* 14 (4), 681–687. doi:10.1017/s1751731119002544
- van Es, M. A., van Vught, P. W., Blauw, H. M., Franke, L., Saris, C. G., Van Den Bosch, L., et al. (2008). Genetic Variation in DPP6 Is Associated with Susceptibility to Amyotrophic Lateral Sclerosis. *Nat. Genet.* 40 (1), 29–31. doi:10.1038/ng.2007.52
- Yaqun, Z., Daniela, Z., Yayoi, I., Shreya, G., Gudrun, S., Knut, B., et al. (2014). Recessive and Dominant Mutations in COL12A1 Cause a Novel EDS/myopathy Overlap Syndrome in Humans and Mice. *Hum. Mol. Genet.* 23 (9), 2339–2352. doi:10.1093/hmg/ddt627
- Zhang, D., Zhang, X., Li, F., Li, C., La, Y., Mo, F., et al. (2019). Transcriptome Analysis Identifies Candidate Genes and Pathways Associated with Feed Efficiency in Hu Sheep. *Front. Genet.* 10, 1183. doi:10.3389/fgene.2019.01183
- Zhang, L., Liu, J., Zhao, F., Ren, H., Xu, L., Lu, J., et al. (2013). Genome-Wide Association Studies for Growth and Meat Production Traits in Sheep. *Plos One* 8 (6), e66569. doi:10.1371/journal.pone.0066569

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhao, Zhang, Li, Zhang, Zhang, Li, Song, Zhou, Zhao, Wang, Xu, Cheng, Li, Lin, Yang, Zeng and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Genome-Wide Association Study and F_{ST} Analysis Reveal Four Quantitative Trait Loci and Six Candidate Genes for Meat Color in Pigs

Hang Liu^{1,2,3†}, Liming Hou^{1,2†}, Wuduo Zhou¹, Binbin Wang^{1,2}, Pingping Han¹, Chen Gao^{1,2}, Peipei Niu², Zongping Zhang², Qiang Li⁴, Ruihua Huang^{1,2} and Pinghua Li^{1,2*}

¹Institute of Swine Science, Nanjing Agricultural University, Nanjing, China, ²Huaian Academy, Nanjing Agricultural University, Huaian, China, ³Hangzhou Academy of Agricultural Sciences, Hangzhou, China, ⁴Huaiyin Pig Breeding Farm of Huaian City, Huaian, China

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Yulin Jin,
Emory University, United States
Paolo Zambonelli,
University of Bologna, Italy

*Correspondence:

Pinghua Li
lipinghua718@njau.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 01 September 2021

Accepted: 24 January 2022

Published: 08 April 2022

Citation:

Liu H, Hou L, Zhou W, Wang B, Han P,
Gao C, Niu P, Zhang Z, Li Q, Huang R
and Li P (2022) Genome-Wide
Association Study and F_{ST} Analysis
Reveal Four Quantitative Trait Loci and
Six Candidate Genes for Meat Color
in Pigs.
Front. Genet. 13:768710.
doi: 10.3389/fgene.2022.768710

Meat color is the primary criterion by which consumers evaluate meat quality. However, there are a few candidate genes and molecular markers of meat color that were reported for pig molecular breeding. The purpose of the present study is to identify the candidate genes affecting meat color and provide the theoretical basis for meat color molecular breeding. A total of 306 Suhuai pigs were slaughtered, and meat color was evaluated at 45 min and 24 h after slaughter by CIELAB color space. All individuals were genotyped using GeneSeek GGP-Porcine 80K SNP BeadChip. The genomic estimated breeding values (GEBVs), heritability, and genetic correlation of meat color were calculated by DMU software. The genome-wide association studies (GWASs) and the fixation index (F_{ST}) tests were performed to identify SNPs related to meat color, and the candidate genes within 1 Mb upstream and downstream of significant SNPs were screened by functional enrichment analysis. The heritability of L^* 45 min, L^* 24 h, a^* 45 min, a^* 24 h, b^* 45 min, and b^* 24 h was 0.20, 0.16, 0.30, 0.13, 0.29, and 0.22, respectively. The genetic correlation between a^* (a^* 45 min and a^* 24 h) and L^* (L^* 45 min and L^* 24 h) is strong, whereas the genetic correlation between b^* 45 min and b^* 24 h is weak. Forty-nine significant SNPs associated with meat color were identified through GWAS and F_{ST} tests. Among these SNPs, 34 SNPs were associated with L^* 45 min within a 5-Mb region on Sus scrofa chromosome 11 (SSC11); 22 SNPs were associated with a^* 45 min within a 14.72-Mb region on SSC16; six SNPs were associated with b^* 45 min within a 4.22-Mb region on SSC13; 11 SNPs were associated with b^* 24 h within a 2.12-Mb region on SSC3. These regions did not overlap with meat color-associated QTLs reported previously. Moreover, six candidate genes (*HOMER1*, *PIK3CG*, *PIK3CA*, *VCAN*, *FABP3*, and *FKBP1B*), functionally related to muscle development, phosphatidylinositol phosphorylation, and lipid binding, were detected around these significant SNPs. Taken together, our results provide a set of potential molecular markers for the genetic improvement of meat color in pigs.

Keywords: meat color, heritability, marker, GWAS, candidate genes, pigs

INTRODUCTION

In recent years, global meat consumption is increasing year by year (Katare et al., 2020). As an indicator of meat freshness and safety, meat color can directly affect the consumer purchase desire of pork (Tomasevic et al., 2021). The discoloration of meat surface will cause huge economic losses and is harmful to the meat industry (Suman et al., 2014). It is important for producers to use objective and scientific methods to evaluate the meat color (Wu and Sun, 2013). Currently, the CIELAB (Commission Internationale de l'Éclairage LAB) color space is the most commonly used system for assessing meat color. It is a three-dimensional Cartesian space containing three mutually independent parameters, including L^* (lightness), a^* (redness), and b^* (yellowness).

Meat color is influenced by many factors, including genetic, nutrition, and slaughter methods, among which the genetic method has a greater impact (Sellier, 1998). The heritability of meat color is low to moderate and varies among different population. Cabling et al. reported that the heritability of L^* , a^* , and b^* was 0.44, 0.68, and 0.64 in 690 Duroc pigs, respectively (Cabling et al., 2015). However, Miar et al. reported that the heritability of meat color of 2075 offsprings from Duroc x Large White pigs was slightly lower, and the heritability of L^* , a^* , and b^* was 0.28, 0.26, and 0.31, respectively (Miar et al., 2014). Meat quality traits have been declined because the previous swine breeding program has been focused on improving the pig's growth rate and lean meat yield (Chen et al., 2018). However, meat quality traits are now being incorporated into the pig farm breeding objective because of the demand of the consumer market for high-quality pork (Wu et al., 2017). Traditional breeding methods are difficult to improve meat color because the determination of meat color is expensive and can only be performed after slaughter. Currently, molecular breeding technology has been widely used owing to the cost of genome sequencing, and gene chip scanning is reducing. Marker-assisted selection (MAS) is an important method of molecular breeding in which population selection is carried out through molecular markers and quantitative trait loci (QTLs) related to target traits (Borakhatariya, 2017; Visscher and Haley, 1995). The Animal QTLdb has included 651 QTLs related with meat color of pig; these QTLs are mainly distributed on the Sus scrofa chromosomes SSC6, SSC7, SSC15, and SSC16. Previous studies have reported that the *RN* gene and *PRKAG3* gene can affect the a^* value of flesh color and the *RYR1* gene can improve the L^* value of flesh meat (Bertram et al., 2000; Küchenmeister et al., 2000; Gunilla, 2004). Of late, the *MYH3* gene was identified associated with the a^* value of meat by the genome-wide association studies (GWASs) (Cho et al., 2019).

China has more than 83 local pig breeds, and the meat quality of these local pig breeds, especially meat color, is better than Western commercial pigs, such as Landrace or Large White (Jiang et al., 2012; Lebret et al., 2015; Zhang et al., 2015). The Suhui pig is a new cross-bred lean-type pig breed containing 25% lineage of Huai pig and 75% lineage of Large White (Wang et al., 2019). The Huai pig is one of the local pigs in North China and is well-documented for its excellent meat quality and redder meat color,

while Large White is a commercial breed with a fast growth rate and poor meat quality (Yang et al., 2014; Liu et al., 2018). Briefly, after 23 years of artificial selection of the cross-bred offspring of the Large White and Huai pig, a new breed was developed, called the Xinhui pig, which contains 50% Huai pig and 50% Large White (1954–1977). Subsequently, Large White pigs were crossed with Xinhui pigs in 1998, and their offsprings were selected and bred for 12 years to obtain the Suhui pig (1998–2010). The Suhui pig is an excellent experimental population for identifying genes associated with meat color because there is phenotypic variation of meat color existent in Suhui pig population. Moreover, the Suhui pig's lineage contains Huai pig lineage and Large White lineage, and the meat color of the Huai pig is better than that of Large White. These two mixed lineages may result in the differentiation in the regions of the genome that affect the Suhui pig's meat color. This study aims to estimate the heritability and genetic correlation of meat color and identify the candidate genes and molecular markers of meat color in Suhui pigs, which will be beneficial for pig molecular breeding.

MATERIAL AND METHODS

Ethics Statement

All pigs were raised in accordance with the guidelines for the care and use of laboratory animals prepared by The Institute of Animal Welfare and Ethics Committee of Nanjing Agricultural University. All experimental schemes have been approved by the Animal Care and Use Committee of Nanjing Agricultural University (certificate no. SYXK (Su) 2017-0007).

Animals and Phenotype Measurements

Three-hundred and six Suhui pigs (227 sires and 79 dams) were used in this study. The Suhui pigs were all fed in three batches on the Huaiyin breeding farm (Huaian, China) under the same fodder and standard management environment. The animals were slaughtered in three batches on Jinyuan Meat Products Co., Ltd. (Huaian, China). The means and standard errors of slaughter age and carcass weight were 218.3 ± 1.09 (day) and 59.1 ± 0.39 (kg), respectively. After slaughter, ear tissue samples were gathered and stored in 75% alcohol solution, and *Longissimus dorsi* (LD) muscle samples were collected from the last rib of the left half carcasses and immediately stored at 4°C. CIELAB color space of meat color was evaluated by MiniScan EZ (HunterLab Corp., New York, USA) which was calibrated according to a standard white plate. The diameter aperture was 8 mm, and D65 illuminant and 0° standard observer angle were applied. The average of the CIELAB color space from three random positions on the surface of LD muscle samples at 45 min and 24 h after slaughter (L^* 45 min, L^* 24 h, a^* 45 min, a^* 24 h, b^* 45 min, and b^* 24 h) was used for subsequent analyses.

Genotyping and Quality Control

Genomic DNA was extracted from ear tissue samples following the standard phenol-chloroform method (Elder et al., 1983). All DNA samples were genotyped using the GeneSeek GGP-Porcine 80 K SNP BeadChip according to the manufacturer's protocol.

TABLE 1 | Significance of the fixed effects and covariant in the mixed model for the analysis.

Parameters	N	Fixed effects			Covariant	
		Sex	Batch	Season	Age	Cw
L* 45 min	306	NS	**	*	*	*
L* 24 h	306	NS	**	**	*	*
a* 45 min	306	NS	**	NS	NS	NS
a* 24 h	306	NS	**	NS	NS	NS
b* 45 min	306	NS	*	**	**	NS
b* 24 h	306	NS	**	*	*	NS

** $p < 0.05$ * $p < 0.01$

NS, non-significant.

Cw = carcass weight.

Genotype quality control was performed for selected SNPs by the PLINK 1.07 base on the follow criteria: SNP call rate $\geq 95\%$, minor allele frequency (MAF) $> 1\%$ and the p -value chi-square test of Hardy–Weinberg equilibrium $> 10^{-5}$ (Purcell et al., 2007). After the quality control and removing the SNPs from the sex chromosomes, 306 individuals and 52640 SNPs (Sus scrofa 11.1) were remained for subsequent analyses. The raw genotyped data of these 306 samples are available at <https://doi.org/10.6084/m9.figshare.16573700.v4>.

Statistics Analyses

The mixed linear model of SAS 9.4 software (SAS Institute, Inc., Cary, NC, USA) was used to fit the fixed effects and the covariates of each CIELAB color space parameter. The relationship matrix of individuals was built based on the marker genotype information developed by VanRaden (Vanraden, 2008). The additive genetic variance and residuals of CIELAB color space parameters were calculated using AI-REML arithmetic of DMU software (Madsen, 2006), and the genomic estimated breeding values (GEBVs) and residuals of each individual were estimated using the following model:

$$y = \mu + m + c + a + e,$$

where y is phenotypic observation, μ is overall mean, m is the fixed effect (L* and b* used batch and season as fixed effects; a* used batch as fixed effects), c is the covariates (L* used age and carcass weight as covariates; b* used age as covariates), a is random additive genetic effect of animal, and e is random residual error [$e \sim N(0, \sigma_e^2)$]

The covariance between CIELAB color space parameters was calculated using the multitrait model of DMU software. The heritability and genetic correlation between CIELAB color space parameters were calculated by the following formula:

$$h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2), r_{gxy} = \text{cov}_{gxy} / \sqrt{\sigma_{g^x}^2 * \sigma_{g^y}^2},$$

where h^2 is heritability, σ_a^2 is additive genetic variance, σ_e^2 is random residual variance, r_{gxy} is the genetic correlation of trait x and y , cov_{gxy} is genotype covariance of trait x and y , $\sigma_{g^x}^2$ is additive genetic variance of trait x , and $\sigma_{g^y}^2$ is additive genetic variance of trait y .

TABLE 2 | Descriptive statistics of meat color.

Parameters	N	Mean \pm SE	Max	Min	CV (%)	$h^2 \pm \text{SE}$
L* 45 min	306	40.07 \pm 0.22	56.30	32.52	9.62	0.20 \pm 0.10
L* 24 h	306	45.32 \pm 0.24	57.40	31.04	9.17	0.16 \pm 0.11
a* 45 min	306	5.14 \pm 0.09	9.22	1.32	31.32	0.30 \pm 0.12
a* 24 h	306	5.99 \pm 0.10	14.41	2.17	28.22	0.13 \pm 0.10
b* 45 min	306	11.62 \pm 0.08	15.67	8.12	11.81	0.29 \pm 0.11
b* 24 h	306	12.71 \pm 0.09	20.45	9.22	12.33	0.22 \pm 0.10

Genome-wide association studies for meat color were performed using a single-marker regression mixed linear model of Genome-wide Efficient Mixed-Model Association (GEMMA) software (Zhou and Stephens, 2012). The model is as follows:

$$Y = W\alpha + x\beta + \mu + \varepsilon; \mu \sim \text{MVN}_n(0, \lambda_T^{-1}K), \varepsilon \sim \text{MVN}_n(0, T^{-1/n}),$$

where Y is the vector of the corrected phenotype that is the sum of GEBV (genomic-estimated breeding value) and residuals of individuals. W is an matrix of fixed effects that is a column of 1, α is a vector of the corresponding coefficient including the intercept, x is a vector of marker genotypes, β is the effect size of SNP, μ is an vector of random effects, ε is an vector of errors, T^{-1} is the variance of the residual errors, λ is the ratio between the two variance components (genetic variance and environmental variance), K is a known relationship matrix which removed the SNPs in the same chromosome to avoid overfitting of the SNP effect on a chromosome, and MVN_n denotes the dimensional multivariate normal distribution (Zhou and Stephens, 2012).

The significance threshold of the test was corrected by the Bonferroni method for GWAS; the genome-wide significance threshold was defined as $0.05/N = 8.89 \times 10^{-7}$, and the suggestive significance threshold was defined as $1/N = 1.78 \times 10^{-5}$ (N = the number of SNPs using in GWAS, 52640) (Yang et al., 2005).

We sorted the individuals according to the GEBV for each meat color parameter (L* 45 min, L* 24 h, a* 45 min, a* 24 h, b* 45 min, and b* 24 h), and selected the highest and lowest 30 individuals for these six parameters. GENESOP 4.0 was used to calculate the F_{ST} statistic of each SNP for evaluating the degree of genetic differentiation in these groups (Rousset, 2008). The threshold of F_{ST} was 0.2.

Analysis of Gene Ontology and Metabolic Pathways

The SNPs that reached both thresholds of GWAS and F_{ST} tests were used as a collective for subsequent analysis. BioMart software was used to detect candidate genes in the 1-Mb region of theses SNPs up and downstream using the Ensembl database (Hou et al., 2016). Gene Ontology (GO) term annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed on the annotated genes using DAVID version 6.8 (Huang et al., 2007).

TABLE 3 | Genetic correlation \pm standard error between meat color.

Parameters	L* 45 min	L* 24 h	a* 45 min	a* 24 h	b* 45 min
L* 24 h	0.62 \pm 0.04	—	—	—	—
a* 45 min	-0.45 \pm 0.05	-0.14 \pm 0.06	—	—	—
a* 24 h	-0.47 \pm 0.05	-0.07 \pm 0.06	0.65 \pm 0.05	—	—
b* 45 min	-0.14 \pm 0.06	-0.43 \pm 0.04	0.62 \pm 0.04	0.35 \pm 0.05	—
b* 24 h	0.06 \pm 0.06	0.52 \pm 0.05	0.27 \pm 0.06	0.70 \pm 0.04	0.06 \pm 0.07

RESULTS

Description of Phenotypic and Genetic Parameters of Meat Color

The fixed effects and covariates of the mixed linear model for analyzing meat color were evaluated according to the significance of factors. As shown in **Table 1**, the batch showed an effect on L*, a*, and b*; season and age showed an effect on L* and b*; and carcass weight showed an effect on L*. Descriptive statistics and the heritability of CIELAB color space parameters are shown in **Table 2**. The heritability of L* 45 min, L* 24 h, a* 45 min, a* 24 h, b* 45 min, and b* 24 h was 0.20, 0.16, 0.30, 0.13, 0.29, and 0.22, respectively. The coefficient of variation of meat color ranges from 9.17% (L* 24 h) to 31.32% (a* 45 min). The genetic correlation of these parameters is shown in **Table 3**. Apart from b*, L* and a* showed a strong positive genetic correlation at two different time points (45 min and 24 h), which were 0.62 and 0.65, respectively. L* 45 min showed a weak negative genetic correlation with a* 45 min and a* 24 h, which are -0.45 and -0.47, respectively, but showed no genetic correlation with b*. Moreover, L* 24 h showed no genetic correlation with a* but showed genetic correlation with b* 45 min (-0.43) and b* 24 h (0.52). The genetic correlation of a* 45 min and b* were 0.62 (b* 45 min) and 0.27 (b* 24 h), respectively, and the genetic correlation of a* 24 h and b* were 0.35 (b* 45 min) and 0.70 (b* 24 h), respectively.

GWAS and F_{ST} Identified the SNPs Associated With Meat Color

The results of GWAS showed that there are 139 SNPs significantly associated with meat color, including 129 SNPs that reached the suggestive significance threshold (L* 45 min, 32 SNPs; L* 24 h, 5 SNPs; a* 45 min, 38 SNPs; a* 24 h, two SNPs; b* 45 min, 34 SNPs; and b* 24 h, 18 SNPs) and 10 SNPs that reached the genome-wide significance threshold (L* 45 min, six SNPs; a* 45 min, 1 SNP; b* 45 min, one SNP; and b* 24 h, two SNPs) (**Figure 1**, **Supplementary Table S1**). It is to be noted that 34 SNPs significantly associated with L* 45 min were located in a 5.17-Mb region on SSC11 (40.13–45.30 Mb); 22 SNPs significantly associated with a* 45 min were located in a 14.72-Mb region on SSC16 (20.32–35.02 Mb); six SNPs significantly associated with b* 45 min were located in a 4.22-Mb region on SSC13 (117.69–121.91 Mb); and 11 SNPs significantly associated

with b* 24 h were located in a 2.12-Mb region on SSC3 (57.52–59.64 Mb).

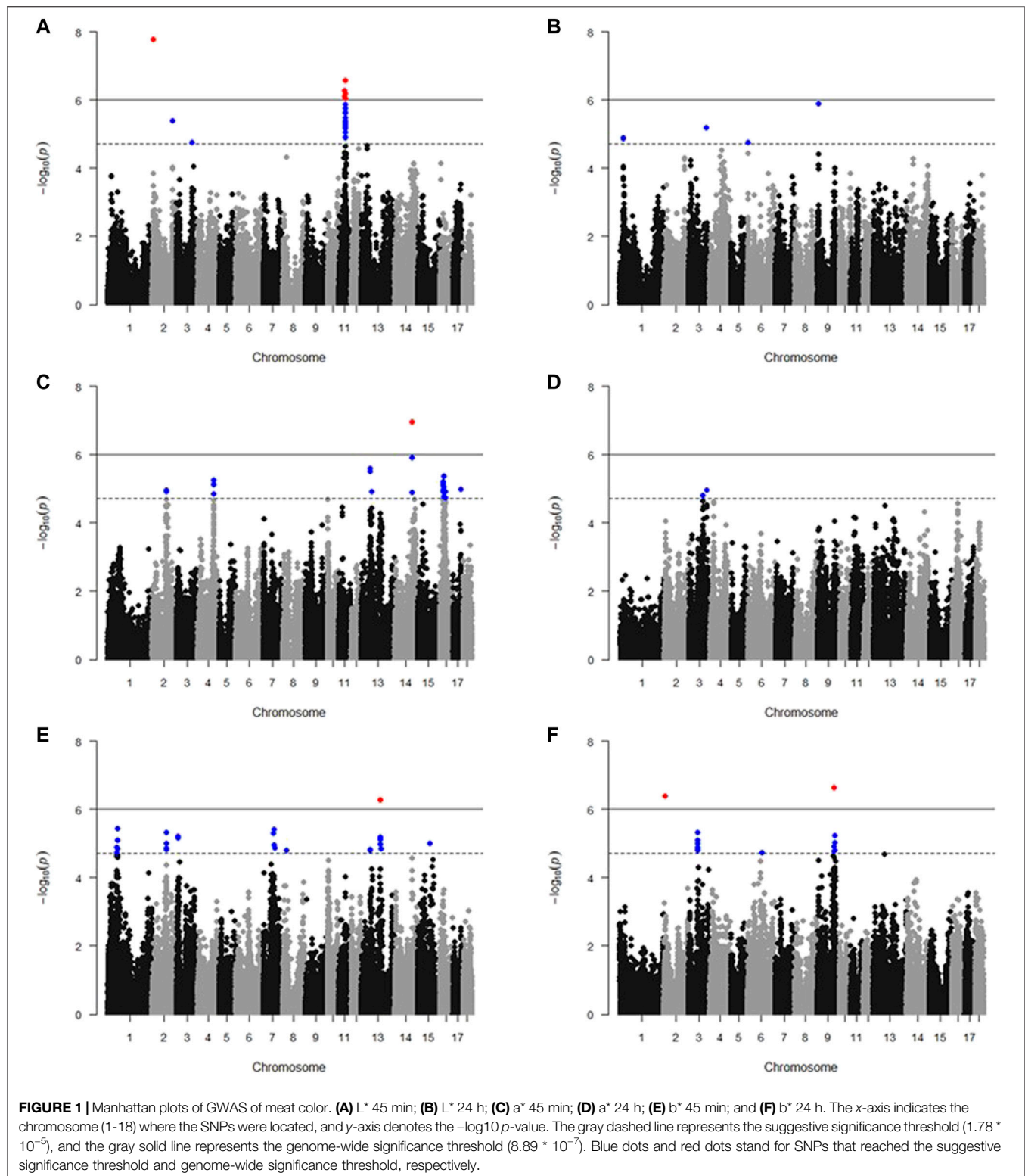
Genome-wide fixation coefficient (F_{ST}) values were calculated for each SNP between the highest and lowest individuals sorted by the GEBV for meat color. A large number of SNPs that reached the threshold (F_{ST} value >0.2) are shown in **Figure 2**. We focused on the overlapping results of GWAS and F_{ST} analyses. In total, 49 significant SNPs were overlapped in both GWAS and F_{ST} tests (**Supplementary Table S2**). Among them, 34 SNPs were identified associated with L* 45 min within a 5.17-Mb region on SSC11. Moreover, one, two, 10, and two SNPs were identified associated with L* 24 h, a* 45 min, b* 45 min, and b* 24 h, respectively.

Identify the Candidate Genes Associated With Meat Color

BioMart software was used to annotate the genes located within the upstream and downstream 1 Mb of significant SNPs, and 163 genes in total were identified (**Supplementary Table S3**). A total of 28 GO terms and six KEGG pathways were enriched by the DAVID platform (**Figure 3**). It is worth noting that five significant GO terms ($p < 0.05$) and one GO term which tends to be significant ($p = 0.0501$) are possibly relevant to meat color (**Table 4**). Six genes were identified in these terms that may affect meat color; a* 45 min (*HOMER1*), b* 45 min (*PIK3CA* and *VCAN*), b* 24 h (*FABP3* and *PIK3CG*), and L* 24 h (*FKBP1B*). These genes can be used as candidate genes of meat color in Suhui pigs. It is noted that most of the SNPs were located in intron and intergenic regions, except rs81361290, which is located in one of the exons of a non-coding transcript (**Supplementary Table S4**).

DISCUSSION

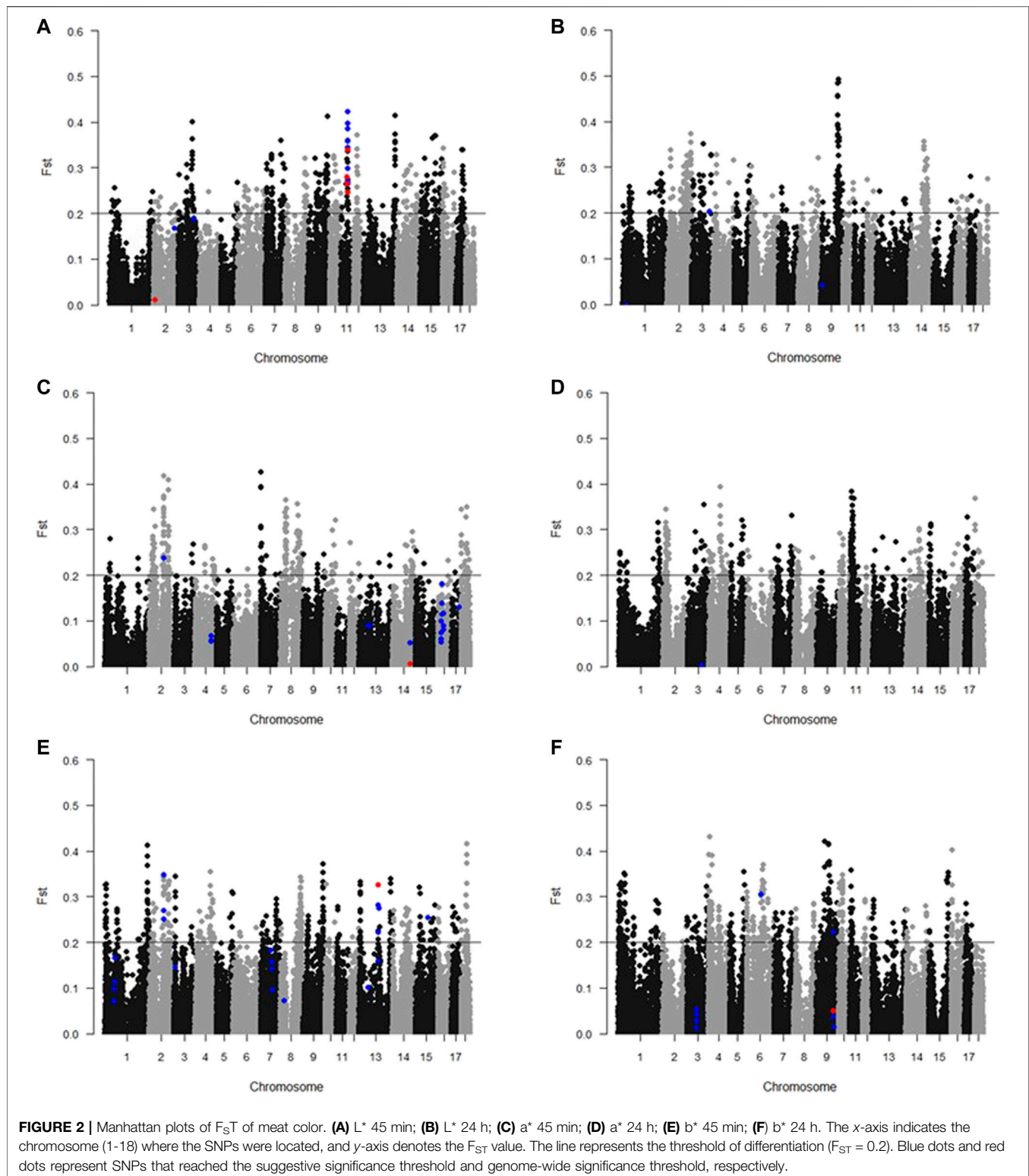
As a direct indicator of pork quality, meat color can significantly affect the economy of the meat market. In this study, the heritability and genetic correlation of meat color were calculated by DMU software, which provided the genetic theoretical basis for molecular breeding of meat color. In order to improve the accuracy and reliability of QTLs for meat color, GWAS and F_{ST} were used in this study, and the overlapping regions identified by these two methods were used to identify candidate genes of meat color (Tang et al., 2020). The GWAS identified the candidate loci by a mixed linear model,



and F_{ST} identified the candidate loci by detecting SNP differentiation between high and low groups according to GEBV. Through a combination of GWAS and F_{ST} tests, the candidate SNPs related to meat color were identified, and

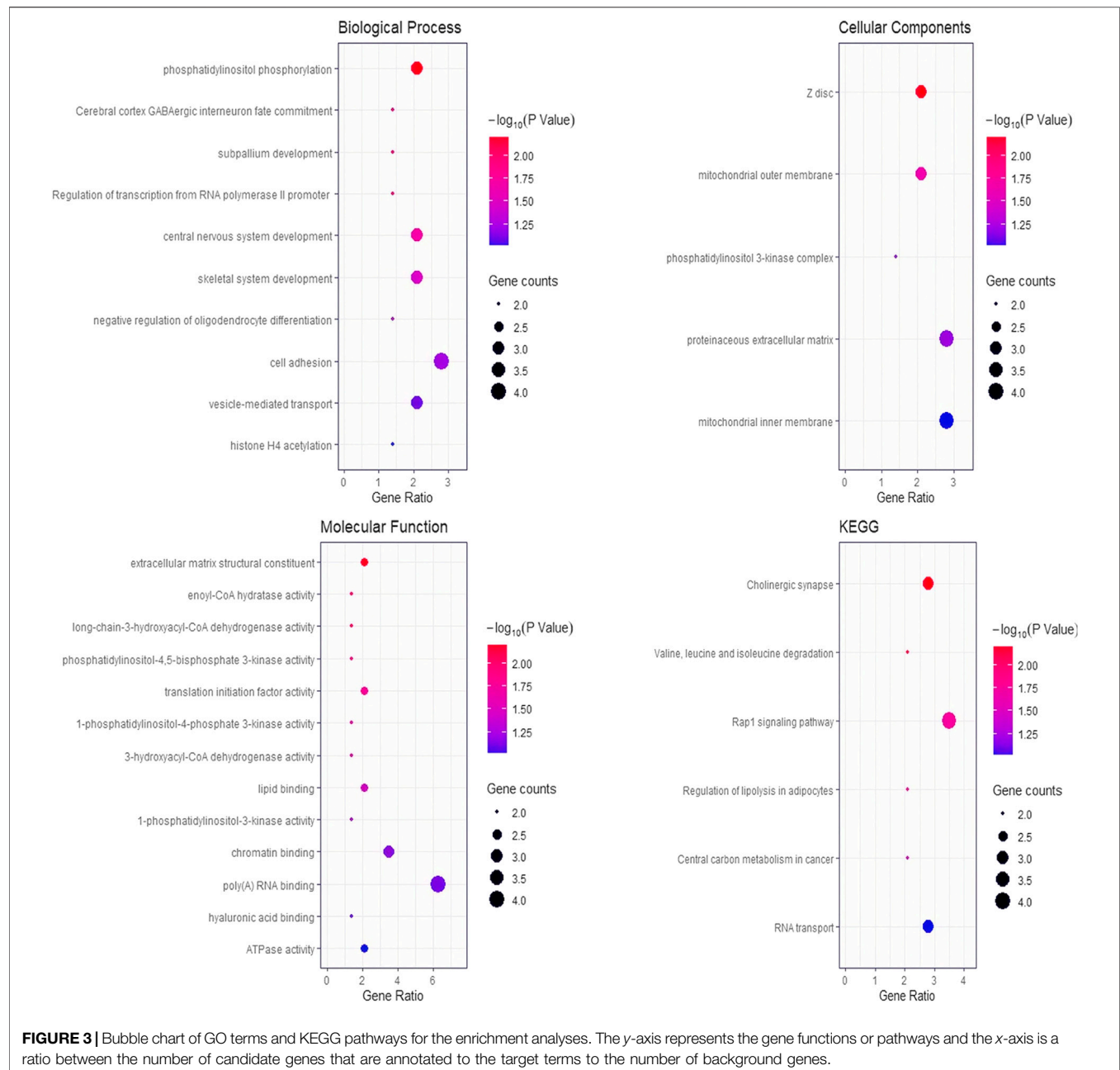
relevant functional candidate genes were detected by bioinformatic analysis.

The meat color at two different time points (45 min and 24 h) after slaughter was measured in this study, which represented the



meat color of fresh meat and chilled meat production with different economic values. For more effectively and accurately measuring meat color, the CIELAB color space was used. In this study, the effects affecting meat color are different, but the batch

showed significant effect on meat color that may be due to the difference of the environment. The season showed significant effect on L^* and b^* but has no significant effect on a^* . a^* is mainly related to the content and state of myoglobin, while L^*

**TABLE 4 |** Enrichment analysis results related with meat color.

Categories	Terms	p-value	Genes
GOTERM_BP_DIRECT	GO:0046854~phosphatidylinositol phosphorylation	0.0063	PIK3CG, PIK3CA, and EFR3B
GOTERM_BP_DIRECT	GO:0001501~skeletal system development	0.0343	HAPLN1, VCAN, and CHRD
GOTERM_CC_DIRECT	GO:0030018~Z disc	0.0463	SYNC, HOMER1, and FKBP1B
GOTERM_MF_DIRECT	GO:0046934~phosphatidylinositol-4,5-bisphosphate 3-kinase activity	0.0268	PIK3CG and PIK3CA
GOTERM_MF_DIRECT	GO:0035005~1-phosphatidylinositol-4-phosphate 3-kinase activity	0.0400	PIK3CG and PIK3CA
GOTERM_MF_DIRECT	GO:0008289~lipid binding	0.0501	PFN4, FABP3, and AP2M1

and b^* are greatly affected by the biochemical reaction of muscles which may be affected by temperature and humidity. It was reported that a^* was related to the proportion of muscle fiber types in the skeletal muscle (Kim, 2010). The value of a^* was relatively high when the skeletal muscle is dominated by slow-oxidative muscle fibers, which have a high content of myoglobin (Vierck et al., 2018). L^* was identified not only related to the proportion of fiber types in the muscle, but it is also related to the glycogen content and the ability of glycolysis in the muscle (Ryu et al., 2008). Meanwhile, L^* , especially L^* 24 h, was affected by the fiber structure in the muscle, which determines the light absorption and reflection ability of the meat (Hughes et al., 2020). Studies have reported that b^* could be affected by lipid (Ha et al., 2017). Meat color was affected by factors which were influenced by the storage environment; therefore, it can be seen that the heritability of meat color at 24 h was less than that at 45 min. The heritability of meat color ranges from 0.1 to 0.3, which belongs to low and middle heritability, and this results are consistent with other reports (Khanal et al., 2019).

We defined $|R| < 0.2$ as irrelevant, $0.2 < |R| < 0.5$ as weak correlation, and $|R| > 0.5$ as strong correlation. Our results showed that the genetic correlation of a^* (a^* 45 min and a^* 24 h) and L^* (L^* 45 min and L^* 24 h) is strong, but b^* 45 min and b^* 24 h have no genetic correlation with each other, indicating that the genetic background of b^* 45 min and b^* 24 h may be different. Our results showed that the main influencing factors for a^* 45 min and a^* 24 h are similar, whereas the main influencing factors of b^* 45 min and b^* 24 h could be different. There was a weak negative genetic correlation between L^* 45 min and a^* (a^* 45 min and a^* 24 h), which may be related to the proportion of muscle fiber types (Hughes et al., 2020). There is no genetic correlation between L^* 24 h and a^* (a^* 45 min and a^* 24 h), indicating that L^* 24 h may be more affected by other factors such as pH, water-holding capacity, and structure of muscle fibers etc. It is worth noting that L^* 24 h has a weak negative genetic correlation with b^* 45 min and strong positive genetic correlation with b^* 24 h, which indicated that the main influencing factors of b^* 45 min and b^* 24 h are different. The a^* and b^* showed strong positive genetic correlation at the same time point and weak positive genetic correlation at different time points, indicating that although b^* is complex, it may have the same genetic background with a^* . Indeed, it is noteworthy that the parameter of meat color could affect each other.

Among the meat color of Suhui pigs, the variation coefficient of a^* is the largest, which is over 30%, while the variation coefficient of L^* is over 9%. Therefore, SNPs and genes affecting meat color could be identified by GWAS in Suhui pigs. In order to reduce the false-positive rate and improve the power of the GWAS model and F_{ST} tests, we used GEBV plus residual as the corrected phenotype. In total, we identified 49 SNPs and both reached the significance threshold of F_{ST} value and GWAS, which could act as the candidate sites associated with meat color in this study. Interestingly, the parameter at the two time points after

slaughter did not share the same significant SNPs, which indicated that the main influencing factors of meat color at 45 min and 24 h after slaughter may be different from a genetic perspective. The meat color at 24 h after slaughter may be mainly affected by metabolic reactions in the muscle, such as glycolysis reaction of the muscle after slaughter; however, the meat color at 45 min after slaughter may be primarily determined by the content of muscle substances such as myoglobin, fat, and moisture etc. These SNPs were not overlapped with the previously reported QTL intervals related with meat color. Meat color is a complex economic trait which is regulated by complex genetic networks, and the genes causing the different meat color in different pig breeds may be located at different regulatory network nodes, which may be the reasons why the current study identified a few new associated genetic regions that were not identified by previous studies.

Although meat color was evaluated using different parameters (L^* , a^* , and b^*) at different time points (45 min and 24 h) after slaughter, the genetic correlation of meat color parameters range from -0.47 to 0.70 (Table 3). Therefore, genes within 1 Mb upstream and downstream of all significant sites were used as a collective for functional enrichment analyses. The results enriched multiple pathways, including muscle development (GO:0001501 and GO:0030018), phosphatidylinositol phosphorylation (GO:0046854, GO:0046934, and GO:0035005) and lipid binding (GO:0008289). Phosphatidylinositol is involved in a variety of physiological functions in the body, including muscle contraction, cell proliferation, and differentiation. The genes within the region on SSC11 (40.13–45.30 Mb) related to L^* 45 min were not enriched in any pathway and were not reported to affect meat color. It is possible that there is a regulatory element in this region that regulates the expression of downstream genes. In total, we identified six candidate genes in these pathways related to meat color. Of these candidate genes, only the *HOMER1* gene was associated with muscle development, and the rs81360833 ($p = 1.21E-05$) was suggestive to be significantly associated with a^* 45 min and was located in the region of the *HOMER1* gene. Homer1 is one of the homer family members that play a role in activity-dependent control of neuronal responses (Worley, 1998). As the scaffolding protein, the lack of Homer1 can cause the dysregulation of transient receptor potential (TRP) channels. It was reported that mice lacking Homer1 showed the decreasing of the muscle fiber cross-sectional area and skeletal muscle force generation, which may cause increasing spontaneous calcium influx (Michel et al., 2004; Stiber et al., 2008). The *HOMER1* gene has different expression patterns in the skeletal muscle of three different pig breeds, including Large White (lean-type), Tongcheng (obese-type), and Wuzhishan (mini-type) (Hou et al., 2016). These studies suggested that *HOMER1* may play an important regulatory role during skeletal muscle growth, which could affect the proportion of muscle fiber types in the skeletal muscle and resulted in different redness (a^*) of the skeletal muscle.

Four candidate genes associated with the b^* were identified, which were involved in the physiological function of fat deposition. The *PIK3CA* gene encoded the P110 α protein, which is a member of the enzyme phosphoinositide 3-kinase (PI3k) family and plays an important role in glucose metabolism, angiogenesis, and cellular growth. *PIK3CA* is a key mediator in insulin signaling, which can regulate glucose and lipid metabolism and the expression of major gluconeogenic-related genes (Sopasakis et al., 2010). The *PIK3CA* gene was differentially expressed in the two groups which were divided according to the degree of fat deposition in the muscle and enriched in the pathways related to the differentiation of adipose tissue (Cánovas et al., 2010). P110 γ , encoded by the *PIK3CG* gene, is the unique catalytic subunit of the PI3K family, and it is involved in the Akt pathway of glucose transport and fat production (Puig-Oliveras et al., 2014). Studies related to the *PIK3CG* gene were mainly focused on signal transduction of inflammation, and p110 γ is a major driver of metabolic diseases, such as fatty liver disease and type-2 diabetes (Van Greevenbroek et al., 2013). The *PIK3CG* gene has been identified as a candidate gene affecting intramuscular fat (IMF) and fatty acid (FA) in the swine muscle of Iberian X Landrace backcross animals (Puig-Oliveras et al., 2014). Versican (VCAN), is considered critical to several key cellular processes which may influenced the growth of adipose tissue, including cellular adhesion, proliferation, differentiation, migration, and angiogenesis (Du et al., 2011). It has been reported that the VCAN gene is associated with glucose tolerance in obese patients (Minchenko et al., 2013). The VCAN gene is associated with pork quality and fat deposition in pork (Piorkowska et al., 2018). Cardiac fatty acid-binding proteins (FABP3) participate in lipid metabolism by ingesting or utilizing long-chain fatty acids. An SNP located in *FABP3* promoter region was found in purebred Large White, Duroc, and Pietrain populations, which was identified related to adipogenesis (Sweeney et al., 2015). These four candidate genes (*PIK3CA*, *PIK3CG*, *VCAN*, and *FABP3*) have been reported to be involved in the regulation of fat metabolism pathways and affected the changes of fatty acid content and glycogen content in the muscle, which could be one of the reasons for the variation of yellowness (b^*).

Genes located in the region of L^* 45 min-associated SNPs were not enriched into any pathways; thus, further studies are needed to reveal the genetic basis for L^* 45 min in other pig breeds. The *FKBP1B* gene was identified near the significant SNP of L^* 24 h. *FKBP1B* is a member of the peptide-proline isomerase family and can be detected in a variety of cells. Studies have found that mir-34a mimic can regulate fat production by reducing the expression of *FKBP1B* mRNA in preadipocytes, indicating the importance of *FKBP1B* in fat production (Jang et al., 2015). In addition to muscle fiber types and the structure of the muscle fiber, L^* may be affected by *FKBP1B* through fat metabolism.

CONCLUSIONS

The a^* value of meat color has a large degree of variation in Suhuai pigs. The heritability of L^* 45 min, L^* 24 h, a^* 45 min, a^* 24 h, b^*

45 min, and b^* 24 h was 0.20, 0.16, 0.30, 0.13, 0.29, and 0.22, respectively. The genetic correlation between a^* (a^* 45 min and a^* 24 h) and L^* (L^* 45 min and L^* 24 h) is strong. Forty-nine potential meat color-related SNPs were identified using GWAS and F_{ST} tests in Suhuai pigs, and six candidate genes (*HOMER1*, *PIK3CG*, *PIK3CA*, *VCAN*, *FABP3*, and *FKBP1B*), which are functionally related to muscle development, phosphatidylinositol phosphorylation, and lipid binding, were detected around these significant SNPs. These findings provide theoretical and molecular basis for genetic improvement of meat color in pigs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The animal study was reviewed and approved by the Institute of Animal Welfare and Ethics Committee of Nanjing Agricultural University.

AUTHOR CONTRIBUTIONS

Conceptualization, PL, LH, and RH; formal analysis, HL and LH; investigation, HL, LH, WZ, BW, PH, PN, ZZ, and QL; methodology, PL, HL, and RH; project administration, PL, HL and RH; writing-original draft, HL, LH and PL; writing-review and editing, HL, LH and PL.

FUNDING

Jiangsu Agriculture Science and Technology Innovation Fund (JASTIF) (CX(20)1003), Fundamental Research Funds for the Central Universities (KJQN202129), Joint Funds of the National Natural Science Foundation of China (U1904115), Ministry of Agriculture and Rural Affairs Joint Projects for the National High Quality and Lean Pig Breeding (19210387), National Natural Science Foundation of China (32002149), Jiangsu seed industry revitalization project (JBGS[2021]024); National Key R and D Program sub-project (2021YFD1301101, 2021YFD1301105).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.768710/full#supplementary-material>

Supplementary Figure S1 | PCA analysis of experimental population.

REFERENCES

- Bertram, H. C., Petersen, J. S., and Andersen, H. J. (2000). Relationship between RN- Genotype and Drip Loss in Meat from Danish Pigs. *Meat Sci.* 56 (1), 49–55. doi:10.1016/S0309-1740(00)00018-8
- Borakhatariya, D. (2017). Genomic Selection in Dairy Cattles: a Review. *Int. J. Sci. Environ. Technology* 6 (1), 339–347. doi:10.20546/ijemas.2017.608.069
- Cabling, M. M., Kang, H. S., Lopez, B. M., Jang, M., Kim, H. S., Nam, K. C., et al. (2015). Estimation of Genetic Associations between Production and Meat Quality Traits in Duroc Pigs. *Asian Australas. J. Anim. Sci.* 28, 1061–1065. doi:10.5713/ajas.14.0783
- Cánovas, A., Quintanilla, R., Amills, M., and Pena, R. N. (2010). Muscle Transcriptomic Profiles in Pigs with Divergent Phenotypes for Fatness Traits. *BMC Genomics* 11 (372). doi:10.1186/1471-2164-11-372
- Cheng, J., Newcom, D. W., Schutz, M. M., Cui, Q., Li, B., Zhang, H., et al. (2018). Evaluation of Current United States Swine Selection Indexes and Indexes Designed for Chinese Pork Production. *The Prof. Anim. Scientist* 34 (5), 474–487. doi:10.15232/pas.2018-01731
- Cho, I.-C., Park, H.-B., Ahn, J. S., Han, S.-H., Lee, J.-B., Lim, H.-T., et al. (2019). A Functional Regulatory Variant of MYH3 Influences Muscle Fiber-type Composition and Intramuscular Fat Content in Pigs. *Plos Genet.* 15 (10), e1008279. doi:10.1371/journal.pgen.1008279
- Du, W. W., Yang, B. B., Yang, B. L., Deng, Z., Fang, L., Shan, S. W., et al. (2011). Versican G3 Domain Modulates Breast Cancer Cell Apoptosis: a Mechanism for Breast Cancer Cell Response to Chemotherapy and EGFR Therapy. *PLoS One* 6 (11), e26396. doi:10.1371/journal.pone.0026396
- Elder, R. T., Fritsch, F. E., and Maniatis, T. (1983). Cloning Techniques Molecular Cloning: A Laboratory Manual T. Maniatis E. R. Pritsch J. Sambrook. *BioScience* 33, 721–722. doi:10.2307/1309366
- Gunilla, L. (2004). A Second Mutant Allele (V199I) at the PRKAG3 (RN) Locus-II. Effect on Colour Characteristics of Pork Loin. *Meat Sci.* 66 (3), 621–627. doi:10.1016/S0309-1740(03)00180-3
- Ha, J., Kwon, S., Hwang, J. H., Park, D. H., Kim, T. W., Kang, D. G., et al. (2017). Squalene Epoxidase Plays a Critical Role in Determining Pig Meat Quality by Regulating Adipogenesis, Myogenesis, and ROS Scavengers. *Sci. Rep.* 7 (1), 16740–16749. doi:10.1038/s41598-017-16979-x
- Hou, X., Yang, Y., Zhu, S., Hua, C., Zhou, R., Mu, Y., et al. (2016). Comparison of Skeletal Muscle miRNA and mRNA Profiles Among Three Pig Breeds. *Mol. Genet. Genomics* 291, 559–573. doi:10.1007/s00438-015-1126-3
- Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., et al. (2007). DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists. *Nucleic Acids Res.* 35 (2), W169–W175. doi:10.1093/nar/gkm415
- Hughes, J. M., Clarke, F. M., Purslow, P. P., and Warner, R. D. (2020). Meat Color Is Determined Not Only by Chromatic Heme Pigments but Also by the Physical Structure and Achromatic Light Scattering Properties of the Muscle. *Compr. Rev. Food Sci. Food Saf.* 19 (1), 44–63. doi:10.1111/1541-4337.12509
- Jang, Y. J., Jung, C. H., Ahn, J., Gwon, S. Y., and Ha, T. Y. (2015). Shikonin Inhibits Adipogenic Differentiation via Regulation of Mir-34a-Fkbp1b. *Biochem. Biophysical Res. Commun.* 467, 941–947. doi:10.1016/j.bbrc.2015.10.039
- Jiang, Y. Z., Zhu, L., Tang, G. Q., Li, M. Z., Jiang, A. A., Cen, W. M., et al. (2012). Carcass and Meat Quality Traits of Four Commercial Pig Crossbreeds in China. *Genet. Mol. Res.* 11 (44), 4447–4455. doi:10.4238/2012.September.19.6
- Kim, G. D., Jeong, J. Y., Hur, S. J., Yang, H. S., and Joo, S. T. (2010). The Relationship between Meat Color (CIE L* and a*), Myoglobin Content, and Their Influence on Muscle Fiber Characteristics and Pork Quality. *Korean Journal for Food Science of Animal Resources* 30 (4), 626–633. doi:10.5851/kosfa.2010.30.4.626
- Katere, B., Wang, H. H., Lawing, J., Hao, N., Park, T., and Wetzstein, M. (2020). Toward Optimal Meat Consumption. *Am. J. Agric. Econ.* 102 (2), 662–680. doi:10.1002/ajae.12016
- Khanal, P., Maltecca, C., Schwab, C., Gray, K., and Tiezzi, F. (2019). Genetic Parameters of Meat Quality, Carcass Composition, and Growth Traits in Commercial Swine. *J. Anim. Sci.* 97 (9), 3669–3683. doi:10.1093/jas/skz247
- Küchenmeister, U., Kuhn, G., and Ender, K. (2000). Seasonal Effects on Ca2+ Transport of Sarcoplasmic Reticulum and on Meat Quality of Pigs with Different Malignant Hyperthermia Status. *Meat Sci.* 55, 239–245. doi:10.1016/S0309-1740(99)00149-7
- Lebret, B., Ecolan, P., Bonhomme, N., Méteau, K., and Prunier, A. (2015). Influence of Production System in Local and Conventional Pig Breeds on Stress Indicators at slaughter, Muscle and Meat Traits and Pork Eating Quality. *Animal* 9 (8), 1404–1413. doi:10.1017/S1751731115000609
- Liu, Y., Yang, X., Jing, X., He, X., Wang, L., Liu, Y., et al. (2018). Transcriptomics Analysis on Excellent Meat Quality Traits of Skeletal Muscles of the Chinese Indigenous Min Pig Compared with the Large white Breed. *Ijms* 19 (1), 21. doi:10.3390/ijms19010021
- Madsen, P. (2006). “DMU - A Package for Analyzing Multivariate Mixed Models in Quantitative Genetics and Genomics,” in *World Congress on Genetics Applied to Livestock Production*.
- Miar, Y., Plastow, G. S., Moore, S. S., Manafiazar, G., Charagu, P., Kemp, R. A., et al. (2014). Genetic and Phenotypic Parameters for Carcass and Meat Quality Traits in Commercial Crossbred Pigs1. *J. Anim. Sci.* 92 (7), 2869–2884. doi:10.2527/jas.2014-7685
- Michel, R. N., Dunn, S. E., and Chin, E. R. (2004). Calcineurin and Skeletal Muscle Growth. *Proc. Nutr. Soc.* 63 (2), 341–349. doi:10.1079/PNS2004362
- Minchenko, D., Ratushna, O., Bashta, Y., Herasymenko, R., and Minchenko, O. (2013). The Expression of TIMP1, TIMP2, VCAN, SPARC, CLEC3B and E2F1 in Subcutaneous Adipose Tissue of Obese Males and Glucose Intolerance. *CellBio* 02 (02), 45–53. doi:10.4236/cellbio.2013.22006
- Piórkowska, K., Żukowski, K., Ropka-Molik, K., Tyra, M., and Gurgul, A. (2018). A Comprehensive Transcriptome Analysis of Skeletal Muscles in Two Polish Pig Breeds Differing in Fat and Meat Quality Traits. *Genet. Mol. Biol.* 41 (1), 125–136. doi:10.1590/1678-4685-gmb-2016-0101
- Puig-Oliveras, A., Ramayo-Caldas, Y., Corominas, J., Estellé, J., Pérez-Montarelo, D., Hudson, N. J., et al. (2014). Differences in Muscle Transcriptome Among Pigs Phenotypically Extreme for Fatty Acid Composition. *PLoS One* 9 (6), e99720. doi:10.1371/journal.pone.0099720
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795
- Rousset, F. (2008). genepop'007: a Complete Re-implementation of the Genepop Software for Windows and Linux. *Mol. Ecol. Resour.* 8 (1), 103–106. doi:10.1111/j.1471-8286.2007.01931.x
- Ryu, Y. C., Choi, Y. M., Lee, S. H., Shin, H. G., Choe, J. H., Kim, J. M., et al. (2008). Comparing the Histochemical Characteristics and Meat Quality Traits of Different Pig Breeds. *Meat Sci.* 80 (2), 363–369. doi:10.1016/j.meatsci.2007.12.020
- Sellier, P. (1998). Genetics of Meat and Carcass Traits. *The Genet. pig*, 463–510.
- Sopasakis, V. R., Liu, P., Suzuki, R., Kondo, T., Winnay, J., Tran, T. T., et al. (2010). Specific Roles of the P110α Isoform of Phosphatidylinositol 3-Kinase in Hepatic Insulin Signaling and Metabolic Regulation. *Cel Metab.* 11 (3), 220–230. doi:10.1016/j.cmet.2010.02.002
- Stiber, J. A., Zhang, Z.-S., Burch, J., Eu, J. P., Zhang, S., Truskey, G. A., et al. (2008). Mice Lacking Homer 1 Exhibit a Skeletal Myopathy Characterized by Abnormal Transient Receptor Potential Channel Activity. *Mol. Cel Biol* 28 (8), 2637–2647. doi:10.1128/MCB.01601-07
- Suman, S. P., Hunt, M. C., Nair, M. N., and Rentfrow, G. (2014). Improving Beef Color Stability: Practical Strategies and Underlying Mechanisms. *Meat Sci.* 98 (3), 490–504. doi:10.1016/j.meatsci.2014.06.032
- Sweeney, T., O'halloran, A. M., Hamill, R. M., Davey, G. C., Gil, M., Southwood, O. I., et al. (2015). Novel Variation in the FABP3 Promoter and its Association with Fatness Traits in Pigs. *Meat Sci.* 100, 32–40. doi:10.1016/j.meatsci.2014.09.014
- Tang, Z., Fu, Y., Xu, J., Zhu, M., Li, X., Yu, M., et al. (2020). Discovery of Selection-driven Genetic Differences of Duroc, Landrace, and Yorkshire Pig Breeds by EigenGWAS and F St Analyses. *Anim. Genet.* 51 (4), 531–540. doi:10.1111/age.12946
- Tomasevic, I., Djekic, I., Font-i-Furnols, M., Terjung, N., and Lorenzo, J. M. (2021). Recent Advances in Meat Color Research. *Curr. Opin. Food Sci.* 41, 81–87. doi:10.1016/j.cofs.2021.02.012
- Van Greevenbroek, M. M., Schalkwijk, C. G., and Stehouwer, C. D. (2013). Obesity-associated Low-Grade Inflammation in Type 2 Diabetes Mellitus: Causes and Consequences. *Neth. J. Med.* 71, 174–187. doi:10.1007/s11845-013-0958-2

- Vanraden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980
- Vierck, K. R., O'Quinn, T. G., Noel, J. A., Houser, T. A., Boyle, E. A. E., and Gonzalez, J. M. (2018). Effects of Marbling Texture on Muscle Fiber and Collagen Characteristics. *Meat Muscle Biol.* 2 (1), 75–83. doi:10.22175/mmb2017.10.0054
- Visscher, P., and Haley, C. (1995). Utilizing Genetic Markers in Pig Breeding Programmes. *Anim. Breed. Abstr.* 63 (1), 1–8.
- Wang, B., Li, P., Zhou, W., Gao, C., Liu, H., Li, H., et al. (2019). Association of Twelve Candidate Gene Polymorphisms with the Intramuscular Fat Content and Average Backfat Thickness of Chinese Suhui Pigs. *Animals* 9 (11), 858. doi:10.3390/ani9110858
- Worley, P. F. (1998). Homer Regulates the Association of Group 1 Metabotropic Glutamate Receptors with Multivalent Complexes of homer-related, Synaptic Proteins. *Neuron* 21 (4), 707–716. doi:10.1016/S0896-6273(00)80588-7
- Wu, D., and Sun, D.-W. (2013). Colour Measurements by Computer Vision for Food Quality Control - A Review. *Trends Food Sci. Technology* 29 (1), 5–20. doi:10.1016/j.tifs.2012.08.004
- Wu, F., Vierck, K. R., Derouche, J. M., O'Quinn, T. G., Tokach, M. D., Goodband, R. D., et al. (2017). A Review of Heavy Weight Market Pigs: Status of Knowledge and Future Needs Assessment. *Anim. Sci.* 1 (1), 1–15. doi:10.2527/tas2016.0004
- Yang, J., Zhou, L., Liu, X., Ma, H., Xie, X., Xiong, X., et al. (2014). A Comparative Study of Meat Quality Traits between Laiwu and DLY Pigs. *Acta Veterinaria et Zootechnica Sinica* 45, 1752–1759.
- Yang, Q., Cui, J., Chazaro, I., Cupples, L. A., and Demissie, S. (2005). Power and Type I Error Rate of False Discovery Rate Approaches in Genome-wide Association Studies. *BMC Genet.* 6, 134–137. doi:10.1186/1471-2156-6-S1-S134
- Zhang, C., Luo, J. Q., Zheng, P., Yu, B., Huang, Z. Q., Mao, X. B., et al. (2015). Differential Expression of Lipid Metabolism-Related Genes and Myosin Heavy Chain Isoform Genes in Pig Muscle Tissue Leading to Different Meat Quality. *Animal* 9 (06), 1073–1080. doi:10.1017/S1751731115000324
- Zhou, X., and Stephens, M. (2012). Genome-wide Efficient Mixed-Model Analysis for Association Studies. *Nat. Genet.* 44 (7), 821–824. doi:10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Liu, Hou, Zhou, Wang, Han, Gao, Niu, Zhang, Li, Huang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Whole-Genome-Based Web Genomic Resource for Water Buffalo (*Bubalus bubalis*)

Aamir Khan^{1†}, Kalpana Singh^{1†}, Sarika Jaiswal¹, Mustafa Raza¹, Rahul Singh Jasrotia¹, Animesh Kumar¹, Anoop Kishor Singh Gurjar¹, Juli Kumari¹, Varij Nayan², Mir Asif Iquebal^{1*}, U. B. Angadi¹, Anil Rai¹, Tirtha Kumar Datta² and Dinesh Kumar¹

¹Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, ²ICAR-Central Institute for Research on Buffaloes, Hisar, India

OPEN ACCESS

Edited by:

Basharat Ahmad Bhat,
University of Otago, New Zealand

Reviewed by:

Qianjun Zhao,
Institute of Animal Sciences (CAAS),
China
Parveen Kumar,
Lund University, Sweden

*Correspondence:

Mir Asif Iquebal
ma.iquebal@icar.gov.in

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 05 November 2021

Accepted: 14 February 2022

Published: 11 April 2022

Citation:

Khan A, Singh K, Jaiswal S, Raza M,
Jasrotia RS, Kumar A, Gurjar AKS,
Kumari J, Nayan V, Iquebal MA,
Angadi UB, Rai A, Datta TK and
Kumar D (2022) Whole-Genome-
Based Web Genomic Resource for
Water Buffalo (*Bubalus bubalis*).
Front. Genet. 13:809741.
doi: 10.3389/fgene.2022.809741

Water buffalo (*Bubalus bubalis*), belonging to the *Bovidae* family, is an economically important animal as it is the major source of milk, meat, and drought in numerous countries. It is mainly distributed in tropical and subtropical regions with a global population of approximately 202 million. The advent of low cost and rapid sequencing technologies has opened a new vista for global buffalo researchers. In this study, we utilized the genomic data of five commercially important buffalo breeds, distributed globally, namely, Mediterranean, Egyptian, Bangladesh, Jaffarabadi, and Murrah. Since there is no whole-genome sequence analysis of these five distinct buffalo breeds, which represent a highly diverse ecosystem, we made an attempt for the same. We report the first comprehensive, holistic, and user-friendly web genomic resource of buffalo (*BuffGR*) accessible at <http://backlin.cabgrid.res.in/buffgr/>, that catalogues 6028881 SNPs and 613403 InDels extracted from a set of 31 buffalo tissues. We found a total of 7727122 SNPs and 634124 InDels distributed in four breeds of buffalo (Murrah, Bangladesh, Jaffarabadi, and Egyptian) with reference to the Mediterranean breed. It also houses 4504691 SSR markers from all the breeds along with 1458 unique circRNAs, 37712 lncRNAs, and 938 miRNAs. This comprehensive web resource can be widely used by buffalo researchers across the globe for use of markers in marker trait association, genetic diversity among the different breeds of buffalo, use of ncRNAs as regulatory molecules, post-transcriptional regulations, and role in various diseases/stresses. These SNPs and InDels can also be used as biomarkers to address adulteration and traceability. This resource can also be useful in buffalo improvement programs and disease/breed management.

Keywords: bovine, lncRNA, miRNA, molecular markers, web-resource, CircRNAs

INTRODUCTION

Water buffalo, scientifically known as *Bubalus bubalis*, is the major source of milk, meat, and drought in various countries, making it an economically important animal. This livestock species belonging to the *Bovidae* family is mainly distributed in tropical and subtropical regions. Based on morphology and behavior, the two categories of domestic Asian water buffalo are river buffalo (2n = 50) and swamp buffalo (2n = 48) (Iannuzzi, 1994). The global

population of water buffalo is ~217 million in 34 countries (FAOSTAT, 2020) with ~82% and ~18% river buffalo and swamp buffalo, respectively. South Asia holds the majority of water buffalo as India ranks first in buffalo breeding with a share of 50.5%, followed by Pakistan and China (Taşcıoğlu et al., 2020). Buffalo are largely domesticated by small farmers in Asia. This indicates the popularity and dependence of water buffalo as compared to any other species that are domesticated. India has the lion's share (69%) in river buffalo. Milk yield is more in buffalo than cattle. Also, buffalo milk has a higher nutritional value than cattle on account of higher fat content (8.0%), higher unsaturated fatty acid levels, higher protein content (4.5%), and lower phospholipid and cholesterol levels. It is a more preferred milk for dairy products (Du et al., 2019).

The fast decline in DNA sequencing costs has paved the way for researchers across the globe to revolutionize genome analysis. Assembly and decoding of the genome of water buffalo is in continuous progress. The following five genome assemblies of water buffalo exist on the NCBI database (<https://www.ncbi.nlm.nih.gov/assembly/?term=Bubalus>) (Last accessed: July 2021): GCA_003121395.1 of the Mediterranean breed from University of Adelaide, Australia; GCA_019923935.1 of the Murrah breed from National Dairy Development Board, India; GCA_004794615.1 of the Bangladesh breed from BGI-Shenzhen, China; GCA_002993835.1 of the Egyptian buffalo breed from Agriculture Genetic Engineering Research Institute and Nile University, Egypt, and GCA_000180995.3 of the Jafarabadi breed from Anand Agricultural University, India. GCA_003121395.1 and GCA_019923935.1 are chromosome level assemblies with 25 chromosomes, however, RefSeq annotation has not yet been provided for GCA_019923935.1.

Whole-genome sequencing and transcriptome studies provide insights on genetic makeup, numerous trait markers, and their expression in organisms. Simple sequence repeats (SSR) are the information source for genetic diversity among different breeds/varieties of the same species (Patzak et al., 2012). There are 22 buffalo breeds (only river subspecies) distributed all over the world with different characteristics like shape, size, color, weight, and lactation period, etc. Genomic variation results in single nucleotide polymorphisms (SNPs), insertions, and deletions (Surya et al., 2018). These variations are stable and are transferred from one generation to the next. These variations impinge start codon gain or loss, stop codon gain or loss, or frame shift. The presence of such variations in protein coding regions culminates in synonymous or non-synonymous amino acid replacement.

Long non-coding RNAs are a group of RNAs which are greater than 200 nt and lack open reading frames or have <100 amino acids in length. lncRNAs regulate gene expression through methylation and demethylation (Bhat and Jones 2016; Fernandes et al., 2019) and through chromatin modifications by interfering with transcription factors [binding with DNA and regulating transcription (Griffiths et al., 2000)] and miRNAs. lncRNAs perform post-translational regulation through capping, alternative

splicing, editing, transport, translation, degradation, and stability of mRNA targets. Apart from their biological roles, lncRNAs can also function as biomarkers. At the organism level, lncRNAs are known to be abnormally expressed in many diseases therefore playing a role in diagnosis (Kosinska-Selbi et al., 2020).

miRNAs are 18–25 nucleotide-long regulatory sequences, which play an important role in response reactions during an organism's exposure to biotic or abiotic conditions (O'Brien et al., 2018). They regulate gene expression by binding to the target sequence with the help of AGO protein and make an miRNA-induced silencing complex (mi-RISC) (Kawamata and Tomari, 2010). Water buffalo are adapted to higher to lower altitudes, hence they face a wide range of stresses like low/high temperatures (Liu et al., 2019), pathogens (Dhanoo et al., 2019; Lecchi et al., 2019), etc. Previously known miRNAs specific to buffalo have been reported from various transcriptome studies involving such stress conditions (Dhanoo et al., 2019; Lecchi et al., 2019; Liu et al., 2019).

Other regulatory non-coding RNAs, known as circular RNAs (circRNAs), spawn through back-splicing of RNAs. They are more stable than RNAs (Chen et al., 2017; Wang et al., 2017). The functions of circRNAs are not well known but still it is reported that they play a significant role in post-transcriptional regulation of gene expression (Lukiw, 2013). CircRNAs function as a sponge of miRNAs by sequestering them by binding and interacting with lncRNAs (Lei et al., 2021). These are being employed as biomarkers for controlling and treating diseases (Meng et al., 2017; Lu, 2020).

Before release of the buffalo reference genome, most of the studies related to buffalo involving omics analyses were based on the *Bos taurus* reference genome. The available whole-genome assemblies of five buffalo breeds represent a highly diverse ecosystem. Their utilization in whole-genome sequence analyses and in extraction of rapid polymorphic markers at lower costs for the breeders is warranted. In 2018, a buffalo reference genome with 24 chromosomes along with X and MT chromosomes was released by the Italian Buffalo Genome Consortium (https://www.ncbi.nlm.nih.gov/assembly/GCA_003121395.1). For the current study, the different omics studies in buffalo were performed using the GCA_003121395.1 buffalo reference genome to extract non-coding RNAs such as miRNAs, lncRNAs, and circRNAs in the 31 buffalo tissues, which had not been attempted earlier. Also, the various genetic markers such as SSRs, SNPs, and InDels from five breeds of buffalo (Mediterranean, Egyptian, Bangladesh, Jaffarabadi and Murrah) were mined. After extraction of the mentioned molecular markers and non-coding RNAs, a web-based genomic resource, *BuffGR* was developed to facilitate the buffalo research community with user-friendly, single-window retrieval of buffalo omics data to be utilized for further scientific research and studies. This buffalo web resource is state-of-the-art, holistic, and currently the largest collection related to buffalo including the most important breed of India, i.e., Murrah from the latest 2021 assembly as well as the world, i.e., Mediterranean from the latest 2018 assembly.

TABLE 1 | The list of assemblies of buffalo from public domain.

Accession	Breed	Submitter	Assembly level	Remarks
GCA_003121395.1	Mediterranean	University of Adelaide	Chromosome-wise	UOA_WB_1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_003121395.1/)
GCA_019923935.1	Murrah	National Dairy Development Board, India	Chromosome-wise	NDDDB_SH_1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_019923935.1/)
GCA_004794615.1	Bangladesh	BGI-Shenzhen	Scaffold level	Bubbub1.0 (https://www.ncbi.nlm.nih.gov/assembly/GCA_004794615.1/)
GCA_002993835.1	Egyptian	Egyptian Water Buffalo Genome Consortium (Agriculture Genetic Engineering Research Institute and Nile University)	Scaffold level	ASM299383v1 (https://www.ncbi.nlm.nih.gov/assembly/GCA_002993835.1/)
GCA_000180995.3	Jaffrabadi	Anand Agricultural University, Anand, Gujarat, India	Scaffold level	Bubalus_bubalis_Jaffrabadi_v3.0 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000180995.3/)

TABLE 2 | The details of RNA-seq data from the International Water Buffalo Genome Project representing different buffalo tissues along with SRA IDs and mapping %.

Tissue	SRA IDs	Mapping %	Tissue	SRA IDs	Mapping %
Tongue	ERR315616	95.71	Ovary-corpora luteum	ERR315632	94.30
Rumen	ERR315617	93.69	Ovary follicle	ERR315633	97.60
Abomasum	ERR315618	95.91	Oviduct	ERR315634	96.67
Small intestine	ERR315619	93.88	Endometrium	ERR315635	96.59
Large intestine	ERR315620	96.03	Mammary gland	ERR315636	95.18
Obex	ERR315621	94.02	Embryo pool	ERR315637	70.87
Hypophysis	ERR315622	96.84	Embryo single	ERR315638	73.61
Spinal Cord	ERR315623	95.40	Thymus	ERR315639	96.71
WBC	ERR315624	97.04	Mesenteric lymph node	ERR315640	96.47
Cerebellum	ERR315625	90.61	Spleen	ERR315641	96.07
Bone Marrow	ERR315626	95.55	Liver	ERR315642	96.57
Muscle longissimus dorsai	ERR315627	96.21	Pancreas	ERR315643	96.70
Muscle semitendinosus	ERR315628	96.62	Kidney	ERR315644	95.23
Testis	ERR315629	97.40	Lung	ERR315645	96.53
Thyroid	ERR315630	96.19	Testis	SRR527266-72	90.02
Heart	ERR315631	94.68	Milk	SRR7091387-98	94.88

MATERIALS AND METHODS

Data Retrieval and Processing

In order to extract the breed-wise molecular markers and variants like SSRs, SNPs, and InDels, in five buffalo breeds, *namely*, Mediterranean, Egyptian, Bangladesh, Jaffrabadi, and Murrah, their genome assemblies were retrieved from NCBI (**Table 1**).

For extraction of cirRNAs, SNPs, and InDels, RNA-seq data of a total of 31 buffalo tissues were retrieved from NCBI, which were mapped with the GCF_003121435.1 genome assembly of the Mediterranean breed using Bowtie2 (Langmead and Salzberg, 2012), while HISAT2 (Kim et al., 2019) was used in the case of lncRNAs (**Table 2**). For the extraction of miRNAs, cirRNAs, and lncRNAs, the genome assembly of the Mediterranean breed was used (GCF_003121435.1).

Identification of SNPs and InDels

For extraction of variants, *namely*, SNPs and InDels, the four buffalo breeds (Murrah, Jaffrabadi, Bangladesh, and Egyptian) were mapped to the water buffalo reference genome of the Mediterranean breed (GCA_003121395.1, the UOA_WB_1 assembly). These mapped reads of RNA-seq data were first

sorted and indexed using Samtools (Li et al., 2009; Li, 2011) along with the indexed reference genome (GCA_003121395.1). Then, coverage extraction of each nucleotide was performed using Samtools *mpileup*. Further, SNPs and InDels were extracted using bcftools (Danecek et al., 2021) *call*. Finally, significant SNPs were filtered using bcftools *view* at p-value <0.05, read depth >10, quality depth >30, minimum root mean square mapping >40, and flanking sequence length =50. This was followed by functional annotation of extracted SNPs and InDels using Perl script utilizing the annotation file of the genome of Mediterranean buffalo (GCA_003121395.1).

Identification of SSRs Markers

MicroSatellite (MISA) (Beier et al., 2017) was used to extract SSRs from genome assemblies of all the five breeds utilizing parameters such as ≥ 10 , ≥ 6 , ≥ 5 , ≥ 4 , and ≥ 4 repeats for mono, di, tri, tetra, and penta nucleotide (nt) motifs, respectively along with length of compound SSRs ≤ 100 nt and minimum distance between two SSRs ≥ 50 nt (Zhao et al., 2017). The functional annotation of mined SSR markers was performed using Perl scripts utilizing the annotation of the Mediterranean buffalo RefSeq genome (GCA_003121395.1). Finally, based on the

result of MISA, primer3 software (Untergasser et al., 2012) was used to design the primer pairs at default parameters, taking the flanking sequences of SSRs of the Mediterranean breed.

Identification of microRNAs

For the prediction of miRNAs, first-known miRNAs and pre-miRNAs of *Bos taurus* from miRBase (Griffiths-Jones et al., 2006) were collected and duplicates were removed using CD-HIT (Huang et al., 2010). The pre-miRNA sequences of non-redundant *Bos taurus* miRNAs were aligned with the buffalo RefSeq genome (GCA_003121395.1) using BLASTn and sequences with 0 gap and ≤ 3 mismatches were taken along with 500 nt up and downstream stretches, making these >1000 nt length sequences (Altschul et al., 1990). Further, 200 nt fragments were taken from these sequences by using 25 nt sliding windows using the SegKit tool (Shen et al., 2016). The obtained sequences were again clustered using CD-HIT to obtain non-redundant sequences. Non-redundant sequences were used to predict the secondary structure by RNAfold (Lorenz et al., 2011) at minimum free energy (MFE) > -20. Further, sequences with <60 nt, non-AUGC, and multi-loop in structure, and pseudo pre-miRNAs were removed by Triplet-SVM classifier (Xue et al., 2005). These putative pre-miRNAs were used for further prediction of mature miRNAs using MiRdup (Leclercq et al., 2013). Finally, psRANTarget (Dai and Zhao, 2011) was used at an expectation value of 2 to predict mRNA targets of predicted miRNAs.

Identification of Circular RNAs

For the identification of circRNAs from the mapped RNA-seq reads of 31 buffalo tissues, CIRI v2.0.4 (Gao et al., 2015) was used. As circRNA-looping sites cannot be aligned directly to the genome, find_circ (Memczak et al., 2013) was used for the first 20 base pairs of each read end that were incompatible with the genome to anchor independent reads, thus map them with the buffalo reference genome (GCA_003121395.1), and finally to find only the mapped site. If the two anchors aligned in the linear region were in the reverse direction, anchor reads were extended until circRNA junctions were found. The sequence was considered a circRNA if the two sides of sequences corresponded to GT/AG splicing signals as mentioned by Fu et al. (2018). CIRI was also used to annotate circRNAs by using the annotation file of the GCF_003121395.1 genome assembly.

Identification of Long Non-Coding RNAs

For the identification of lncRNAs, from RNA-seq data of 31 buffalo tissues, first, mapping was performed using HISAT2 (Kim et al., 2019), followed by assembly using Stringtie v1.3.5 (Pertea et al., 2015). Then, putative lncRNAs were predicted from assembled reads using CPC2 (Kang et al., 2017) and passed through subsequent steps (a and b) for further validation as non-coding transcripts, i.e., 1) the transcripts with length ≥ 200 bp, open reading frame (ORF) ≤ 100 aa, strand information (+/-strand), and CPC2 score <0.5 were selected using OrfPredictor (Min et al., 2005) and passed through annotation using the annotation file of the GCF_003121395.1 genome assembly by

GffCompare (Burset and Guigo, 1996). 2) These were then searched against the NCBI-nr protein database through blastx (E value 0.01, coverage >80%, and identity >90%) and the Pfam protein database through HMMER (Finn et al., 2011). Finally, the validated lncRNAs were classified based on origin of lncRNAs as *i* (within a reference intron), *j* (alternative lncRNAs isoforms of known genes), *o* (lncRNAs with exonic overlap with a known transcript), *u* (intergenic lncRNAs), and *x* (exonic overlap on the opposite stand) as classified by Roberts et al. (2011). Transcripts with FPKM ≥ 0.5 for multi-exon transcripts and FPKM ≥ 1 for single-exon transcripts were selected as lncRNAs.

Development of Buffalo Web Genomic Resource, BuffGR

The Buffalo Genomic Resource Database, BuffGR is a 'three tier architecture' relational database developed using client, server, and database tiers. The analyzed datasets were catalogued in BuffGR on a Linux server. The following steps were involved in the development of BuffGR (Figure 1A): 1) Extraction of SNPs/InDels, SSR markers, lncRNAs, miRNAs, and circRNAs from the reference genomes of different breeds of buffalo and SRA data of 31 tissues of buffalo. These data are absolute, rather than having relative quantification. 2) Development of relational database in MySQL version 10.4.17, which includes 11 tables for all the fields, namely, for SNPs/InDels, SSR markers, lncRNAs, miRNAs, and circRNAs (Figure 1B); 3) development of web interface in PHP, HTML, and Java. Web hosting of this interface was done by Apache2 server version 3.2.4. A request was sent to the web server from the user's system in PHP. A query was generated following the request of the user on the web server and sent to MySQL. The database response was prepared in MySQL and sent back to the web server. Finally, a response prepared in PHP was displayed in the user's system.

RESULTS

Identification of SNPs and InDels

A total of 6028881 SNPs and 613403 InDels were extracted from the set of 31 buffalo tissues. The highest number of SNPs and InDels was extracted from milk tissue (1625901 SNPs/174256 InDels) followed by testis (448640 SNPs/46172 InDels) and large intestine (152608 SNPs/17552 InDels) (Figure 2A). However, the variants detected breed-wise showed a maximum number of SNPs and InDels in the Murrah breed (6313245 SNPs/510515 InDels), followed by Bangladesh (906446 SNPs/114319 InDels) and Egyptian (447224 SNPs/5920 InDels), while the least was seen in Jaffarabadi (60207 SNPs/3370 InDels) (Figure 2B). Table 3 represents the extracted tissue-wise genes showing abundance of SNPs and InDels by functional annotation. A total of 7727122 SNPs and 634124 InDels were collectively distributed in the four breeds of buffalo (Murrah, Bangladesh, Jaffarabadi, and Egyptian) with reference to the Mediterranean breed. From functional annotation of breed-wise SNP/InDels, 12326/8469, 15152/2044, 4798/1100, and 21762/17222 genes were found to have abundance of SNPs/InDels in Bangladesh,

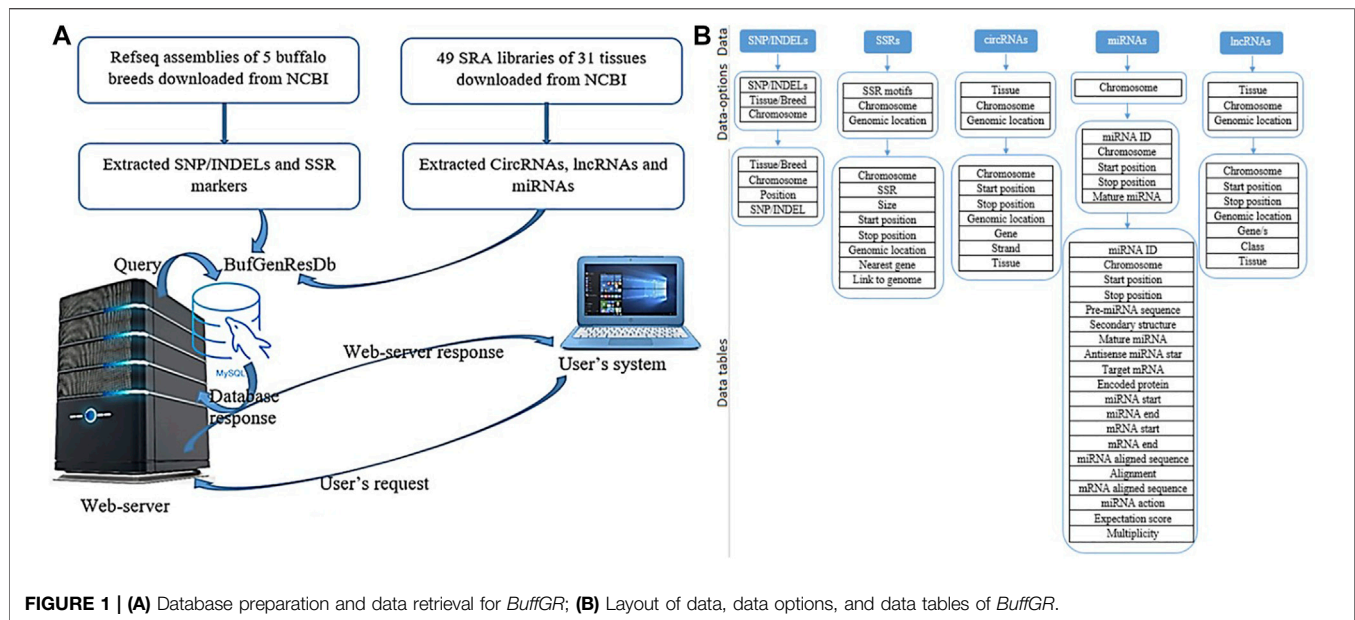


FIGURE 1 | (A) Database preparation and data retrieval for *BuffGR*; **(B)** Layout of data, data options, and data tables of *BuffGR*.

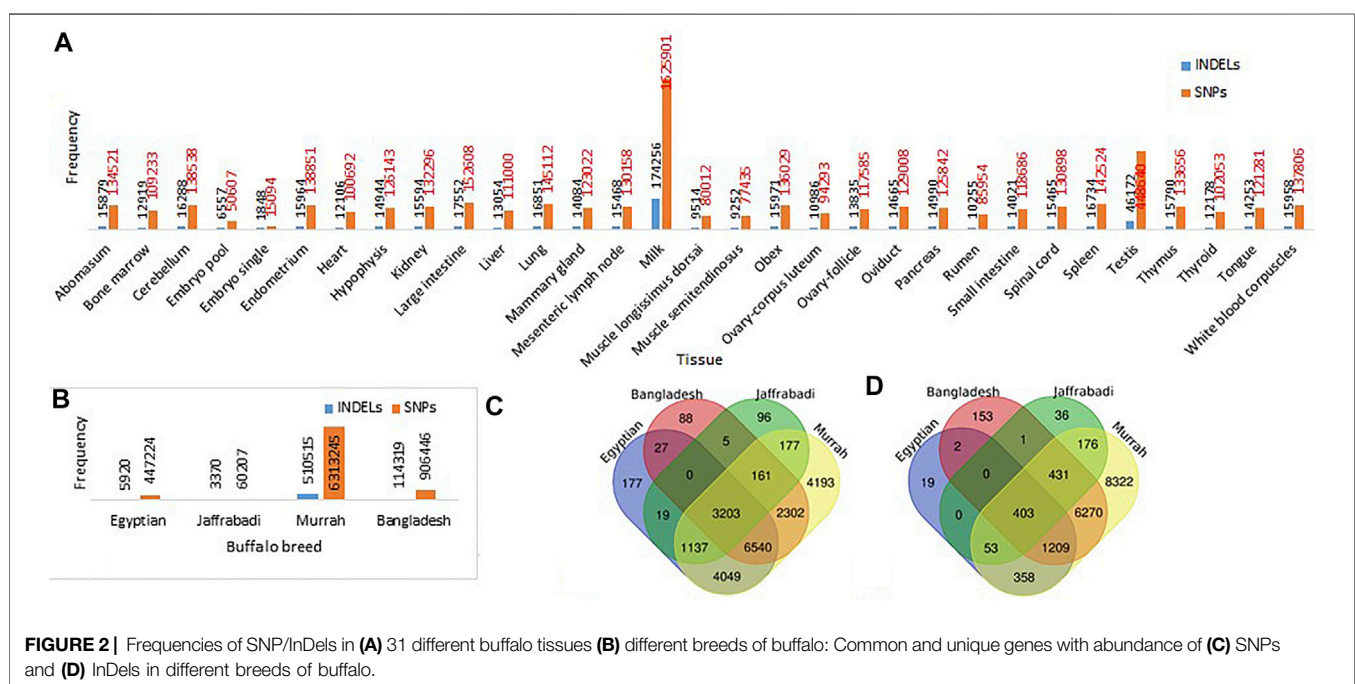


FIGURE 2 | Frequencies of SNP/InDels in **(A)** 31 different buffalo tissues **(B)** different breeds of buffalo: Common and unique genes with abundance of **(C)** SNPs and **(D)** InDels in different breeds of buffalo.

Egyptian, Jaffrabadi, and Murrah breeds, respectively (**Figure 2C** for SNPs and **Figure 2D** for InDels). SNP discovery plays an important role in obtaining varying alleles associated with different traits of interest (Mishra et al., 2020). This can be useful in marker trait association studies for various traits (Pareek et al., 2008).

A total of 12 genes (SPP1: chr7, SCD: chr23, SREBF1: chr3, STAT1: chr2, TG: chr15, LALBA: chr4, INSIG2: chr2, GHRL: chr21, DGAT1: chr15, CSN1S1: chr7, BTN1A1: chr2, ADRA1A: chr3) with abundance of milk tissue SNPs from the present study

were found to be common out of 19 candidate genes reported to be associated with milk production trait by Du et al. (2019) (**Table 4**). We also found 10 genes (COL1A2, APOB, GDF7, KLHL29, NRXN1, RGS22, VPS13B, MFSD14A, SLC35A3, PALMD) with abundance of SNPs of different breeds from the present study to be common out of 12 candidate genes for different QTL traits such as milk yield, fat yield, protein yield, fat %, and protein % identified from GWAS analysis of Italian Mediterranean buffalo using the SNP-ChIP technique by Iamartino et al. (2017) and Liu et al. (2017) (**Table 4**).

TABLE 3 | Annotated genes with abundance of extracted SNP/InDels from buffalo tissues.

Tissue	Genes with SNPs	Genes with InDels	Tissue	Genes with SNPs	Genes with InDels
Tongue	14392	6944	Muscle longissimus dorsai	12381	5149
Rumen	13503	5751	Muscle semitendinosus	12428	5038
Obex	15038	7404	Small intestine	14322	6867
WBC	13422	6813	Large intestine	15438	7879
Testis	16121	8588	Ovary-corpus luteum	13538	5821
Thyroid	14227	6429	Ovary follicle	14208	6751
Heart	13279	6125	Cerebellum	14711	7385
Thymus	14602	7179	Endometrium	14882	7376
Oviduct	14728	7109	Mesenteric lymph node	14445	7189
Spleen	14629	7479	Mammary gland	14674	7052
Liver	13969	6593	Spinal cord	14583	7229
Pancreas	14620	7128	Bone marrow	13376	6252
Kidney	14726	7303	Embryo pool	9531	3510
Lung	14989	7600	Embryo single	6008	1338
Testis	16121	8588	Hypophysis	14763	7144
Milk	16090	9308	Abomasum	14877	7514

TABLE 4 | Genes with abundance of extracted tissue/breed SNPs found to be common within the reported candidate genes of QTL traits.

Genes with abundance of SNPs: Chromosome (reported candidate genes)	Total SNPs (within respective genes)	Tissue/breed of extracted SNPs	QTL trait	Reference
SPP1: chr7, SCD: chr23, SREBF1: chr3, STAT1: chr2, TG: chr15, LALBA: chr4, INSIG2: chr2, GHRL: chr21, DGAT1: chr15, CSN1S1: chr7, BTN1A1: chr2, ADRA1A: chr3, COL1A2: chr8, APOB: chr12	15, 22, 17, 53, 122, 04, 18, 03, 19, 13, 05, 01	Milk tissue	Milk production	Du et al. (2019)
GDF7: chr12	112, 193	Murrah, Bangladesh, Egyptian, Mediterranean	Milk yield	Iamartino et al. (2017)
KLHL29: chr12	1598	Murrah, Bangladesh, Mediterranean	Milk yield	Iamartino et al. (2017)
RGS22: chr15, VPS13B: chr15	1458	Murrah, Bangladesh, Egyptian, Jaffrabadi, Mediterranean	Milk yield	Iamartino et al. (2017)
MFSD14A: chr6, SLC35A3: chr6, PALMD: chr6	3249	Murrah, Bangladesh, Egyptian, Jaffrabadi, Mediterranean	Milk yield, fat yield, protein yield	Liu et al. (2017)
	344			
	60, 41, 215	Murrah, Bangladesh, Egyptian, Mediterranean	Fat %, protein %	Liu et al. (2017)

Identification of SSR Markers

Maximum number of SSRs were observed in Jaffrabadi (1028180), followed by Bangladesh (9463410) and Mediterranean (908402), while the least was found in Egyptian (726405) (**Figure 3A**). The number of SSRs based on repeat types (mono, di, tri, tetra, penta, hexa-nucleotide repeats) along with their proportions, frequency of SSRs per Mb, and distance between the two SSRs are listed in **Table 5**. In all the breeds, abundance of mononucleotides was observed which might be because of the inherent limitation of the chemistry employed in next-generation sequencing for data generation (Haseneyer et al., 2011) (**Figure 3B**). A similar higher proportion of mono repeats has been found in other animals like cattle, horse, and camel (Ma, 2016; Khalkhali-Evrigh et al., 2019). The relative distributions of various SSR motif lengths in genomes differ from species to species (Sharma et al., 2007). A total of 4329, 4284, 1435, 29822, and 4326 putative genes with an abundance of SSRs in Bangladesh, Egyptian, Jaffrabadi, Mediterranean, and Murrah

breeds, respectively, were annotated. **Figure 3C** shows the common genes with abundance of SSRs in different breeds. The reported putative molecular markers can be used in marker trait association studies for buffalo genetic improvement programs (Sikka and Sethi, 2008; Bhuyan et al., 2010; Kannur et al., 2017).

Identification of microRNAs

We identified a total of 938 miRNAs from the genome assembly of the Mediterranean breed. The pre-miRNA sequences, secondary structure, target information, and location of origin were extracted for each miRNA along with mature miRNA sequence and anti-miRNA star sequence. It was observed that chromosome 11 had the maximum frequency of miRNAs (132 miRNAs) followed by chromosomes 23 (81 miRNAs) and 13 (80 miRNAs) (**Figure 4B**). A target search for 938 miRNAs was performed, out of which 88 miRNAs were found to have 3451 mRNA targets (predicted mode of action of miRNAs was

TABLE 5 | Breed-wise frequencies of SSRs, their proportions, SSR density, and distance between two SSRs in different repeat motifs.

Breeds	Repeats	Number	Proportion %	Frequency of SSRs per Mb	Distance between two SSRs in Kb
Mediterranean	Mono	515343	57.01	191.64	5.22
	Di	176276	19.24	65.55	15.25
	Tri	113425	12.40	42.18	23.71
	Tetra	10120	1.10	3.76	265.72
	Penta	13514	1.48	5.03	198.98
	Hexa	289	0.03	0.11	9304.68
	Compound	79435	8.75	29.54	33.85
Egyptian	Mono	436413	60.08	145.18	6.89
	Di	152723	21.02	50.81	19.68
	Tri	71979	9.91	23.95	41.76
	Tetra	6521	0.90	2.17	460.96
	Penta	5122	0.71	1.70	586.87
	Hexa	107	0.01	0.04	28092.99
	Compound	53541	7.37	17.81	56.14
Jaffrabadi	Mono	580010	56.41	154.26	6.48
	Di	209586	20.38	55.74	17.94
	Tri	127585	12.41	33.93	29.47
	Tetra	11688	1.14	3.11	321.70
	Penta	14154	1.38	3.76	265.65
	Hexa	343	0.03	0.09	10962.04
	Compound	84815	8.25	22.56	44.33
Murrah	Mono	516017	57.63	196.77	5.08
	Di	174481	19.49	66.53	15.03
	Tri	112283	12.54	42.82	23.36
	Tetra	10097	1.13	3.85	259.73
	Penta	13527	1.51	5.16	193.87
	Hexa	300	0.03	0.11	8741.53
	Compound	68658	7.67	26.18	38.20
Bangladesh	Mono	533868	56.41	192.71	5.19
	Di	190507	20.13	68.77	14.54
	Tri	113377	11.98	40.93	24.43
	Tetra	10575	1.12	3.82	261.96
	Penta	12246	1.29	4.42	226.22
	Hexa	268	0.03	0.10	10336.79
	Compound	85501	9.03	30.86	32.40

cleavage of mRNA targets to destroy them or binding with mRNA targets to sequester them) and included in the web resource. Protein encoded by target mRNA, aligned as paired-unpaired sequences of the binding site between mRNA target and miRNA, were also mentioned in the web resource. The miRNAs have the future prospective to be used as biomarkers and for disease management and treatment. miRNAs can be used as a powerful tool to understand the regulatory mechanisms related to disease pathogenesis (Singh et al., 2020; Do et al., 2021).

Identification of Circular RNAs

Out of the total 1702 circRNAs extracted from the 31 buffalo tissues, 1458 were unique circRNAs. **Figure 4A** shows that the maximum number of circRNAs was found in milk (833) tissues followed by embryo pool (153), testis (88), and tongue (52) tissues. Information of genomic localization into intron, exon, and intergenic regions of circRNAs along with genes of origin and strand of origin was extracted by functional annotation of

circRNAs from different tissues which were catalogued in the web resource. The chromosome-wise distribution of circRNAs showed that most of the circRNAs originated from chromosome 2 (227), followed by chromosomes 3 (160) and 4 (155) (**Figure 4B**). circRNAs have multiple regulatory roles which can enrich breeding and improve economic traits related to buffalo (Fu et al., 2018; He et al., 2021; Yang et al., 2021).

Identification of Long Non-Coding RNAs

A total of 44221 lncRNAs were identified in the 31 buffalo tissues. Abundance of lncRNAs was observed in milk tissue (17387) followed by testis (5048) and pooled embryo (4419) (**Figure 4A**). Genomic annotation based on the site of origin of lncRNAs found distribution of 37712 unique lncRNAs into five classes such as intron (14252), isoform/pseudogene (1308), exon (1358), intergenic (17134), and antisense exon (3659) regions. Protein and transcript information was also included for genic origin of lncRNAs. Genomic annotation of unique lncRNAs from

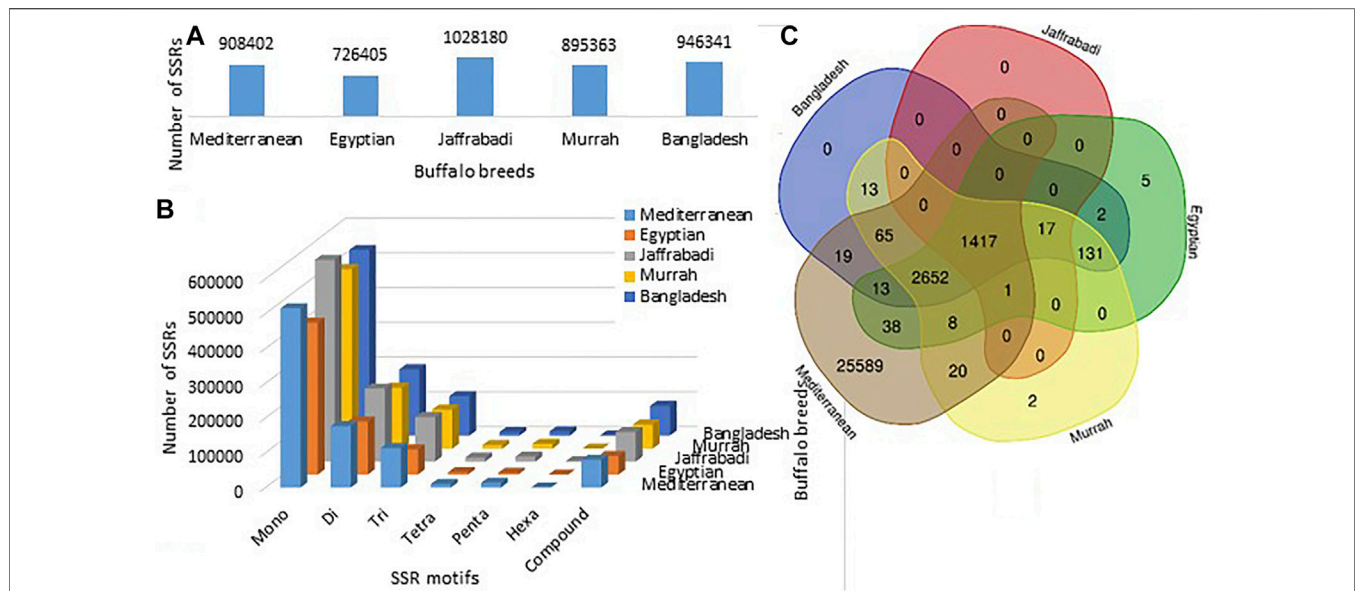


FIGURE 3 | (A) Breed-wise frequencies of SSRs. **(B)** Breed-wise representation of different repeat motifs. **(C)** Common and unique genes with abundance of SSRs in the five breeds of buffalo.

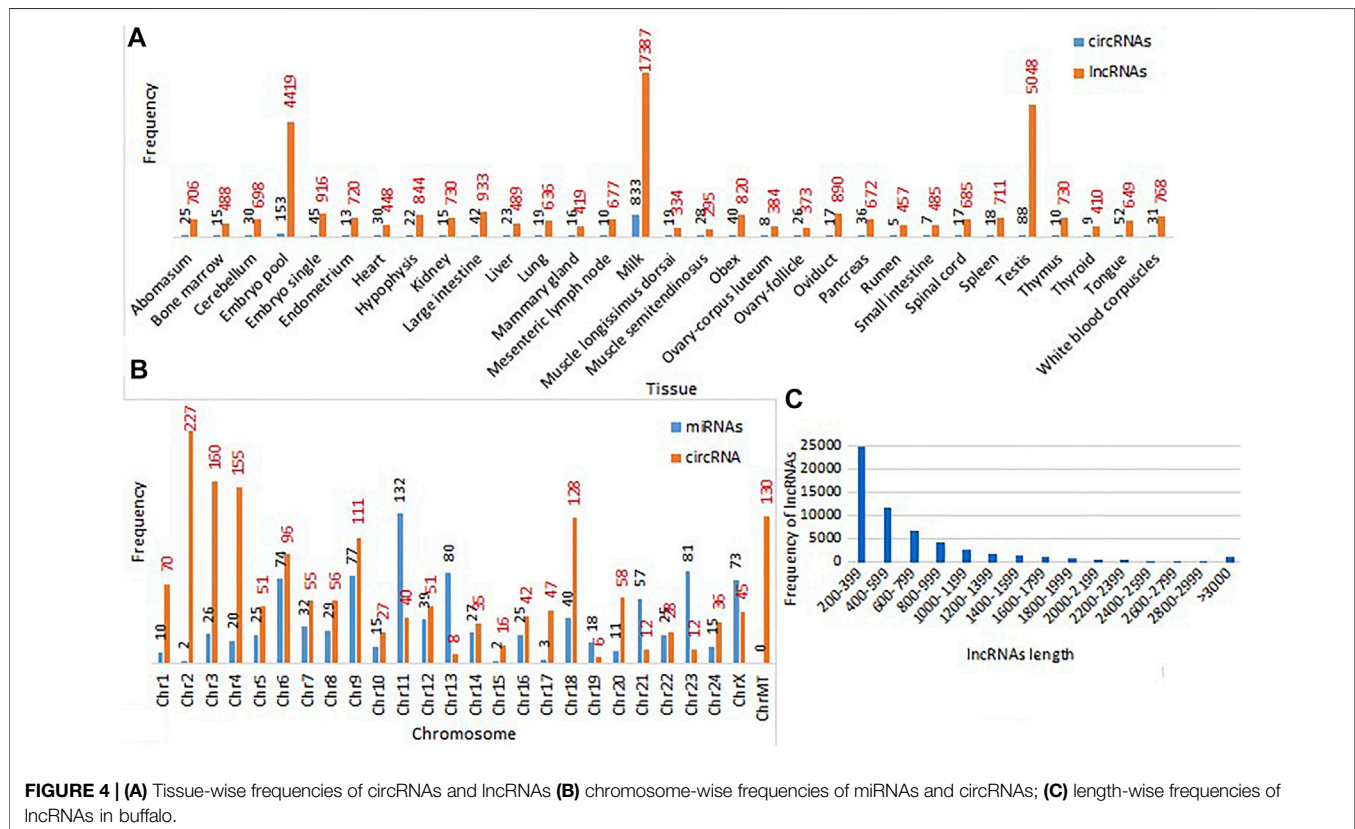
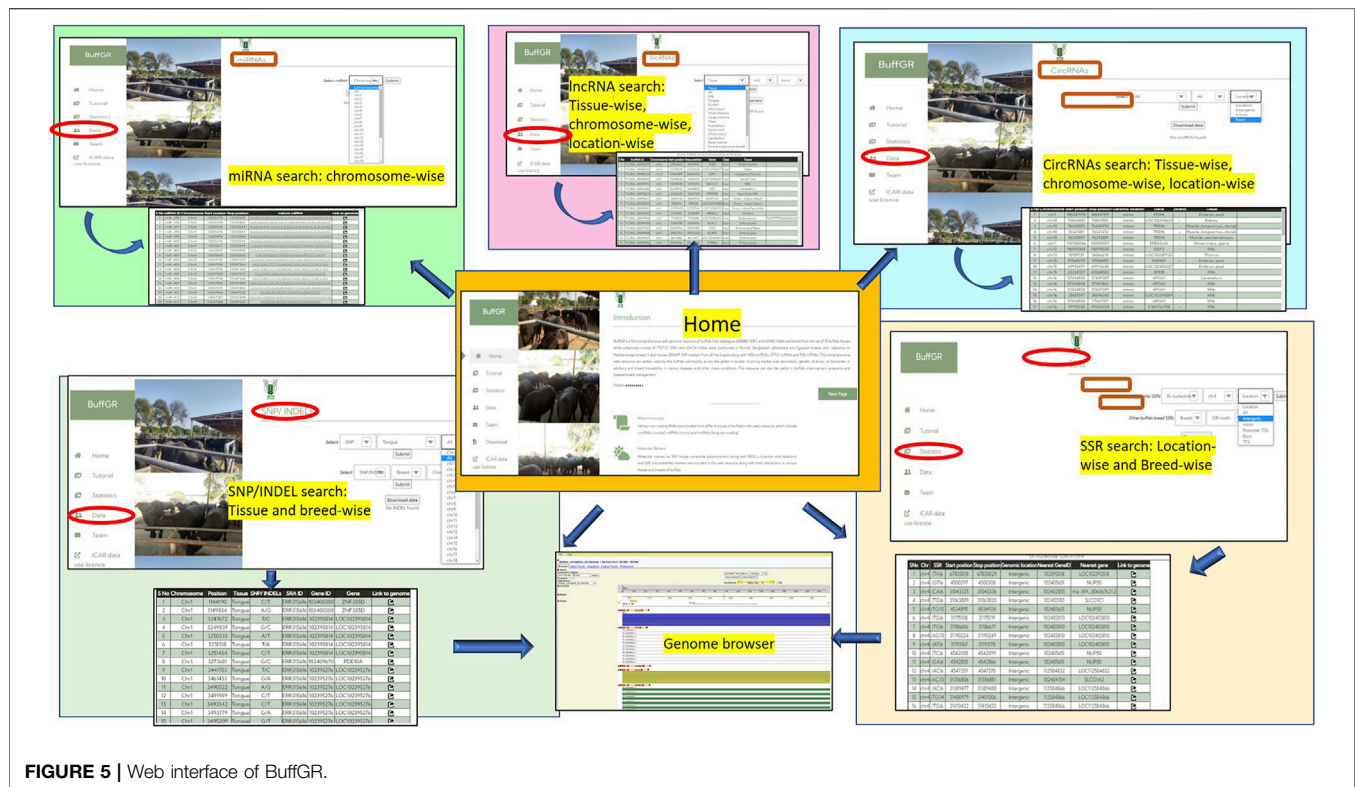


FIGURE 4 | (A) Tissue-wise frequencies of circRNAs and lncRNAs **(B)** chromosome-wise frequencies of miRNAs and circRNAs; **(C)** length-wise frequencies of lncRNAs in buffalo.

all tissues depicted abundance in the intergenic (17134), followed by intron 14252) regions in our study. The graphical representation of lncRNA frequencies based on their length

showed that most lncRNAs had a length of 200–399 bps and had a decreasing trend in frequency with increase in lncRNA length (**Figure 4C**). The role of lncRNAs in genomic studies has



been found to be critical in linking the gap between livestock genotype and phenotype (Kosinska-Selbi et al., 2020).

Development of Buffalo Web Genomic Resource

BuffGR is a comprehensive, first-of-its-kind web resource, with a holistic collection of buffalo molecular markers and variants of five buffalo breeds (Murrah, Mediterranean, Jaffarabadi, Bangladesh, and Egyptian). It is a user-friendly web resource, which catalogues SNPs, InDels, and SSRs along with ncRNAs such as microRNAs, lncRNAs, and circRNAs from the five buffalo breeds and 31 tissues. It has a left vertical section which provides access to varying sections of the web page including Home, Statistics, Data, and Team. The *Home* page includes the brief introduction of the buffalo web genomic resource along with a description about RNA/transcripts and molecular markers of buffalo. The *Statistics* section provides the statistics of extracted buffalo genomic data represented in the form of various graphs and pie charts.

The *Data* section includes hyperlinked images of each data point included in the web resource, and by clicking on the image, the user navigates to the next page of the respective data which provides the user varying options including type of tissue or chromosome number or breed, etc. (as shown in detail in **Figure 1B**). After selecting the combination of options, the user gets a complete table of the related data. The last column of each table provides a hyperlink to the genome

browser, which navigates to the genomic location of the respective marker or ncRNA. In the case of miRNAs, each miRNA sequence is hyperlinked, which navigates to its mRNA target/s wherever available; the *Team* page includes the name of the team members with their profile. The *Tutorial* page guides users regarding the use of this web genomic resource (**Figure 5**).

Utility of Buffalo Web Genomic Resource

The computational approach of discovery of SSR markers, SNPs, and InDels along with miRNAs, lncRNAs, and circRNAs utilizing the available genomic data of different breeds resulted in a ready-to-use, user-friendly, rapid, and economical approach for genomic resource development. The developed web resource, *BuffGR* can be of immense use to the international buffalo research community, which can utilize the information of genomic attributes from five breeds from India (Murrah and Jaffarabadi), Italy (Mediterranean), Bangladesh (Bangladesh), and Egypt (Egyptian). The catalogued SNP/InDel markers from different breeds could be used to study genetic diversity among different breeds of buffalo (Camargo et al., 2015; Deng et al., 2016; El-Halawany et al., 2017; Iamartino et al., 2017; Liu et al., 2017; Dutta et al., 2020). Highly variable SSR markers extracted in the present study could be utilized to find genetic diversity (Barker et al., 1997; Zhang et al., 2020). The SSR markers from different breeds could be used to find polymorphic SSRs (Moore et al., 1995) and their utilization in the study of genetic diversity of respective breeds (Moioli et al., 2001; Merdan et al., 2019; Vohra et al.,

2021; Ünal et al., 2021). We also extracted ~270000 polymorphic SSRs in the Mediterranean buffalo breed with respect to Murrah, Bangladesh, Jaffrabadi, and Egyptian breeds. The species-specific genetic markers (SNP/InDels and SSRs) can also be used as biomarkers of species to be used in the meat industry to trace adulteration or trafficking/traceability (Kannur et al., 2017).

Two coding variants were detected in the ASIP gene by Dutta et al. (2020), one synonymous variant at chr14:19947421 and another non-synonymous variant at chr14:19947429. Dutta et al. (2020) also reported that the alternative allele at the synonymous variant was not observed in Murrah, Surti, or Mediterranean breeds. The potential of extracted SNPs from this study as biomarkers can be seen from the example that the Murrah and Mediterranean breeds in the present study only had one non-synonymous SNP at 19947429 in the ASIP gene on chr14 as reported by Dutta et al. (2020). Significant SNPs could be utilized to find candidate genes specific to a certain function. The variants and SSRs can also be utilized in GWAS (El-Halawany et al., 2017) and later in MTA (marker trait association) analysis and QTL analysis by interval mapping (Deng et al., 2016; Mishra et al., 2020). The present study also shows the potential utilization of extracted markers in marker trait association as few of the genes with abundance of extracted tissue/breed SNPs were found in common with the candidate genes of the few reported QTL traits determined from GWAS studies. We found 12 genes with abundance of milk tissue SNPs to be in common with candidate genes of milk trait, and 10 genes with abundance of SNPs from different breeds to be in common with candidate genes of QTL traits such as milk yield, fat yield, protein yield, fat %, and protein % from other GWAS analyses (Iamartino et al., 2017; Liu et al., 2017; Du et al., 2020).

Tissue-specific lncRNAs could be helpful in studying post-transcriptional regulation by targeting certain mRNAs by cleaving or binding (Zhang et al., 2021) with target mRNAs. lncRNAs could be competitors of miRNAs, which targeted certain mRNAs, where lncRNAs sequestered miRNAs by binding to them and preventing miRNA from cleaving the respective mRNA (Li et al., 2020). Also, tissue-specific lncRNAs could be helpful in utilization in transcriptional regulation by targeting or modulating transcription regulatory proteins by facilitating their binding to a certain site or blocking binding at their target site (Cai et al., 2019; Pan et al., 2021). Another important fact is that the provided tissue-wise lncRNAs are the largest reported group of annotated lncRNAs of buffalo in a single study while several studies report tissue-specific lncRNAs in various species of livestock such as *Bos taurus*, *Gallus gallus*, *Sus scrofa* (Kosinska-Selbi et al., 2020), and *Bos indicus* (Alexandre et al., 2020). The TCONS_00011978 lncRNA, identified from muscle tissue in the present study, was reported to have regulatory potential in muscle with the highest degree of connectivity within the muscle network by Alexandre et al. (2020), reaffirming the potential of our extracted lncRNAs to be utilized in various future studies of buffalo. The buffalo miRNAs and their target mRNAs

extracted in the present study can be utilized in post-transcriptional regulation of certain mRNAs and their encoding proteins by cleaving or binding with their target mRNAs (MacFarlane and Murphy, 2010; Hammond, 2015; Chen et al., 2020; Singh et al., 2020) along with recognition, decapping, and degradation of 3' UTR, and de-adenylation and adenylation of 3' UTR of mRNAs (Shukla et al., 2011). The miRNAs could be used to find their lncRNAs target; action of miRNAs on lncRNAs could be sequestering them by binding or destroying them by cleaving (Assmann et al., 2019; Xie et al., 2020). The tissue-wise extracted circRNAs in the present study could be utilized in the studies of tissue-specific post-transcriptional regulation involving circRNAs and their role in various buffalo diseases (Gao et al., 2018; He et al., 2021; Lei et al., 2021; Yang et al., 2021).

SNP and SSR markers can also be used in parentage and relatedness testing required in breeding and conservation programs (Labuschagne et al., 2015). SNP markers can also be used in estimating inbreeding and effective population sizes required in conservation management monitoring genetic diversity (Panetta et al., 2017). They can be used to compute global co-ancestries of un-pedigreed populations. Such an approach can be of immense use in formulation of selective mating plans based on minimum co-ancestry mating and minimizing inbreeding (Fernández et al., 2005). Both SSR and SNP markers can be used in individual animal identification and breed traceability (Zhao et al., 2020). Water buffalo miRNAs and SNPs can be further used as genomic resources. Such use has been reported in cattle where SNPs and miRNAs have been found associated with bovine phenotypes to be used in breed improvement (Sousa et al., 2021).

CONCLUSION

Through this study, we report the first comprehensive and user-friendly web genomic resource for buffalo (*BuffGR*) including genomic findings of five commercially important buffalo breeds, namely Mediterranean, Egyptian, Bangladesh, Jaffrabadi, and Murrah. *BuffGR* catalogues a total of 6028881 SNPs and 613403 InDels extracted from the set of 31 buffalo tissues. Collectively, a total of 7727122 SNPs and 634124 InDels were distributed in the four breeds of buffalo (Murrah, Bangladesh, Jaffrabadi, and Egyptian) with reference to the Mediterranean breed. The web resource has 4504691 SSR markers from all the breeds, 1458 unique circRNAs and 37712 lncRNAs from 31 buffalo tissues, and 938 miRNAs from the genome assembly of the Mediterranean breed. This information can be widely used by the buffalo researchers across the globe for studying the genetic diversity among the different breeds of buffalo, studies involving post-transcriptional regulation, and their role in various buffalo diseases. The provided markers can be used as biomarkers in the meat industry to trace adulteration, trafficking, and breed traceability. These can be used not only for knowledge discovery research but also for marker trait association, which will be helpful in the improvement and management of buffalo breeds.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

MI, DK, and SJ conceived the theme of the study. AaK, KS, SJ, MR, RJ, AnK, AG, JK, MI, and UA performed the computational analysis and developed genomic resources. KS, AaK, MI, SJ, and DK drafted the manuscript. VN, AR, TD, and DK edited the manuscript. All co-authors read and approved the final manuscript.

REFERENCES

- Alexandre, P. A., Reverter, A., Berezin, R. B., Porto-Neto, L. R., Ribeiro, G., Santana, M. H. A., et al. (2020). Exploring the Regulatory Potential of Long Non-coding RNA in Feed Efficiency of Indicine Cattle. *Genes* 11 (9), 997. doi:10.3390/genes11090997
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/s0022-2836(05)80360-2
- Assmann, T. S., Milagro, F. N., and Martínez, J. (2019). Crosstalk between microRNAs, the Putative Target Genes and the lncRNA Network in Metabolic Diseases. *Mol. Med. Rep.* 20 (4), 3543–3554. doi:10.3892/mmr.2019.10595
- Barker, J. S. F., Moore, S. S., Hetzel, D. J. S., Evans, D., Byrne, K., and Tan, S. G. (1997). Genetic Diversity of Asian Water buffalo (*Bubalus Bubalis*): Microsatellite Variation and a Comparison with Protein-Coding Loci. *Anim. Genet.* 28 (2), 103–115. doi:10.1111/j.1365-2052.1997.00085.x
- Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a Web Server for Microsatellite Prediction. *Bioinformatics* 33 (16), 2583–2585. doi:10.1093/bioinformatics/btx198
- Bhat, S., and Jones, W. D. (2016). An Accelerated miRNA-Based Screen Implicates Atf-3 in *Drosophila* Odorant Receptor Expression. *Sci. Rep.* 6 (1), 1–8. doi:10.1038/srep20109
- Bhuyan, D. K., Sangwan, M. L., Gole, V. C., and Sethi, R. K. (2010). Studies on DNA Fingerprinting in Murrah Buffaloes Using Microsatellite Markers. *Indian J. Biotechnol.* 9 (4), 367–370.
- Burset, M., and Guigó, R. (1996). Evaluation of Gene Structure Prediction Programs. *Genomics* 34 (3), 353–367. doi:10.1006/geno.1996.0298
- Cai, R., Tang, G., Zhang, Q., Yong, W., Zhang, W., Xiao, J., et al. (2019). A Novel Lnc-RNA, Named Lnc-ORA, Is Identified by RNA-Seq Analysis, and its Knockdown Inhibits Adipogenesis by Regulating the PI3K/AKT/mTOR Signaling Pathway. *Cells* 8 (5), 477. doi:10.3390/cells8050477
- Chen, L., Zhang, S., Wu, J., Cui, J., Zhong, L., Zeng, L., et al. (2017). circRNA_100290 Plays a Role in Oral Cancer by Functioning as a Sponge of the miR-29 Family. *Oncogene* 36 (32), 4551–4561. doi:10.1038/onc.2017.89
- Chen, Z., Xie, Y., Luo, J., Chen, T., Xi, Q., Zhang, Y., et al. (2020). Milk Exosome-Derived miRNAs from Water buffalo Are Implicated in Immune Response and Metabolism Process. *BMC Vet. Res.* 16 (1), 123–125. doi:10.1186/s12917-020-02339-x
- Dai, X., and Zhao, P. X. (2011). psRNA Target: a Plant Small RNA Target Analysis Server. *Nucleic Acids Res.* 39 (Suppl. 1_2), W155–W159. doi:10.1093/nar/gkr319
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *GigaScience* 10 (2), giab008. doi:10.1093/gigascience/giab008

ACKNOWLEDGMENTS

The authors are thankful to the Indian Council of Agricultural Research (ICAR), Ministry of Agriculture and Farmers' Welfare, Government of India for providing financial assistance in the form of a CABIN grant as well as the use of the Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility at ICAR-IASRI, New Delhi, India created under the National Agricultural Innovation Project, and funded by the World Bank. The authors further acknowledge the supportive role of the Director of ICAR-IASRI, New Delhi and Director, ICAR-CIRB, Hisar, India. The grant of Junior Research Fellowship to AaK by the Indian Council of Agricultural Research is duly acknowledged. Authors are also thankful to Lal Bahadur Shastri Outstanding Young Scientist Scheme, ICAR for necessary support.

- de Camargo, G., Aspilcueta-Borquis, R., Fortes, M., Porto-Neto, R., Cardoso, D., Santos, D., et al. (2015). Prospecting Major Genes in Dairy Buffaloes. *BMC Genomics* 16 (1), 1–14. doi:10.1186/s12864-015-1986-2
- Deng, T. X., Pang, C. Y., Liu, M. Q., Zhang, C., and Liang, X. W. (2016). Synonymous Single Nucleotide Polymorphisms in the MC4R Gene that Are Significantly Associated with Milk Production Traits in Water Buffaloes. *Genet. Mol. Res.* 15, 1–8. doi:10.4238/gmr.15028153
- Dhanoa, J. K., Singh, J., Singh, A., Arora, J. S., Sethi, R. S., and Mukhopadhyay, C. S. (2019). Discovery of isomiRs in PBMCs of Diseased Vis-À-Vis Healthy Indian Water Buffaloes. *ExRNA* 1 (1), 1–12. doi:10.1186/s41544-019-0013-1
- Do, D. N., Dudemaine, P.-L., Mathur, M., Suravajhala, P., Zhao, X., and Ibeagha-Awemu, E. M. (2021). miRNA Regulatory Functions in Farm Animal Diseases, and Biomarker Potentials for Effective Therapies. *Ijms* 22 (6), 3080. doi:10.3390/ijms22063080
- Du, C., Deng, T., Zhou, Y., Ye, T., Zhou, Z., Zhang, S., et al. (2019). Systematic Analyses for Candidate Genes of Milk Production Traits in Water buffalo (*Bubalus Bubalis*). *Anim. Genet.* 50 (3), 207–216. doi:10.1111/age.12739
- Dutta, P., Talenti, A., Young, R., Jayaraman, S., Callaby, R., Jadhav, S. K., et al. (2020). Whole Genome Analysis of Water buffalo and Global Cattle Breeds Highlights Convergent Signatures of Domestication. *Nat. Commun.* 11 (1), 1–13. doi:10.1038/s41467-020-18550-1
- El-Halawany, N., Abdel-Shafy, H., Shawky, A.-E. -M. A., Abdel-Latif, M. A., Al-Tohamy, A. F. M., and Abd El-Moneim, O. M. (2017). Genome-wide Association Study for Milk Production in Egyptian buffalo. *Livestock Sci.* 198, 10–16. doi:10.1016/j.livsci.2017.01.019
- FAOSTAT (2020). The Food and Agriculture Organization (FAO) of the United Nations Statistics Division. Available at: <https://www.fao.org/faostat/en/#home> (Accessed August 2021).
- Fernandes, L. G. V., Guaman, L. P., Vasconcellos, S. A., Heinemann, M. B., Picardeau, M., and Nascimento, A. L. T. O. (2019). Gene Silencing Based on RNA-Guided Catalytically Inactive Cas9 (dCas9): a New Tool for Genetic Engineering in *Leptospira*. *Sci. Rep.* 9 (1), 1–14. doi:10.1038/s41598-018-37949-x
- Fernández, J., Villanueva, B., Pong-Wong, R., and Toro, M. A. (2005). Efficiency of the Use of Pedigree and Molecular Marker Information in Conservation Programs. *Genetics* 170 (3), 1313–1321. doi:10.1534/genetics.104.037325
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* 39 (Suppl. 1_2), W29–W37. doi:10.1093/nar/gkr367
- Fu, Y., Jiang, H., Liu, J.-B., Sun, X.-L., Zhang, Z., Li, S., et al. (2018). Genome-wide Analysis of Circular RNAs in Bovine Cumulus Cells Treated with BMP15 and GDF9. *Sci. Rep.* 8 (1), 1–10. doi:10.1038/s41598-018-26157-2
- Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an Efficient and Unbiased Algorithm for De Novo Circular RNA Identification. *Genome Biol.* 16 (1), 1–16. doi:10.1186/s13059-014-0571-3

- Gao, Y., Wu, M., Fan, Y., Li, S., Lai, Z., Huang, Y., et al. (2018). Identification and Characterization of Circular RNAs in Qinchuan Cattle Testis. *R. Soc. Open Sci.* 5 (7), 180413. doi:10.1098/rsos.180413
- Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., and Gelbart, W. M. (2000). *Transcription: An Overview of Gene Regulation in Eukaryotes*. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK21766/>.
- Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA Sequences, Targets and Gene Nomenclature. *Nucleic Acids Res.* 34 (Suppl. 1), D140–D144. doi:10.1093/nar/gkj112
- Hammond, S. M. (2015). An Overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi:10.1016/j.addr.2015.05.001
- Haseneyer, G., Schmutzer, T., Seidel, M., Zhou, R., Mascher, M., Schön, C.-C., et al. (2011). From RNA-Seq to Large-Scale Genotyping - Genomics Resources for rye (*Secale Cereale* L.). *BMC Plant Biol.* 11 (1), 1–13. doi:10.1186/1471-2229-11-131
- He, T., Chen, Q., Tian, K., Xia, Y., Dong, G., and Yang, Z. (2021). Functional Role of circRNAs in the Regulation of Fetal Development, Muscle Development, and Lactation in Livestock. *Biomed. Res. Int.* 2021, 5383210. doi:10.1155/2021/5383210
- Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics* 26 (5), 680–682. doi:10.1093/bioinformatics/btq003
- Iamartino, D., Nicolazzi, E. L., Van Tassel, C. P., Reecy, J. M., Fritz-Waters, E. R., Koltes, J. E., et al. (2017). Design and Validation of a 90K SNP Genotyping Assay for the Water buffalo (*Bubalus Bubalis*). *PLoS One* 12 (10), e0185220. doi:10.1371/journal.pone.0185220
- Iannuzzi, L. (1994). Standard Karyotype of the River buffalo (*Bubalus Bubalis* L., 2n = 50). Report of the Committee for the Standardization of Banded Karyotypes of the River buffalo. *Cytogenet. Cell. Genet.* 67 (2), 102–113. doi:10.1159/000133808
- Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. *Nucleic Acids Res.* 45 (W1), W12–W16. doi:10.1093/nar/gkx428 https://www.ncbi.nlm.nih.gov/assembly/GCF_003121395.1/ https://www.ncbi.nlm.nih.gov/assembly/GCA_000180995.3/ https://www.ncbi.nlm.nih.gov/assembly/GCA_004794615.1/ https://www.ncbi.nlm.nih.gov/assembly/GCA_002993835.1/ https://www.ncbi.nlm.nih.gov/assembly/GCF_019923935.1/
- Kannur, B. H., Fairoze, M. N., Girish, P. S., Karabasanavar, N., and Rudresh, B. H. (2017). Breed Traceability of buffalo Meat Using Microsatellite Genotyping Technique. *J. Food Sci. Technol.* 54 (2), 558–563. doi:10.1007/s13197-017-2500-4
- Kawamata, T., and Tomari, Y. (2010). Making Risc. *Trends Biochemical Sciences* 35 (7), 368–376. doi:10.1016/j.tibs.2010.03.009
- Khalkhali-Evrigh, R., Hedayat-Evrigh, N., Hafezian, S. H., Farhadi, A., and Bakhtiarzadeh, M. R. (2019). Genome-wide Identification of Microsatellites and Transposable Elements in the Dromedary Camel Genome Using Whole-Genome Sequencing Data. *Front. Genet.* 10, 692. doi:10.3389/fgene.2019.00692
- Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype. *Nat. Biotechnol.* 37, 907–915. doi:10.1038/s41587-019-0201-4
- Kosinska-Selbi, B., Mielczarek, M., and Szyda, J. (2020). Review: Long Non-coding RNA in Livestock. *Animal* 14 (10), 2003–2013. doi:10.1017/s1751731120000841
- Labuschagne, C., Nupen, L., Kotzé, A., Grobler, P. J., and Dalton, D. L. (2015). Assessment of Microsatellite and SNP Markers for Parentage Assignment in Ex Situ African Penguin (*Spheniscus demersus*) Populations. *Ecol. Evol.* 5 (19), 4389–4399. doi:10.1002/ece3.1600
- Langmead, B., and Salzberg, S. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923
- Lecchi, C., Catozzi, C., Zamarian, V., Poggi, G., Borriello, G., Martucciello, A., et al. (2019). Characterization of Circulating miRNA Signature in Water Buffaloes (*Bubalus Bubalis*) during Brucella Abortus Infection and Evaluation as Potential Biomarkers for Non-invasive Diagnosis in Vaginal Fluid. *Sci. Rep.* 9 (1), 1945. doi:10.1038/s41598-018-38365-x
- Leclercq, M., Diallo, A. B., and Blanchette, M. (2013). Computational Prediction of the Localization of microRNAs within Their Pre-miRNA. *Nucleic Acids Res.* 41 (15), 7200–7211. doi:10.1093/nar/gkt466
- Lei, Z., Wu, H., Xiong, Y., Wei, D., Wang, X., Luoreng, Z., et al. (2021). ncRNAs Regulate Bovine Adipose Tissue Deposition. *Mol. Cell Biochem* 476 (7), 2837–2845. doi:10.1007/s11010-021-04132-2
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352
- Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* 27 (21), 2987–2993. doi:10.1093/bioinformatics/btr509
- Li, H., Huang, K., Wang, P., Feng, T., Shi, D., Cui, K., et al. (2020). Comparison of Long Non-coding RNA Expression Profiles of Cattle and buffalo Differing in Muscle Characteristics. *Front. Genet.* 11, 98. doi:10.3389/fgene.2020.00098
- Liu, J. J., Liang, A. X., Campanile, G., Plastow, G., Zhang, C., Wang, Z., et al. (2017). Genome-wide Association Studies to Identify Quantitative Trait Loci Affecting Milk Production Traits in Water buffalo. *J. Dairy Sci.* 101, 433–444. doi:10.3168/jds.2017-13246
- Liu, S., Ye, T., Li, Z., Li, J., Jamil, A. M., Zhou, Y., et al. (2019). Identifying Hub Genes for Heat Tolerance in Water buffalo (*Bubalus Bubalis*) Using Transcriptome Data. *Front. Genet.* 10, 209. doi:10.3389/fgene.2019.00209
- Lorenz, R., Bernhart, S. H., Höner Zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6 (1), 26–14. doi:10.1186/1748-7188-6-26
- Lu, M. (2020). Circular RNA: Functions, Applications and Prospects. *ExRNA* 2 (1), 1–7. doi:10.1186/s41544-019-0046-5
- Lukiw, W. J. (2013). Circular RNA (circRNA) in Alzheimer's Disease (AD). *Front. Genet.* 4, 307. doi:10.3389/fgene.2013.00307
- Ma, Z.-J. (2016). Abundance and Characterization of Perfect Microsatellites on the Cattle Y Chromosome. *Anim. Biotechnol.* 28 (3), 157–162. doi:10.1080/10495398.2016.1243551
- MacFarlane, L.-A., and Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. *Cg* 11 (7), 537–561. doi:10.2174/138920210793175895
- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency. *Nature* 495 (7441), 333–338. doi:10.1038/nature11928
- Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., et al. (2017). CircRNA: Functions and Properties of a Novel Potential Biomarker for Cancer. *Mol. Cancer* 16 (1), 1–8. doi:10.1186/s12943-017-0663-2
- Merdan, S. M., El-Zarei, M. F., Ghazy, A., Ayoub, M. A., Al-Shawa, Z. M., and Mokhtar, S. A. (2019). Genetic Differentiation between Egyptian Buffalo Populations Using Microsatellite Markers. *J. Anim. Poult. Fish Prod.* 8 (1), 21–28.
- Min, X. J., Butler, G., Storms, R., and Tsang, A. (2005). OrfPredictor: Predicting Protein-Coding Regions in EST-Derived Sequences. *Nucleic Acids Res.* 33 (Suppl. 1), W677–W680. doi:10.1093/nar/gki394
- Mishra, D. C., Sikka, P., Yadav, S., Bhati, J., Paul, S. S., Jerome, A., et al. (2020). Identification and Characterization of Trait-specific SNPs Using ddRAD Sequencing in Water buffalo. *Genomics* 112 (5), 3571–3578. doi:10.1016/j.ygeno.2020.04.012
- Moioli, B., Georgoudis, A., Napolitano, F., Catillo, G., Giubilei, E., Ligda, C., et al. (2001). Genetic Diversity between Italian, Greek and Egyptian buffalo Populations. *Livestock Prod. Sci.* 70 (3), 203–211. doi:10.1016/S0301-6226(01)00175-0
- Moore, S. S., Evans, D., Byrne, K., Barker, J. S. F., Tan, S. G., Vankan, D., et al. (1995). A Set of Polymorphic DNA Microsatellites Useful in Swamp and River buffalo (*Bubalus Bubalis*). *Anim. Genet.* 26 (5), 355–359. doi:10.1111/j.1365-2052.1995.tb02674.x
- O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of microRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol.* 9, 402. doi:10.3389/fendo.2018.00402
- Pan, Y., Yang, S., Cheng, J., Lv, Q., Xing, Q., Zhang, R., et al. (2021). Whole-Transcriptome Analysis of LncRNAs Mediated ceRNA Regulation in Granulosa Cells Isolated from Healthy and Atresia Follicles of Chinese Buffalo. *Front. Vet. Sci.* 8, 680182. doi:10.3389/fvets.2021.680182
- Panetto, J. C. D. C., Machado, M. A., da Silva, M. V. G. B., Barbosa, R. S., dos Santos, G. G., Leite, R. d. M. H., et al. (2017). Parentage Assignment Using SNP Markers, Inbreeding and Population Size for the Brazilian Red Sindhi Cattle. *Livestock Sci.* 204, 33–38. doi:10.1016/j.livsci.2017.08.008

- Pareek, C. S., Czarnik, U., Pierzchała, M., and Zwierchowski, L. (2008). An Association between the C> T Single Nucleotide Polymorphism within Intron IV of Osteopontin Encoding Gene (SPP1) and Body Weight of Growing Polish Holstein-Friesian Cattle. *Anim. Sci. Pap. Rep.* 26 (4), 251–257.
- Patzak, J., Paprštein, F., Henychová, A., and Sedlák, J. (2012). Comparison of Genetic Diversity Structure Analyses of SSR Molecular Marker Data within Apple (*Malus domestica*) Genetic Resources. *Genome* 55 (9), 647–665. doi:10.1139/G2012-054
- Perte, M., Perte, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122
- Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of Novel Transcripts in Annotated Genomes Using RNA-Seq. *Bioinformatics* 27 (17), 2325–2329. doi:10.1093/bioinformatics/btr355
- Sharma, P. C., Grover, A., and Kahl, G. (2007). Mining Microsatellites in Eukaryotic Genomes. *Trends Biotechnology* 25 (11), 490–498. doi:10.1016/j.tibtech.2007.07.013
- Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* 11 (10), e0163962. doi:10.1371/journal.pone.0163962
- Shukla, G. C., Singh, J., and Barik, S. (2011). MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Mol. Cell. Pharmacol.* 3 (3), 83–92. PMC3315687.
- Sikka, P., and Sethi, R. K. (2008). Genetic Variability in Production Performance of Murrah Buffaloes (*Bubalus Bubalis*) Using Microsatellite Polymorphism. *Indian J. Biotechnol.* 7, 103–107.
- Singh, J., Dhanoa, J. K., Choudhary, R. K., Singh, A., Sethi, R. S., Kaur, S., et al. (2020). MicroRNA Expression Profiling in PBMCs of Indian Water Buffalo (*Bubalus Bubalis*) Infected with Brucella and Johne's Disease. *ExRNA* 2, 1–13. doi:10.1186/s41544-020-00049-y
- Sousa, M. A. P. D., de Athayde, F. R. F., Maldonado, M. B. C., Lima, A. O. D., Fortes, M. R. S., and Lopes, F. L. (2021). Single Nucleotide Polymorphisms Affect miRNA Target Prediction in Bovine. *PLoS One* 16 (4), e0249406. doi:10.1371/journal.pone.0249406
- Surya, T., Vineeth, M. R., Sivalingam, J., Tania, M. S., Dixit, S. P., Niranjana, S. K., et al. (2019). Genomewide Identification and Annotation of SNPs in Bubalus Bubalis. *Genomics* 111 (6), 1695–1698. doi:10.1016/j.ygeno.2018.11.021
- Taşcıoğlu, Y., Akpınar, M. G., Gül, M., Karli, B., and Bozkurt, Y. (2020). Determination of Optimum Agricultural Policy for buffalo Breeding. *Revista Brasileira de Zootecnia* 49, 1–10. doi:10.37496/rbz4920200120
- Ünal, E. Ö., Işık, R., Şen, A., Geyik Kuş, E., and Soysal, M. İ. (2021). Evaluation of Genetic Diversity and Structure of Turkish Water Buffalo Population by Using 20 Microsatellite Markers. *Animals* 11 (4), 1067. doi:10.3390/ani11041067
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new Capabilities and Interfaces. *Nucleic Acids Res.* 40, e115. doi:10.1093/nar/gks596
- Vohra, V., Singh, N. P., Chhotaray, S., Raina, V. S., Chopra, A., and Kataria, R. S. (2021). Morphometric and Microsatellite-Based Comparative Genetic Diversity Analysis in Bubalus Bubalis from North India. *PeerJ* 9, e11846. doi:10.7717/peerj.11846
- Wang, K., Gan, T.-Y., Li, N., Liu, C.-Y., Zhou, L.-Y., Gao, J.-N., et al. (2017). Circular RNA Mediates Cardiomyocyte Death via miRNA-dependent Upregulation of MTP18 Expression. *Cell Death Differ.* 24 (6), 1111–1120. doi:10.1038/cdd.2017.61
- Xie, F., Liu, Y. L., Chen, X. Y., Li, Q., Zhong, J., Dai, B. Y., et al. (2020). Role of MicroRNA, LncRNA, and Exosomes in the Progression of Osteoarthritis: a Review of Recent Literature. *Orthop. Surg.* 12 (3), 708–716. doi:10.1111/os.12690
- Xue, C., Li, F., He, T., Liu, G. P., Li, Y., and Zhang, X. (2005). Classification of Real and Pseudo microRNA Precursors Using Local Structure-Sequence Features and Support Vector Machine. *BMC Bioinformatics* 6 (1), 310–317. doi:10.1186/1471-2105-6-310
- Yang, Z., He, T., and Chen, Q. (2021). The Roles of CircRNAs in Regulating Muscle Development of Livestock Animals. *Front. Cell. Dev. Biol.* 9, 163. doi:10.3389/fcell.2021.619329
- Zhang, R., Wang, J., Xiao, Z., Zou, C., An, Q., Li, H., et al. (2021). The Expression Profiles of mRNAs and lncRNAs in Buffalo Muscle Stem Cells Driving Myogenic Differentiation. *Front. Genet.* 12, 1048. doi:10.3389/fgene.2021.643497
- Zhang, Y., Colli, L., and Barker, J. S. F. (2020). Asian Water buffalo: Domestication, History and Genetics. *Anim. Genet.* 51 (2), 177–191. doi:10.1111/age.12911
- Zhao, C., Qiu, J., Agarwal, G., Wang, J., Ren, X., Xia, H., et al. (2017). Genome-wide Discovery of Microsatellite Markers from Diploid Progenitor Species, *Arachis Duranensis* and *A. Ipaensis*, and Their Application in Cultivated Peanut (*A. hypogaea*). *Front. Plant Sci.* 8, 1209. doi:10.3389/fpls.2017.01209
- Zhao, J., Li, A., Jin, X., and Pan, L. (2020). Technologies in Individual Animal Identification and Meat Products Traceability. *Biotechnol. Biotechnological Equipment* 34 (1), 48–57. doi:10.1080/13102818.2019.1711185

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Khan, Singh, Jaiswal, Raza, Jasrotia, Kumar, Gurjar, Kumari, Nayan, Iqbal, Angadi, Rai, Datta and Kumar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Improvement of Genomic Predictions in Small Breeds by Construction of Genomic Relationship Matrix Through Variable Selection

Enrico Mancin[†], Lucio Flavio Macedo Mota[†], Beniamino Tuliozi^{*}, Rina Verdiglione, Roberto Mantovani[‡] and Cristina Sartori[‡]

Department of Agronomy, Food, Natural Resources, Animals and Environment, University of Padua, Legnaro, Italy

OPEN ACCESS

Edited by:

Mudasir Ahmad Syed,
Sher-e-Kashmir University of
Agricultural Sciences and Technology,
India

Reviewed by:

Yi Liu,
University of Chicago, United States
Breno De Oliveira Fragomeni,
University of Connecticut,
United States

*Correspondence:

Beniamino Tuliozi
beniamino.tuliozi@unipd.it

[†]These authors have contributed
equally to this work and share first
authorship

[‡]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 13 November 2021

Accepted: 22 March 2022

Published: 18 May 2022

Citation:

Mancin E, Mota LFM, Tuliozi B,
Verdiglione R, Mantovani R and
Sartori C (2022) Improvement of
Genomic Predictions in Small Breeds
by Construction of Genomic
Relationship Matrix Through
Variable Selection.
Front. Genet. 13:814264.
doi: 10.3389/fgene.2022.814264

Genomic selection has been increasingly implemented in the animal breeding industry, and it is becoming a routine method in many livestock breeding contexts. However, its use is still limited in several small-population local breeds, which are, nonetheless, an important source of genetic variability of great economic value. A major roadblock for their genomic selection is accuracy when population size is limited: to improve breeding value accuracy, variable selection models that assume heterogeneous variance have been proposed over the last few years. However, while these models might outperform traditional and genomic predictions in terms of accuracy, they also carry a proportional increase of breeding value bias and dispersion. These mutual increases are especially striking when genomic selection is performed with a low number of phenotypes and high shrinkage value—which is precisely the situation that happens with small local breeds. In our study, we tested several alternative methods to improve the accuracy of genomic selection in a small population. First, we investigated the impact of using only a subset of informative markers regarding prediction accuracy, bias, and dispersion. We used different algorithms to select them, such as recursive feature eliminations, penalized regression, and XGBoost. We compared our results with the predictions of pedigree-based BLUP, single-step genomic BLUP, and weighted single-step genomic BLUP in different simulated populations obtained by combining various parameters in terms of number of QTLs and effective population size. We also investigated these approaches on a real data set belonging to the small local Rendena breed. Our results show that the accuracy of GBLUP in small-sized populations increased when performed with SNPs selected via variable selection methods both in simulated and real data sets. In addition, the use of variable selection models—especially those using XGBoost—in our real data set did not impact bias and the dispersion of estimated breeding values. We have discussed possible explanations for our results and how our study can help estimate breeding values for future genomic selection in small breeds.

Keywords: genomic selection accuracy, single-step GBLUP, SNP selection methods, machine learning, local breed cattle, Rendena, genomic selection

INTRODUCTION

Genomic information has been successfully implemented in animal breeding due to its effectiveness in bringing significant improvements in accuracy (Blasco and Toro, 2014). These improvements in accuracy can lead to an increase in the rate of genetic gains and have reduced the cost of progeny testing by allowing to preselect animals with great genetic merit early (Meuwissen et al., 2001). Combining these advancements with the progressively reduced cost of genotyping makes single-nucleotide polymorphism (SNP) panels a promising tool to select small local breeds (Biscarini et al., 2015).

SNP marker information allows for better modeling of Mendelian sampling than the traditional pedigree-based best linear unbiased prediction (PBLUP) (VanRaden, 2008a), which used only pedigree information. The genomic BLUP (GBLUP) method was developed to replace the pedigree-based relationships for genomic relationships estimated from SNP markers, which captured the genomic similarity between animals but are limited to the use of only genotyped animals (Habier et al., 2013). In addition, Legarra et al. (2009) proposed a naive method, single-step GBLUP (ssGBLUP), in which genotyped and non-genotyped animals are jointly combined under the assumption that the genomic and pedigree relationship matrixes are multivariate and normally distributed. Due to its straightforward computational approach (Misztal et al., 2013) and unbiased breeding values predictions, compared to the GBLUP with its multistep approach (Masuda et al., 2018), the ssGBLUP has become a routine method for genomic evaluations in many livestock breeds and species (Aguilar et al., 2010; Christensen and Lund, 2010).

However, one major challenge in using (ss)GBLUP remains the accuracy of estimation when phenotyped animals are limited in number, such as in local breeds (Meuwissen et al., 2001). For example, Karaman et al. (2016) reported that GBLUP showed lower performance than that of models using only SNPs selected through a Bayesian hierarchical model as Bayes B and Bayes C, but only when phenotyped animals were few. Indeed, when presented with a small number of animals and many SNP markers ($n < p$), models that select a number of priority SNPs (variable selection models) and models that assume heterogeneous variance can lead to improvements in EBV accuracy. These models can accomplish this by reducing the number of variables to estimate and by preventing overfitting linked to high-dimensional data (Gianola 2013). Frouin et al. (2020) went as far as deriving the prediction accuracy of GBLUP as a function of the ratio n/p , while Pocrnic et al. (2019) regarded the accuracy of GBLUP as not only strictly dependent on the number of SNPs but also on the number of independent chromosome segments.

Several studies thus focused on relaxing the assumption of ssGBLUP that all SNPs must show a common variance by applying different weights to the SNPs when the G matrix is calculated. Methods such as weighted ssGBLUP (WssGBLUP) (Wang et al., 2014) were widely reported to outperform ssGBLUP's accuracy of prediction (Gualdrón Duarte et al., 2014; Gualdrón Duarte et al., 2020; Mehrban et al., 2021; Ren

et al., 2021), but their use led to a proportional increase of breeding value bias and dispersion (Mancin et al., 2021b; Botelho et al., 2021; Cesarani et al., 2021; Mehrban et al., 2021).

Moreover, it is unclear how models considering heterogeneous variances account for selection since only k -fold cross-validation is usually applied (Zhu et al., 2021). In real-life breeding scenarios, time cross-validation should be considered (Liu, 2010) because this validation method mimics the true accumulation of information across time. The estimated breeding values (EBVs) are in fact used to select young bulls, and after 3–5 years, the bulls will receive daughter information; it is thus desirable that EBVs would highly correlate to the final EBVs. However, the few studies that evaluated the impact of WssGBLUP using time cross-validation with small samples of individuals (e.g., Cesarani et al., 2021) found higher bias and overdispersion. These mutual increases are relevant when a low number of phenotypes and high shrinkage values are present, and the reasons behind the loss of these unbiased properties in heterogeneous SNP regression or GBLUP are still not entirely clear.

This issue is not trivial as the bias and the slope of the regression (dispersion) need to be considered, especially when proven, and young animals are mixed in the population as young candidates will have unfair EBVs (Legarra and Reverter, 2017).

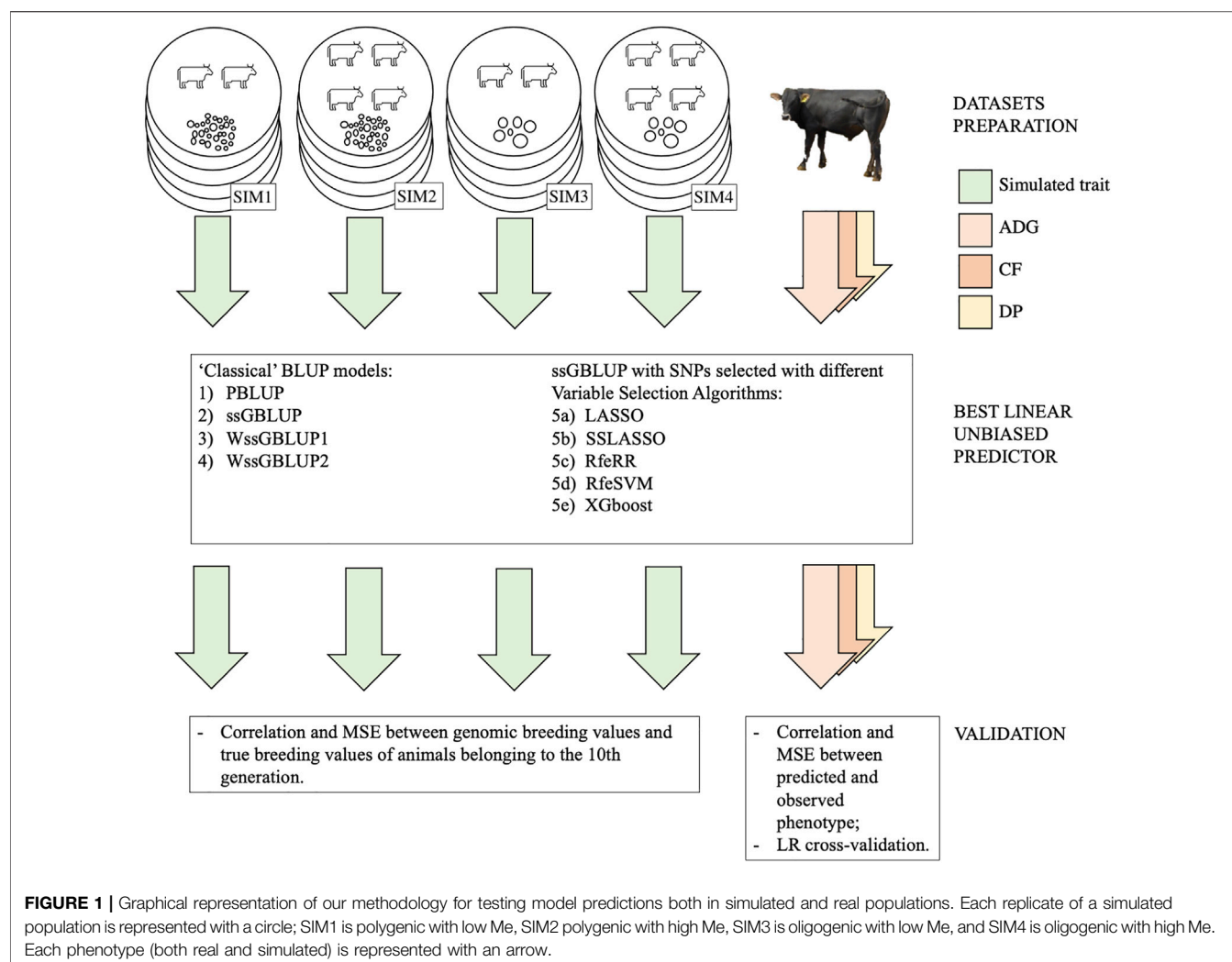
Thus, the abovementioned issues of lack of accuracy of ssGBLUP when used in contexts with a few animals have not been conclusively resolved. For this reason, in the present study, we intend to explore alternative methods to improve accuracy in small populations within a single-step framework. A possible solution could come from implementing a naive approach, where instead of giving each SNP a specific weight, we removed the non-informative ones or variable selection models. Thus, we aimed to investigate the impact, in terms of accuracy of predictions, dispersion, and bias, of reducing the dimensionality of the G matrix by constructing it using only a subset of informative markers.

In order to accomplish this, we tried different machine learning and variable selection algorithms with the aim to identify the most informative SNPs by indirect prediction. These algorithms were as follows: least absolute shrinkage and selection operator (LASSO), spike-and-slab LASSO (SSLASSO), recursive feature elimination using ridge regression (RfeRR), recursive feature elimination using support vector machine regression (RfeSVM), and extreme gradient boost (XGBoost).

We aimed to test suitable procedures for genomic estimation by considering both the abovementioned variable selection models ssGBLUP and the predictions of BLUP, classical ssGBLUP, and WssGBLUP. To do that, we created different simulated populations and also considered a local population, the Rendena cattle. We then used different cross-validation methods to assess our results.

MATERIALS AND METHODS

For a graphical representation of our methodology for testing BLUP models, see **Figure 1**.



Data sets

Simulated Data sets

Simulations were performed with the QMSim simulation program (Sargolzaei and Schenkel, 2009). A total of four different populations were simulated based on various combinations of quantitative trait locus (QTL) number and effective population size (N_e). Each simulation was replicated five times.

All simulations were generated starting from the historical population using a similar structure to that used by Pocrnic et al. (2019): we created an initial bottleneck contracting the historical population size from 5,000 to 1,000 animals in 1,250 generations and then expanded it to 25,000. In the first generation, 10 bovine autosomes were simulated, placing evenly spaced 80,000 ca. biallelic SNPs with equal allele frequencies and a recurrent mutation rate of $2.5e^{-5}$ per generation. The number of SNPs per chromosome was set to 8,000, while the QTL number changed according to different simulation strategies. In two of the four simulations, one biallelic and randomly distributed QTL per chromosome was sampled from a gamma distribution with a shape parameter equal to 0.4 (oligogenic scenarios). In the other

TABLE 1 | Number of QTLs and effective population size in the four different simulated populations.

	QTL	N_e
SIM1	1,000	40
SIM2	10	350
SIM3	1,000	40
SIM4	10	350

two simulations, 100 QTLs per chromosome were generated using the same parameter (polygenic scenarios). In all these simulations, 10 discrete generations were created by randomly mating 750 females and a different number of sires according to the simulation strategies. In two scenarios, one oligogenic and one polygenic, we assumed a large N_e , with 100 males per generation used as sires, while in simulations with a low N_e , only 10 males per generation were used as sires. The following four populations were, thus, created by mixing different numbers of QTL and different N_e values, and five replicates for each population were generated:

TABLE 2 | Population structure of simulated and real data sets.

	Simulated		Real
	SIM1–SIM3 ^a	SIM2–SIM4 ^a	
Number of records	1,500	1,500	1,691
Number of animals in the pedigree	3,413	3,794	6,926
Number of genotyped animals	2,250	2,250	1,739
Number of genotyped animals with records	1,500	1,500	687
Inbreeding from pedigree	0.0126	0.0009	0.0316

^aSince population structure is the same for SIM1 and SIM3 and for SIM2 and SIM4, populations were grouped together in pairs in the table.

- SIM1 polygenic population with small Ne
- SIM2 polygenic population with large Ne
- SIM3 oligogenic population with small Ne
- SIM4 oligogenic population with large Ne

The effective population size and number of QTLs in the four different simulated populations are reported in **Table 1**, and numbers of -genotyped animals are reported in **Table 2** (2,250 animals). We simulated a single trait with heritability of 0.3, close to the heritability of the traits in the real data set further described. To do that, we obtained a single phenotype record per animal by adding an overall mean of 1.0 to the sum of the QTL effects together with a residual effect. As in the study by Pocrnic et al. (2019), only phenotypes from generations 8 to 9 were retrieved, and genomic information of animals belonging to generations 8 to 10 was used for further analysis ($750 \times 3 = 2,250$ animals). The structure of simulated populations is reported in **Table 2**. Before proceeding with genomic prediction, SNPs with a minor allele frequency (MAF <0.01) and with high linkage disequilibrium ($LD > 80$) were removed using the SNPrune program (Calus and Vandenplas, 2018).

Real Data set

A real data set containing information from the performance test evaluations of young bulls belonging to the Rendena cattle breed was provided by the National Breeders Association of Rendena (ANARE). ANARE also provided herdbook information about the whole population traced back to the 1950s, whereas genomic data of bulls were, in part, provided by ANARE (PSRN DualBreeding, www.dualbreeding.it) and, in part, obtained under academic funding (SID Project, BIRD183281). Rendena is a small local population (6,384 heads for 249 breeding males and 6,135 breeding females belonging to 202 herds censed on 31.12.2020; fao/dad.is.org) bred for the dual-purpose attitude of milk and meat. Rendena is native to the Northeastern Alps in Italy but is now widespread also in the adjacent lowland territory on the right side of the Brenta River in the Veneto region (Po Valley; Guzzo et al., 2018).

The real phenotypes considered in this study were single individual records of average daily gain (ADG), *in vivo* estimates of carcass fleshiness (CF) and dressing percentage (DP) collected in the years 1985–2020. These traits have been extensively described in Guzzo et al. (2019) and Mancin et al. (2021b). The Illumina Bovine LD GGP v3, comprising 26,497 SNP markers (low-density panel: LD), and Bovine 150K Array GGP v3 Bead Chip, including 138,974 SNPs (Illumina Inc, San Diego, CA, United States; high-density panel: HD), were used for genotyping Rendena cattle.

The LD panel belonging to 1,416 individuals with 26,497 SNPs was imputed on the HD panel with 138,974 SNPs belonging to 554 bulls. The overlap between the two panels was about 60%. Information about data quality control and imputation is reported in greater detail by Mancin et al. (2022). In addition to the previous study, further quality control was performed by removing SNPs with high linkage disequilibrium (>80), using SNPrune (Calus and Vandenplas, 2018): this removed a total of 28,049 SNPs. An amount of 85,331 SNPs was finally retained for analysis. Overall, the study considered 1,691 young bulls with only phenotypic information, 1,739 animals with only genotypic information, and 687 animals with both phenotypic and genotypic information. The data structure of the real data set used for genomic prediction is reported in **Table 2**.

Prediction Models

The breeding values for the single trait of the four simulated populations and the three performance test traits of the real Rendena data set were estimated using several BLUP models. First, we used four ‘classical’ BLUP models:

- 1) standard pedigree best linear unbiased prediction (PBLUP, described in *Pedigree Best Linear Unbiased Predictor*);
- 2) single-step genomic BLUP (ssGBLUP, described in *Single-Step Genomic Best Linear Unbiased Predictor*);
- 3) small shrinkage weighted single-step genomic BLUP (WssGBLUP1, described in *Weighted Single-Step Genomic Best Linear Unbiased Predictor*);
- 4) high shrinkage weighted single-step genomic BLUP (WssGBLUP2, described in *Weighted Single-Step Genomic Best Linear Unbiased Predictor*).

Then, we performed five ssGBLUPs with preselected SNPs (described in 2.2.4). SNP selection was achieved using the following algorithms:

- 5a) least absolute shrinkage and selection operator [LASSO, described in *Least Absolute Shrinkage and Selection Operator (LASSO)*];
- 5b) spike-and-slab LASSO (SSLASSO, described in *Spike-and-Slab LASSO*);
- 5c) recursive feature elimination using ridge regression [RfeRR, described in *Recursive Feature Elimination Using Ridge Regression (RfeRR)*];

- 5d) recursive feature elimination using support vector machine [RfeSVM, described in *Recursive Feature Elimination Using Support Vector Machine (RfeSVM)*];
- 5e) extreme gradient boosting (XGBoost, described in *Boosting Ensemble*).

Pedigree Best Linear Unbiased Predictor

PBLUP was the first method used to estimate predictors, and it is described by the following equation (Henderson, 1975):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where \mathbf{y} is the vector of phenotypic observations, \mathbf{X} is the matrix of the incidence of fixed effects, and \mathbf{b} is the vector of these effects. In the real data set, fixed effects are represented by the contemporary group (young bulls tested at the same period in the same pen; 142 levels) and parity group of dams in four classes (Guzzo et al., 2019). In the simulated data sets, \mathbf{X} was substituted by a vector of $\mathbf{1}$'s; thus, \mathbf{b} stands for the mean of the models. Matrix \mathbf{Z} represents the incidence matrix that relates the random genetic additive effect, included in vector \mathbf{a} , to the phenotype. The random residual error was included in a vector \mathbf{e} showing normal distribution $N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance. The vector of additive genetic effects is distributed as $N(0, \mathbf{A}\sigma_a^2)$, where σ_a^2 is the genetic variances and \mathbf{A} is the identical by descent (IBD) relationship matrix constructed from pedigree data.

Single-Step Genomic Best Linear Unbiased Predictor

We used ssGBLUP as a benchmark to evaluate the impact of other models (see further, WssGBLUP and ssGBLUP with selected SNPs). The ssGBLUP method presents the same structure of equation as in 2.2.1, except for the (co)variance matrix of random genetic effects, which is substituted by \mathbf{H} , as described by Aguilar et al. (2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where \mathbf{A} and \mathbf{A}_{22}^{-1} are the reverse of the IBD matrix for all animals and for only genotyped animals, respectively, and \mathbf{G} is the genomic matrix including the genomic relationships among animals.

The \mathbf{G} matrix was built using the first methods proposed by VanRaden (2008b):

$$\mathbf{G}_0 = \frac{\mathbf{M}\mathbf{M}'}{2\sum p_i(1-p_i)}$$

where p is the allele frequency of the i th locus and \mathbf{M} is a matrix of SNP content centered by twice the current allele frequencies. Since the frequencies of the current genotyped population are used to center \mathbf{G} , pedigree and genomic matrices have different bases, \mathbf{G} was adjusted so the average diagonal and off-diagonal matched the averages of diagonal and off-diagonal in \mathbf{A}^{22} , as described by Vitezica et al. (2011).

Weighted Single-Step Genomic Best Linear Unbiased Predictor

The WssGBLUP is the third method we used (two models, each with a different CT value, as explained below). This approach is equal to model 2.2.2, except for the matrix \mathbf{G} , built following the second method of VanRaden (2008a), as shown below:

$$\mathbf{G}_0 = \frac{\mathbf{M}\mathbf{D}\mathbf{M}'}{2\sum p_i(1-p_i)}$$

where p is the allele frequency of the i th locus, \mathbf{M} is a matrix of SNP content centered by twice the current allele frequencies, and \mathbf{D} is the diagonal matrix in which SNP specific weights are contained. The iterative algorithm reported by Zhang et al. (2016) has been used as a weighting strategy. The SNP weights presented in \mathbf{D} were obtained as a function of the estimated SNP effect (\hat{u}). The weighting function used in this study was called non-linear A, as reported by Fragomeni et al. (2019). This method was preferred over other weighting strategies due to its stability among the iterations. The iterative algorithm applied followed the steps reported below:

1. The initial parameter was set as $t = 1$, $\mathbf{D}_{(t)} = \mathbf{I}$, $\mathbf{G}_{(t)} = \frac{\mathbf{M}\mathbf{D}_{(t)}\mathbf{M}'}{2\sum p_i(1-p_i)}$, where \mathbf{I} is an identity matrix;
2. GEBV (\hat{a}) is obtained, where \hat{a} is the vector of solutions of additive genomic breeding value using the ssGBLUP algorithm;

3. The SNP effect (\hat{u}) is obtained, as in Gualdrón Duarte et al. (2014):

$$\hat{u} = \frac{1}{2\sum p(1-p)} \mathbf{D}\mathbf{M}' [\mathbf{M}\mathbf{D}\mathbf{M}']^{-1} \hat{a}.$$

4. $d_{i(t+1)}$, as in Fragomeni et al. (2019), is transformed in $\text{CT} \frac{|u_i|}{sd(u)}^{-2}$, where CT is a shrinkage factor determining how much the SNP effect distribution deviates from normality;
5. The weight of SNPs is standardized by maintaining a constant genetic variance among iterations:

$$\mathbf{D}_{(t+1)} = \frac{\text{tr}(\mathbf{D}_{(1)})}{\text{tr}(\mathbf{D}_{(t+1)})} \mathbf{D}_{(t+1)}.$$

6. Matrix \mathbf{G} is then recreated by including the new weights: $\mathbf{G}_{(t+1)} = \frac{\mathbf{M}\mathbf{D}_{(t+1)}\mathbf{M}'}{2\sum p_i(1-p_i)}$;
7. Set $t = t + 1$ and go to point 2 for a new iteration.

We created two different WssGBLUP models with two different CT values: WssGBLUP1 had a CT value of 1.105, while WssGBLUP2 had a CT value of 1.250. This process was carried out to grant WssGBLUP1 the lowest possible shrinkage effect and WssGBLUP2 the highest possible shrinkage effect. For both models, the maximum number of iterations was set to five. For simplicity, we reported only two WssGBLUP predictions instead of the 10 analyzed in this study (combination of two CT values and five iterations). Thus, we retained two opposite

WssGBLUP scenarios: WssGBLUP1, which presents the lowest SNPs shrinkage effect, and WssGBLUP2, which provides the highest shrinkage effect.

Single-Step Linear Unbiased Predictor With Only Informative SNPs

The last group of models (five models) consisted of ssGBLUP in which the **G** matrix of 2.2.2 was constructed using SNPs obtained after applying the different variable selection algorithms (described below, **Section 2.4**). The number of columns in **Z** is, thus, different for each trait and each data set.

Model Computations

A was built with the pedigree information tracking back up to three generations in all models. In addition, according to Cesarani et al. (2019), the variance components of each data set were estimated under PBLUP models by tracing back all animals in the pedigree. Variance components were estimated using the AIREml algorithm (Gilmour et al., 1995). All genetic and genomic prediction analyses were performed using the BLUPF90 family of programs (Aguilar et al., 2018). The consistency of all this information is reported in **Table 2**. Preliminary analysis, such as LD calculation, was conducted using preGSf90 (Aguilar et al., 2018, belonging to the BLUPF90 family of programs).

Featured Selection Algorithms

The EBVs of the target trait were used to map the major SNP markers associated with the phenotype, using five different statistical approaches. The genome content was considered a covariance matrix, while EBVs of genotyped animals (\hat{a}) (estimated using models in 2.2.2) were considered as the observed variable. The genome content was scaled in advance. Hyperparameter search and the choice of best models were performed by dividing the data set into a training group and a test group. In the real data set, young animals born after 2015 belonged to the test group, while older animals belonged to the training group. In the simulation, animals of 8th to 9th generations were part of the training group, while animals of the 10th generation belonged to the test group.

Least Absolute Shrinkage and Selection Operator

In the high-dimensional information literature, many penalized likelihood approaches have been proposed. Given the baseline $y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + e_i$, a variant of the penalized likelihood approach can be described as follows:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} - \frac{1}{2} \left\| \sum_{i=1}^N \left\{ y_i - \left(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j \right) \right\}^2 + \operatorname{pen}_{\lambda}(\beta) \right.$$

where N is the number of animals for each trait, β_0 is model mean, β_j is SNP contribution, p is the number of columns in **x**, N is the number of data, λ is the regularization parameter; and $\operatorname{pen}_{\lambda}(\beta)$ is a penalty function. In LASSO (Tibshirani, 1996), the penalty is as follows:

$$\operatorname{pen}_{\lambda}(\beta) = -\lambda \sum_{j=1}^p |\beta_j|$$

A grid search was performed to find the optimal values obtained by testing values from 0 to 20 in increments of 0.1. These values were used to maximize the LASSO model performance, based on the highest coefficient of determination and the lowest mean squared error (MSE) in the training set. To carry out this calculation, we used the glmnet R package (Friedman et al., 2010).

Spike-and-Slab LASSO

Spike-and-slab LASSO (SSLASSO) was proposed by Ročková and George (2018). It is based on the idea that every penalized likelihood has a Bayesian interpretation (Bai et al., 2021). For instance, the LASSO penalization is equivalent to a Laplace distribution regulated by hyperparameter λ , where the posterior mode of β is as follows:

$$p(\beta|\lambda) = \prod_{j=1}^p \frac{\lambda}{2} e^{-\lambda|\beta_j|}$$

The SSLASSO is the equivalent to a two-point mixture of Laplace distributions defined as follows:

$$p(\beta|\lambda) = \prod_{j=1}^p \left[(1 - \gamma_j) \left(\frac{\lambda}{2} e^{-\lambda_0|\beta_j|} \right) + \gamma_j \left(\frac{\lambda}{2} e^{-\lambda_1|\beta_j|} \right) \right]$$

where $p(\gamma|\theta) = \prod_{j=1}^p [\theta^{\gamma_j} (1 - \theta)^{1-\gamma_j}]$ and $p(\theta) \sim \operatorname{Beta}[a, b]$.

The Bayesian prior can be rearranged in a penalized likelihood context by taking this marginal logarithm as a prior (Bai et al., 2021); after some derivation, the following can be obtained:

$$\lambda_{\theta}(\beta_j) = \lambda_1 p_{\theta}(\beta_j) + \lambda_0 [1 - p_{\theta}(\beta_j)]$$

where

$$p_{\theta}(\beta_j) = \frac{1}{1 + \frac{(1-\theta)}{\theta} \frac{\lambda_0}{\lambda_1} \exp[-|\beta_j|(\lambda_0 - \lambda_1)]}$$

SSLASSO was computed using the SSLASSO R package (Ročková and George, 2018), error variances were assumed to be unknown, and a self-adaptive penalty was set. In so doing, θ was assumed to be random and different shrinkage was applied to each β_j .

Recursive Feature Elimination Using Ridge Regression

Similar to LASSO, ridge regression is based on a principle of penalized likelihood, with a penalty equal to $\lambda \sum_{j=1}^p \beta_j^2$. Before proceeding with recursive feature elimination, the optimal values of λ were obtained as in LASSO selection. The glmnet R package was used (Friedman et al., 2010).

After that, a recursive feature elimination using penalized ridge regression was performed as follows. In each iteration, the SNP effect β_j was estimated based on training data. Then, 10% of the variable with lowest $|\beta_j|$ was removed from the subsequent iterations. The variable (SNP) present in the iteration with the lowest mean squared error (MSE) was considered for the prediction. MSE was calculated as $(y_{\text{test}} - \hat{y}_{\text{test}})^2$, where y_{test} is the EBV which belongs to the test database and \hat{y}_{test} is the predicted one.

Recursive Feature Elimination Using the Support Vector Machine

The SVM is a kernel-based supervised learning technique, often used for regression analysis. Depending on the kernel function considered, the SVM can map linear or nonlinear relationships between phenotypes and SNP markers. The best kernel function to map genotype to phenotype was determined in different training subsets: a five-fold split was used to determine which kernel function was a better fit for the data, either with linear, polynomial, or radial basis. We found that performing the SVM with a linear basis function outperformed the polynomial and radial basis function of about 12.5% in predictive ability.

The general model for the SVM (Evgeniou and Pontil, 2005; Hastie et al., 2009) can be described as follows:

$$y_i^* = b + h(m) * w + e$$

where $h(m)$ represents the linear kernel basis function ($h(m) = m'm$) used to transform the original predictor variables (i.e., SNP marker information (m)), b denotes model bias, and w represents the unknown weight vector. In the SVM model, the learn function $h(m)$ was given by minimizing the loss function as follows: $C \sum_{i=1}^N L(y_i^* - \hat{y}_i^*) + \frac{1}{2} \|w\|^2$. The C represents a regularization parameter, which controls the trade-off between predictor error and model complexity, and w^2 denotes the squared norm under a Hilbert space. The SVM model was fitted using an epsilon-support vector regression that ignores residual absolute value ($|y_i^* - \hat{y}_i^*|$) smaller than some constant (ϵ) and penalizes larger residuals (Vapnik, 2000). The parameters C and ϵ were defined using the training data set as proposed by Cherkassky and Ma (2004): $C = \max(|\bar{y}^* + 3\sigma_{y^*}|, |\bar{y}^* - 3\sigma_{y^*}|)$ and $\epsilon = 3\sigma_{y^*} (\sqrt{\ln(n)/n})$, where \bar{y}^* and σ_{y^*} are the mean and standard deviation of the target EBV for the traits on the training population, respectively, and n represents the number of animals in the training set. The SVM was performed using the e1071 R package (Meyer et al., 2020).

After that, recursive feature elimination using the SVM was performed using the same procedure described for RfeRR in the study by Sanz et al. (2018).

Boosting Ensemble

The boosting approach (XGBoost) is an ensemble technique that combines gradient descent error minimization with boosting, aiming to convert weak regression tree models into strong learners (Hastie et al., 2009; Natekin and Knoll, 2013). This ensemble process combines different predictor variables sequentially in the regression tree model, using regularization *via* selection and shrinkage of the predictors to control the residual from the previous model (Friedman, 2002). In addition, the XGBoost can use parallel computation to use more regularized models to prevent overfitting. The XGBoost approach can be described as follows:

$$y = \sum_{w=1}^W \beta_w h(x, \gamma_w) + e$$

where y is the vector of the target EBV; W is the number of iterations (expansion coefficients); β_w is shrinkage factor, also

known as “boost”; $h(x, \gamma_w)$ is base learner, a function of the multivariate argument x with a set of parameters $\gamma_w = \{\gamma_1, \gamma_2, \dots, \gamma_w\}$; and e is the vector of the residuals. Expansions of the coefficients $\{\beta_w\}_1^W$ and parameters $\{\gamma_w\}_1^W$ are used to map the predictor variables (x), that is, SNP markers to the target EBV (y) considering the joint distribution of all values (y, x) and minimizing the loss function $L\{y_i, F(x)\}$ given as $[y, F_{w-1}(x_i) + h(y_i; x_i, p_w)]$, where p_w is the predictor to minimize $\sum_{i=1}^n L[y, F_{m-1}(x_i) + h(y_i; x_i, p_m)]$. Our XGBoost follows the algorithm specified by Chen and Guestrin, 2016. In the XGBoost method, a regularization term is added in the loss function, representing the weight vectors learned in the loss function: this term penalizes the ponderation of large weights. This regularization term is defined as follows: $\sum_{i=1}^n L[y, F_{m-1}(x_i) + h(y_i; x_i, p_m)] + \sum \Omega(f_n)$, where L is the error between the true value of the target trait and the predicted value and $\Omega(f_n)$ is the regularization function used to prevent overfitting: $\Omega(f_n) = \gamma T + 0.5 \lambda \omega^2$, where T is the number of leaves in the regression tree f_n and ω represents the weight for the leaves in each tree (i.e., the predicted values stored at the leaf nodes). Including in the objective function makes the tree less complex, which minimizes the loss function and helps reduce overfitting; γT is a constant penalty for each additional tree leaf, and $\lambda \omega^2$ penalizes extreme weights. The γ and λ are the regularization terms L1 and L2, respectively (Mitchell and Frank, 2017). The random search for XGBoost was performed considering the four most important parameters able to increase prediction accuracy and minimize the prediction error. These hyperparameters were Ntree (total number of trees in the sequence used in the model), learning rate (determines the contribution of each tree to the final model and performs shrinkage to avoid variable overfitting), maximum tree depth (controls the depth of the individual trees to be considered in the model), and minimum samples per leaf (controls the complexity of each tree). The Ntree values ranged from 600 to 5,000 in intervals of 200; the learning rate was in the range of 0.05–1 in intervals of 0.05; the maximum tree depth was determined with a value ranging from 5 to 80 in intervals of 5; the minimum sample per leaf was set from 5 to 100 in intervals of 5 and considering lambda and alpha regularization values ranging from 0 to 1 in intervals of 0.05. The random grid search XGBoost was performed using the *h2o.grid* function of the h2o R package (<https://cran.r-project.org/web/packages/h2o/>), considering as fixed parameter a maximum of 150 models with random combinations of the hyperparameters over 60 min.

Effective Population Size Calculations

The effective population size (N_e) has been computed from the individual increase in inbreeding (ΔF) (Falconer and Mackay, 1996) to compare real and simulated data properly. Individual ΔF was calculated as follows:

$$\Delta F = \frac{F_n - F_{n-1}}{1 - F_{n-1}}$$

$$N_e = \frac{1}{2\Delta F}$$

TABLE 3 | Description of different validation sets used in cross-validation. The first and last years of birth of animals in the training data set and the number of animals (individuals) used in the validation cohort are reported.

First	Last	Individuals
2012	2020	178
2013	2020	154
2014	2020	130
2015	2020	109
2016	2020	106
2017	2020	72
2018	2020	45

where F_n is the inbreeding in the n th generation. N_e was calculated using the *purge* R package (<https://cran.r-project.org/web/packages/purgeR>).

Validation

Validation in the Simulated Data set

Quality of prediction was measured as the correlation and MSE between the genomic breeding values estimated under different models and the true breeding values for animals belonging to the 10th generation, that is, the last generation of animals, including individuals without phenotypes but with genotype.

Validation in the Real Data set

In the real data set, two different cross-validation methods were applied. The first method we used to cross validate predictive ability was to calculate both the correlation and MSE between predicted and observed phenotypes. In this case, five-fold cross-validation with 10 iterations was performed. Since not all animals were genotyped in each iteration, 1/5 of non-genotyped and 1/5 of genotyped animals were masked. The current study considered predictive ability metrics only for genotyped animals; however, results about non-genotyped animals were also obtained (**Supplementary Figure S1**).

Linear regression (LR) (Legarra and Reverter, 2018) was used as the second cross-validation method. It compares the prediction performances of different models on groups of focal individuals born after a given date, in this case, the young bulls. LR is particularly suited to the specific needs of the Rendena population since predicting the future performance of young bulls without phenotype is one of the main objectives of the breeding plans for performance tests (Mancin et al., 2021a).

The LR method evaluates the goodness of a model by comparing its performance in a complete data set and a partial data set. The complete data set contains the whole amount of information or it is the data set used for prediction. A partial data set is referred to as the complete data set with some animals with the phenotype removed, usually young animals known as candidates to selection. According to Macedo et al. (2020), we built partial data sets by excluding phenotypes since a target recent birth year of young bulls (since 2012–2020; since 2014–2020; and since 2017–2020) to describe possible variations and random deviations of the estimator; consistencies are reported on **Table 3**. LR considered the following three parameters: bias, dispersion, and accuracy. Bias is the difference between the expected breeding values

estimated under the complete and partial data sets. Dispersion was calculated as the regression coefficient considering the breeding values from the complete data set on the ones estimated from the partial data and accuracy as correlations between the two breeding values.

RESULTS

Genomic Structure

Genomic Structure in Simulated Data sets

Figure 2 highlights the different genomic assets of small N_e populations (SIM1 and SIM3; 10 sires per generation) and large N_e populations (SIM2 and SIM4; 200 sires per generation). Since the different number of QTLs assumed for the populations with the same N_e (that is, 10 vs. 1000 QTL) did not impact **G** matrix dimensionality, only SIM1 and SIM2 were plotted for simplicity. In SIM1, 193 eigenvalues were necessary to explain 98% of **G** matrix variance, while in SIM2, 795 eigenvalues were necessary to explain 98% of **G** matrix variance. When only ten sires per generation were used, it was possible to observe different subpopulations (**Supplementary Figure S2**); however, no population structure was found when plotting the first two eigenvalues. On the other hand, SIM2 appeared homogenous, and individuals seemed almost unrelated. In addition, when LD per chromosome was calculated, a greater value was observed in SIM1 (0.161 ± 0.076) than that in SIM2 (0.067 ± 0.054 ; data not shown). An N_e value of 81.18 ± 4 was determined for SIM1 and SIM3 and 1869 ± 546 for SIM2 and SIM4.

Genomic Structure in the Real Data set

We also investigated **G**'s dimensionality on the real data set of the Rendena cattle population (**Figure 3**). The real data set presented a situation closer to SIM1 and SIM3 than to SIM2 and SIM4. It showed, indeed, an average N_e value of 108.2 ± 0.74 calculated from pedigree data. It is possible to observe a few clusters in the genomic relationship matrix (**Figure 3**); however, they are not as

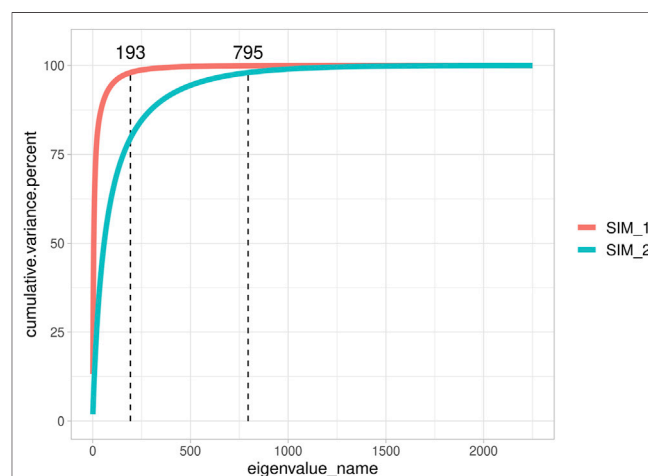
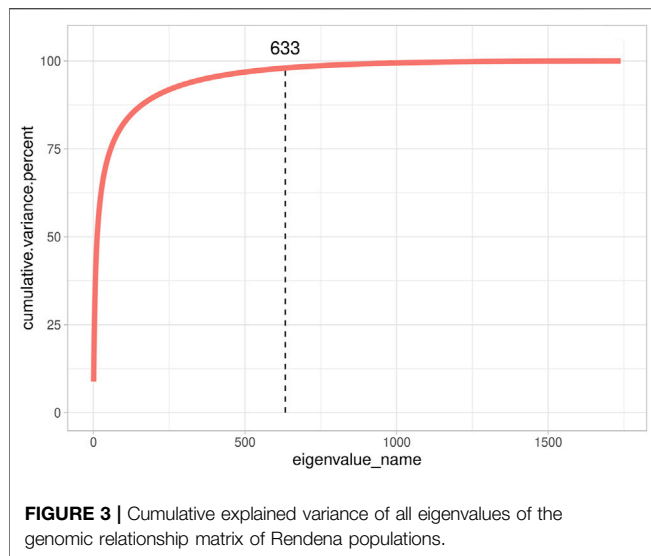


FIGURE 2 | Cumulative explained variance of all eigenvalues of the genomic relationship matrix of two representative simulated populations.



straightforward as in SIM. We, therefore, can note that no population structure is present in Rendena breed (i.e., no subpopulations), which is in line with previous research (Mancin et al., 2022). Only 633 eigenvalues explained 98% of G variance; thus, the scenario was closer to SIM2 than SIM1. In addition, we observed an average LD of 0.187 ± 0.107 per chromosome (Mancin et al., 2022).

SNPs Retained by Variable Selection Models

SNPs Retained in Simulated Data sets

Figure 4 reports the impact of the different algorithms in terms of the number of informative markers retained. Specifically, we were interested in identifying the impact that different **G** matrix dimensionality and number of QTLs had on the number of SNPs considered informative. In all simulations, LASSO and SSLASSO retained the lowest number of SNPs (roughly 2,000 SNPs averaged across simulations), and they presented lower intra- and between-scenario variability. On the contrary, RfeSVM and RfeRR algorithms retained higher numbers of SNPs, on average 12,000 for RfeRR and 7,000 for RfeSVM. RfeSVM also presented an extreme variability across scenarios (**Figure 4**). XGBoost retained an intermediate number of SNPs, with an average of 3,000 SNPs retained across simulations. As shown in **Figure 4**, different numbers of QTLs did not affect the number of SNPs retained by each algorithm. In fact, no difference was observed between SIM1 vs. SIM3 and SIM2 vs. SIM4; only LASSO and SSLASSO algorithms seem to be slightly affected by the number of QTLs. Interestingly, the dimensionality of the **G** matrix seems to be more influential as scenarios with higher N_e presented a higher number of SNPs (SIM1 and SIM2). The XGBoost is the only algorithm where this trend was not seen. Crucially, we observed that the negative gap in model accuracy present in simulations with lower QTL (SIM3 and SIM4) fades when variable selection models are introduced.

SNPs Retained in the Real Data set

We showed the impact of variable selection methods regarding the number of informative markers retained in the Rendena population in **Figure 5**. Although the number of initial SNPs was similar to that of the simulated populations, in general, the algorithms retained a higher number of SNPs in the real data set. Similar to what was reported in the simulated data, LASSO and SSLASSO were the most restrictive algorithms of SNP selection, with an average of 2,000 SNPs retained across the simulations. The XGBoost was the second most restrictive algorithm in terms of SNPs retained by the models, about 3,000 on average. RfeSVM and RfeRR algorithms retained almost half of the SNPs presented in the panels (40,000 SNPs). No clear patterns were identified across different phenotypes: some algorithms found a greater number of SNPs in certain traits and some in others. For example, the lowest number of informative markers retained by RFE algorithms was identified on the DP trait, but the opposite situation occurred for XGBoost, where the number of informative SNPs retained for DP was almost twice the number of informative SNPs retained for other traits.

Breeding Value Prediction

We compared the prediction accuracy of four ‘classical’ models for BLUP and ssGBLUP with five different SNP preselection strategies. The models are detailed in *Materials and Methods* and summarized as follows: 1) PBLUP; 2) single ssGBLUP; 3) WssGBLUP1; 4) WssGBLUP2; 5a) ssGBLUP with SNPs preselected via LASSO; 5b) ssGBLUP with SNPs preselected via SSLASSO; 5c) ssGBLUP with SNPs preselected via RfeRR; 5d) ssGBLUP with SNPs preselected via RfeSVM; and 5e) ssGBLUP with SNPs preselected via XGBoost. **Table 4** provides a qualitative summary of the results, described in the following paragraphs.

Breeding Value Prediction in Simulated Data sets

Different prediction model accuracies are reported in **Figure 6**, with correlation and MSE as comparison metrics. MSE values were comparable to those obtained for correlations. Standard BLUP models achieved the lowest accuracy. A substantial increase in accuracy was observed in ssGBLUP models, that is, when genomic data were integrated: this increase of accuracy was more relevant for populations with small N_e (SIM1 and SIM3).

A slightly greater accuracy than that in ssGBLUP was observed when a heterogeneous distribution of SNPs was considered within the matrix **G** (WssGBLUP). The gap in accuracy was greater in the populations with few QTLs (SIM3 and SIM4), especially for WssGBLUP2. On the other hand, the increase in accuracy for SIM1 and SIM2 under WssGBLUP was almost close to zero. A substantial variation in accuracy values was observed when ssGBLUP was performed with **G** matrixes constructed with selected SNPs; however, the accuracy of the prediction performance of each variable selection model changed according to the simulation structure. Generally, SSLASSO presented the highest increase in accuracy among the genetic models in all simulations, except for SIM2, where we observed a dramatic drop in accuracy. On the other hand, LASSO achieved

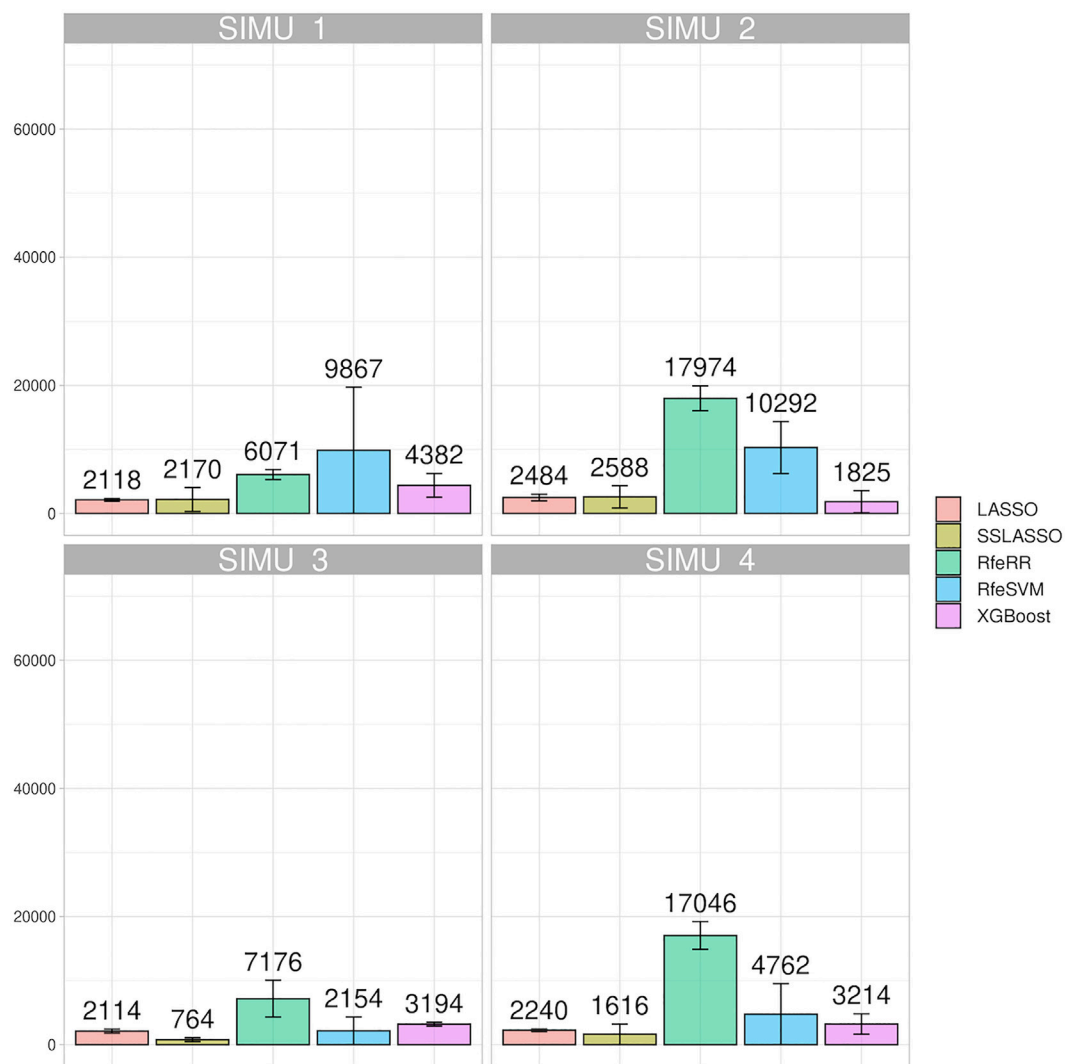


FIGURE 4 | Bar plot representing the number of SNPs retained by each algorithm on the four simulated population; error bars represent standard deviation.

greater accuracy on both SIM1 and SIM2. Other algorithms presented an intermediate increase in accuracy among the genetic models in all simulations, namely, RfeRR, RfeSVM, and XGBoost, with different rankings in different scenarios.

Breeding Value Prediction in Real Data set

With our real data sets, we were first interested in evaluating the performance of these models in terms of prediction; then, we wanted to evaluate the feasibility of introducing them in a real breeding plan scenario. This point was achieved using LR cross-validation methods (Legarra and Reverter, 2018). **Figure 7** presents the results of repeated five-fold cross-validation. The integrations of genomic data led again to a substantial increase in accuracy: the PBLUP presented the overall lowest correlation (r) values (r from 0.36 to 0.53). The ssGBLUP presented the lowest correlation values among genomic models (r from 0.46 to 0.62), while a slight increment was observed for WssGBLUP1 (from 0.55 to 0.67) and for WssGBLUP2 (from 0.67 to 0.75). As with

simulated data, variable selection models improved model accuracy substantially. Again, the highest correlations were found for LASSO and SSLASSO, with values of r ranging from 0.83 to 0.92, while other algorithms presented intermediate values (r around 0.70). This pattern was observed across all traits. MSE reflected the results obtained with correlations.

LR methods evaluated dispersion and bias in addition to accuracy. **Figure 8** represents the different results obtained using LR cross-validation methods in various validation sets from 2015–2020. This set of years was chosen as representative of all seven validation cohorts. **Figure 9** reports the summary statistics of all seven validation cohorts.

Accuracy trends of the real data set measured with the LR method were similar to the accuracies obtained with five-fold cross-validation. However, looking at the other statistics (slope and bias), we can observe that LASSO, SSLASSO RfeRR, and RfeSVM cannot be considered suitable variable selection approaches in real breeding plans due to their higher bias and

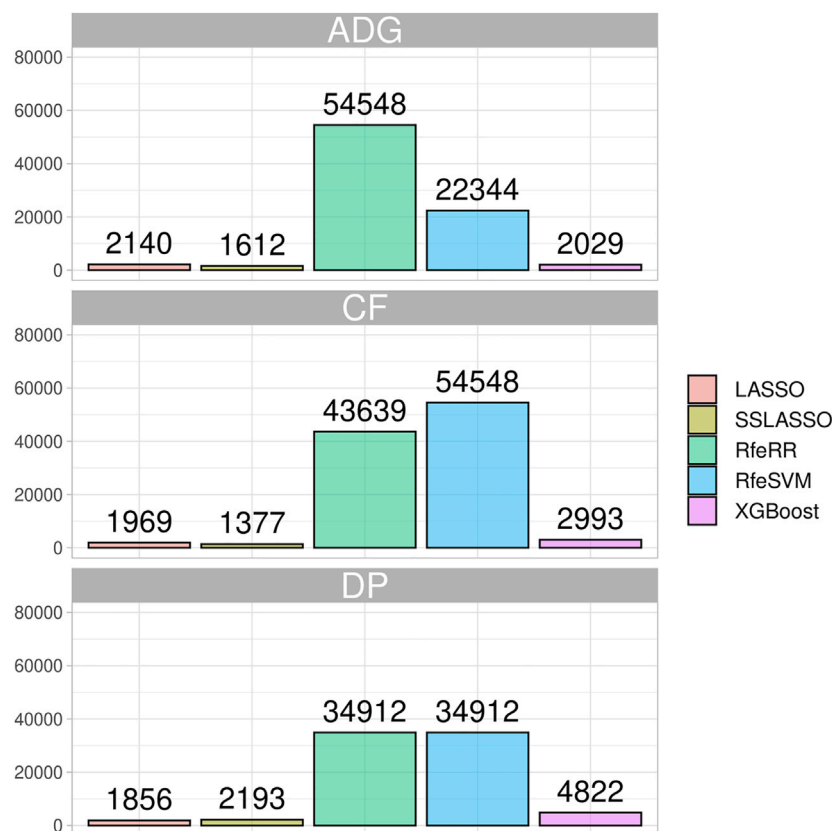


FIGURE 5 | Bar plot representing the number of SNPs retained by each algorithm on the three phenotypes of the Rendena population.

TABLE 4 | Summary of results obtained using the nine models considered in the study and the cross-validations applied.

Method name	Accuracy across simulations (Correlation/MSE)	Accuracy in real data set (Correlation/MSE)	Bias/Slope in LR cross-validation in real data set
PBLUP	Poor	Poor	Good
ssGBLUP	Medium	Medium	Best
WssGBLUP1	Medium	Medium	Good
WssGBLUP2	Medium	Good	Poor
LASSO-selected ssGBLUP	Best	Best	Poor
SSLASSO-selected ssGBLUP	Best	Best	Poor
RfeRR-selected ssGBLUP	Good	Good	Poor
RfeSVM-selected ssGBLUP	Good	Good	Poor

dispersion values, especially if compared with ssGBLUP. Conversely, XGBoost was the only model with similar or even lower bias and dispersion values than ssGBLUP but with greater accuracy. As seen in **Figure 9**, we demonstrate that these trends are consistent over different validation cohorts.

DISCUSSION

The present study had two objectives: testing if reducing the number of SNPs used to construct **G** could lead to an increase in

the accuracy of (ss)GBLUP and whether this method could be introduced in genomic evaluations of a real population with a small size, such as the Rendena breed.

In our study, using both simulated and real data sets, we demonstrated that the accuracy of (ss)GBLUP increases when it is performed when the number of SNPs to construct **G** was reduced. This finding agrees with that of the extensive literature supporting the increased accuracy of Bayesian variable selection models in many different species (Lourenco et al., 2014; Mehrban et al., 2021; Yoshida et al., 2018; Zhu et al., 2021). For example, Akbarzadeh et al. (2021) integrated only a subset of chosen

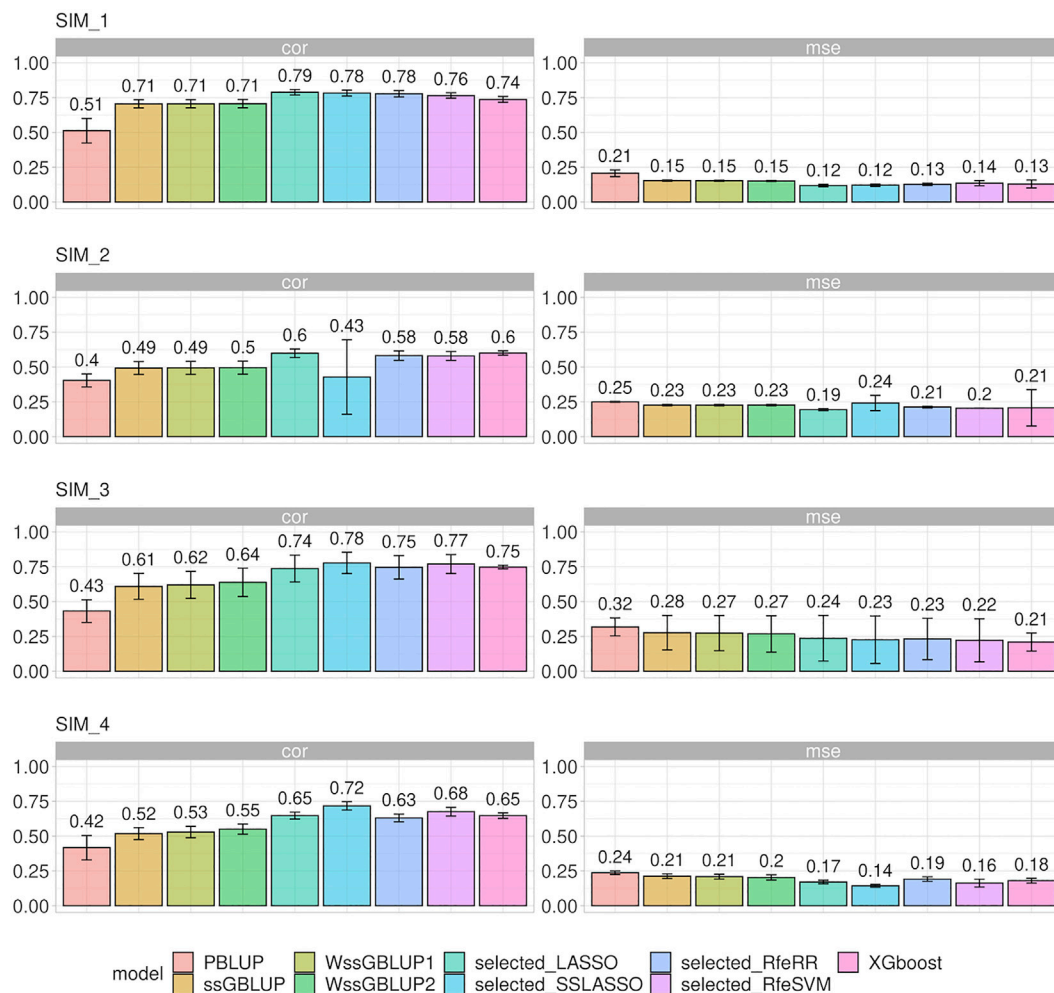
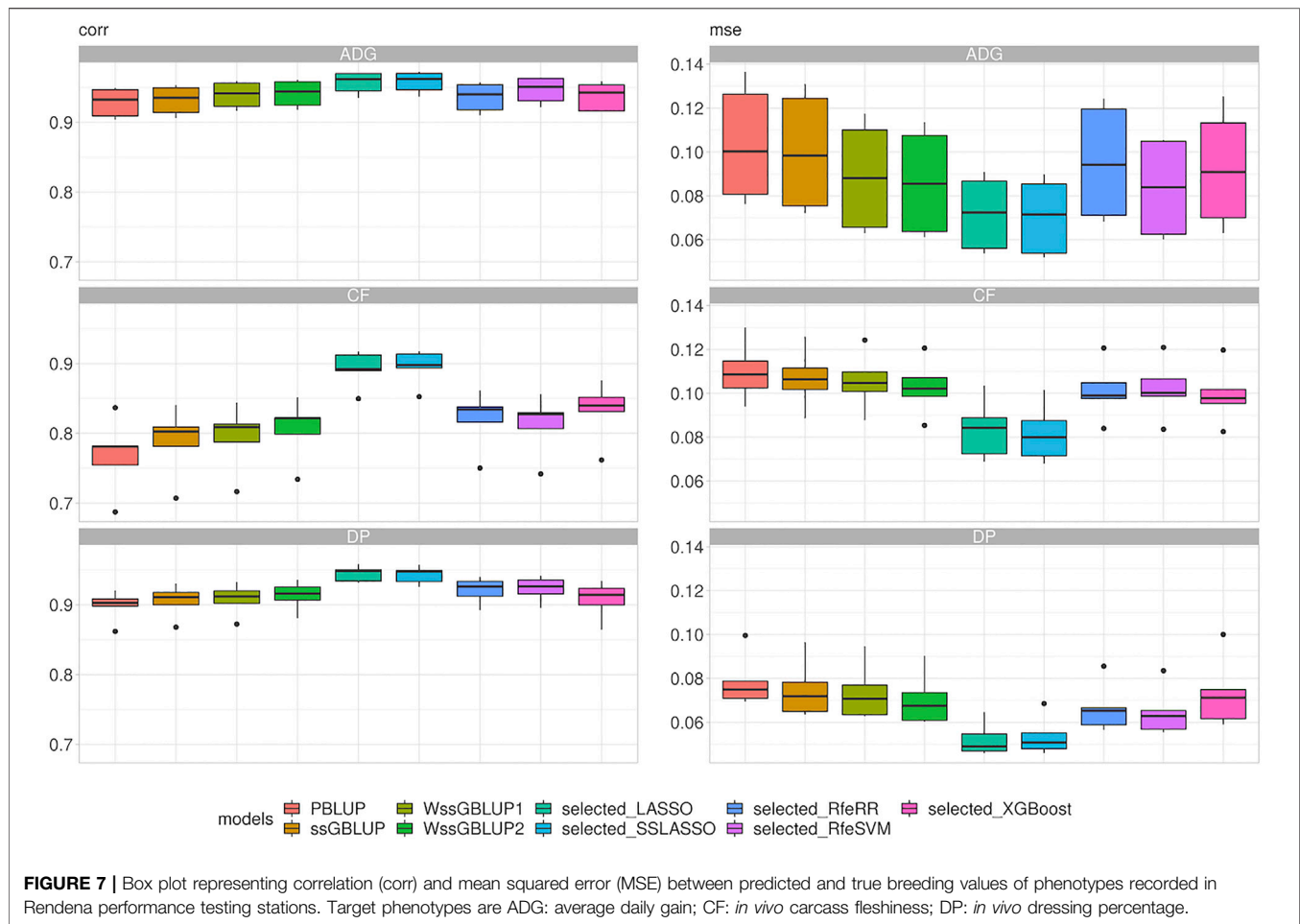


FIGURE 6 | Bar plot representing correlation (corr) and mean squared error (MSE) between predicted and true breeding values on the four different simulated populations. Error bars represent standard deviations.

SNPs into the GBLUP framework based on a classical GWAS analysis (i.e., 1, 5, 10, and 50% of significant SNPs). A slightly greater accuracy than that in the canonical GBLUP was observed when \mathbf{G} was constructed using only the best 10 and 50% SNPs; contrariwise, models using the 1 and 5% of the SNP prediction underperformed. Furthermore, Akbarzadeh et al. (2021) reported a dramatic decline in performance when the same percentage of SNPs was randomly chosen. We tried preliminary tests of a similar approach—construction of the \mathbf{G} matrix using the top 500, 1,000, and 50,000 SNPs ranked by their absolute SNP effect values calculated through back solutions—in Rendena breed; however, we immediately discarded this approach because of the extreme bias and inflated breeding value predictions (these findings are reported by Mancin et al., 2022 in press). In addition, choosing so few and unrepresentative SNPs reduced a lot the compatibility between \mathbf{A} and \mathbf{G} matrices, and thus ssGBLUP properties were affected (Misztal et al., 2013).

Li et al. (2018) and then Piles et al. (2021) showed how using different methods to select the most informative SNPs could

significantly improve the performance of the variable selection models. Li et al. (2018) constructed the \mathbf{G} matrix using the best 400, 1,000, and 3,000 SNPs, ranking SNPs effects by three different machine learning models. As in the previous case, an increase in accuracy was obtained only with a certain number of selected SNPs (1,000 SNPs), while a lower accuracy than that in the canonical GBLUP was observed with a lower number of SNPs. In addition, Piles et al. (2021) and Azodi et al. (2019) showed that by combining different variable selection algorithms with various parametric and nonparametric prediction models (i.e., ensemble predictions), it is possible to obtain a consistent increase in accuracy compared to models without variable selection. However, our study has not explored these scenarios since prediction methods other than ssGBLUP or ssSNP-BLUP (Fernando et al., 2017) do not seem to bring any concrete improvement for livestock traits (Abdollahi-Arpanahi et al., 2020). Furthermore, ssGBLUP and ssSNP-BLUP are the only methods that allow combining straightforwardly non-genotyped animals with genotyped ones—a crucial feature for a real-life



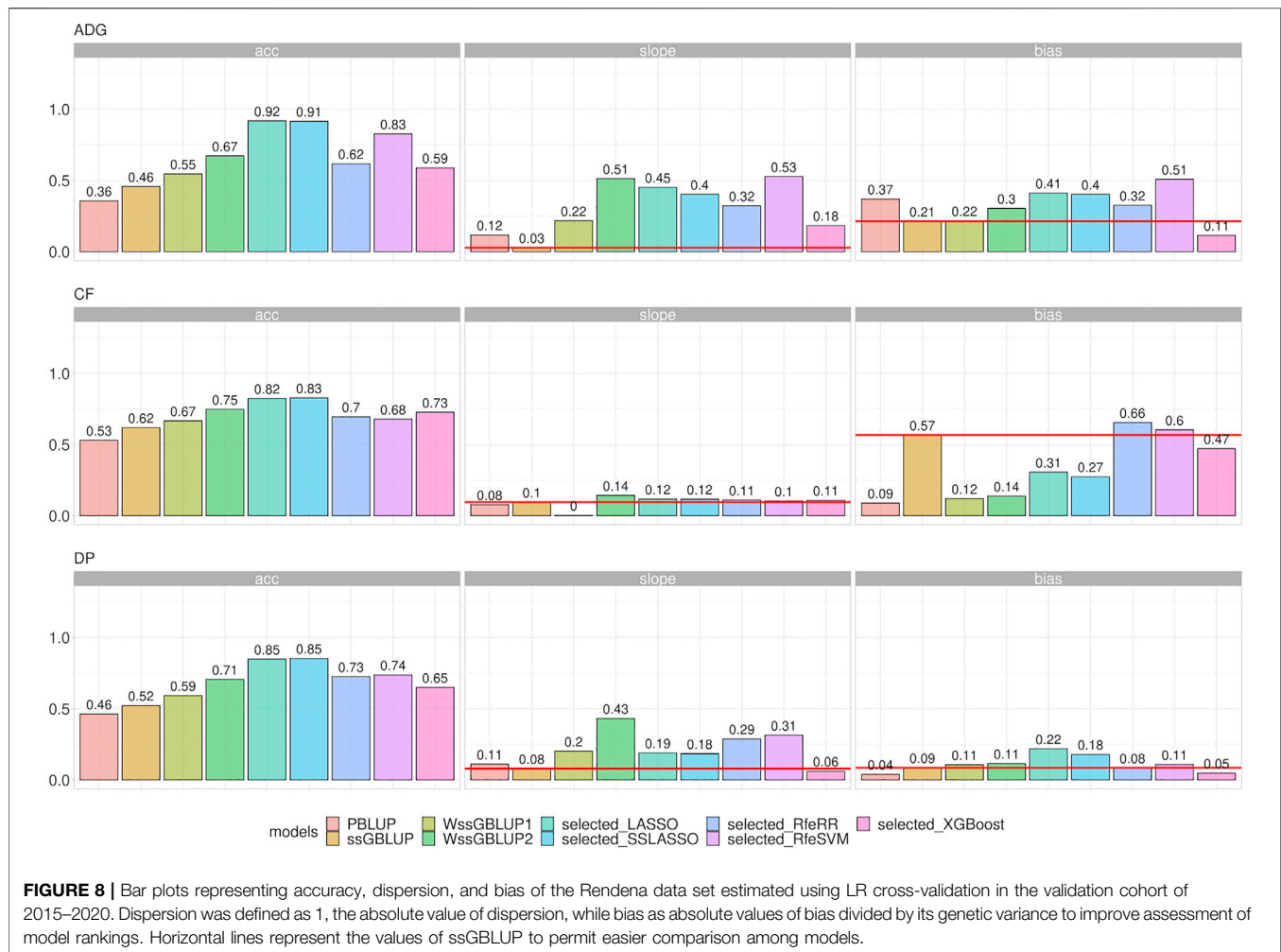
routine selection plan and something that the other algorithms cannot do.

Our result that reducing the number of parameters positively impacts accuracy is also supported by Frouin et al. (2020). In that study, it was demonstrated that the error of the prediction tends to linearly increase when $n > p$ until the “irreducible” error ($1 - h^2$) occurring when $n \gg p$. In addition, Pocrnic et al. (2019), demonstrated that the accuracy of (ss)GBLUP is connected by the distribution of eigenvalues of \mathbf{G} ; thus, “ n ” becomes the number of independent chromosome segments (Me) captured by SNPs (Pocrnic et al., 2019). In highly related populations (small N_e), higher accuracy values can be achieved than in populations with larger N_e because fewer eigenvalues and thus a small “ n ” are necessary to explain \mathbf{G} . As a matter of fact, in large N_e populations, more data are needed to increase accuracy. This issue is also intuitive since prediction error accuracy (Henderson, 1975) is directly proportional to the coefficient C^{aa} (defined below); thus, in highly related populations, C^{aa} tends to have lower values. C^{aa} is the inversion of the coefficient matrix of the mixed model equation where aa is the block referring to the genetic effect of animals. What was reported by Pocrnic et al. (2019) could explain the lower performance identified by Akbarzadeh et al. (2021) when subsets of 1 and 5% of SNPs were considered (Akbarzadeh et al., 2021). Indeed, discarding too many SNPs from the construction of \mathbf{G}

may omit the inclusion of important eigenvalues. From another perspective, Fragomeni et al. (2017) demonstrated the positive impact of removing non-informative SNPs on GBLUP. The authors showed in a simulated data set that better accuracy was found when the \mathbf{G} was built by eliminating all SNPs outside the window where the QTL was situated or using only QTL information. However, a practical limit to this method is that knowing all the QTLs within a genome is nearly impossible, especially when the population is small (Mancin et al., 2021a).

Our simulated results support the abovementioned theory, as simulations with lower N_e presented higher accuracy of ssGBLUP (SIM1, SIM3). Furthermore, differences between scenarios emerge when comparing simulations differing for their number of QTLs. ssGBLUP showed lower performance in SIM3 and SIM4 (QTL10) than in SIM1 and SIM2 (QTL1000); however, this discrepancy in accuracy decreases by applying variable selection. This result agrees with that by Daetwyler et al. (2010), which demonstrated that SNP selection *via* Bayes B presents substantial advantages when the number of QTLs is small compared to the number of independent chromosome segments.

As mentioned above, Bayesian SNP regression, or (ss)GBLUP using a weighted realized relationship matrix (Tiezzi and Maltecca, 2015; Zhang et al., 2016), always provides higher prediction accuracy than models assuming homogenous



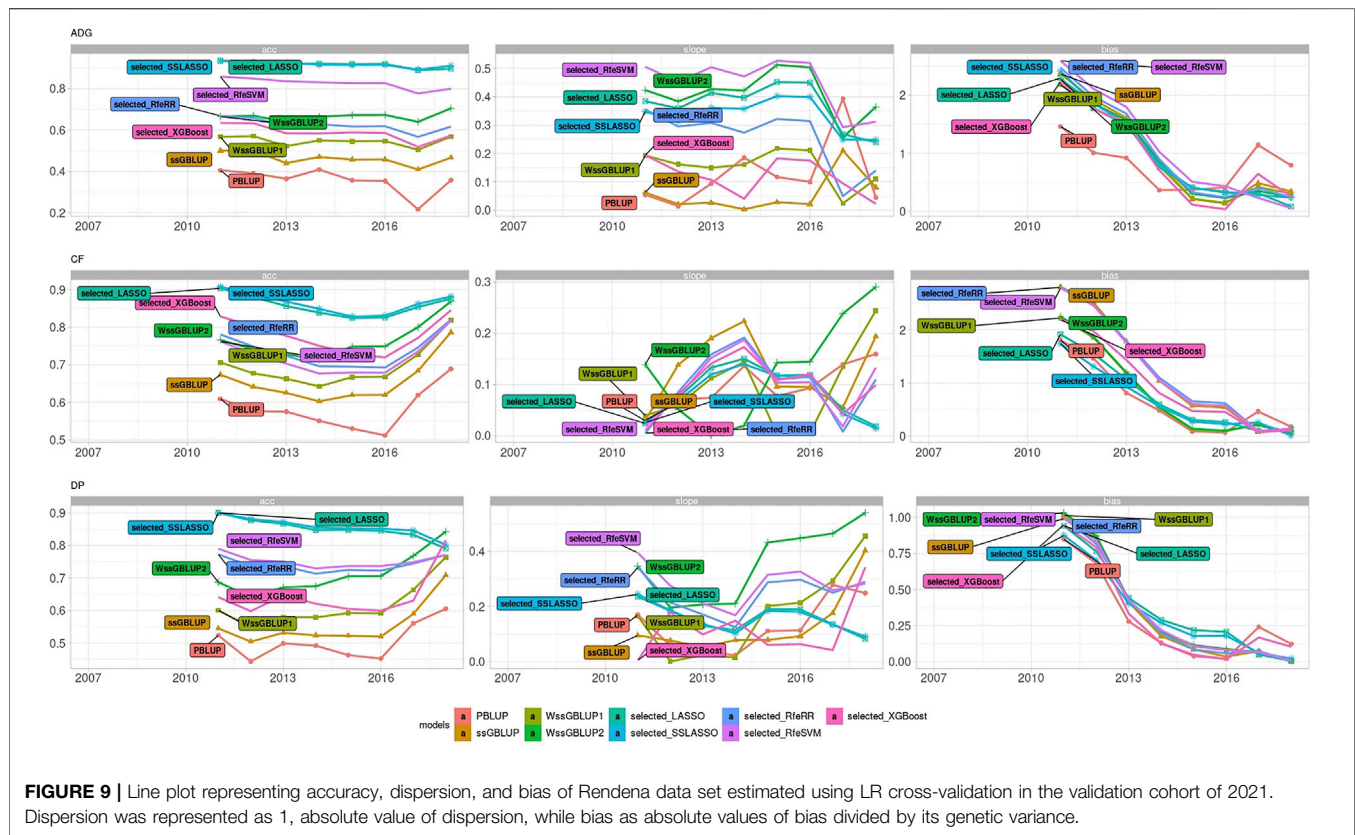
variance among SNPs (GBLUP or SNP-BLUP). However this increase in accuracy is often connected with increases in bias, especially when time cross-validation is used (Mehrban et al., 2021), instead of five-fold or leave-one-out cross-validation (Zhu et al., 2021). However, when the goal is to achieve the “best predictor”, namely, a value closer as possible to real one, models assuming heterogenous variances and models with variable selection can be identified as the best models. They have, indeed, the highest MSE, intended as bias-variance trade-off (Gianola, 2013). In this regard, LASSO and SSLASSO, thus, appeared as “best models” for both simulated and real data. We showed that (SS)LASSO regression performs automatic feature selection, especially in the presence of linearly correlated features, such as SIM1 and SIM3, since their simultaneous presence will increase the value of the cost function. Thus, LASSO regression will try to shrink the coefficient of the less important SNPs to 0 to select the best features.

However, in real-life breeding scenarios, time cross-validation must be considered (Liu, 2010; Legarra and Reverter, 2017) as this procedure simulates the natural accumulation of information across time. Only a few studies evaluated the impact of

heterogenous or variable selection models using time cross-validation with small samples of individuals. Cesarani et al. (2021) and Mancin et al. (2021b) found higher bias and overdispersion values in WssGBLUP than in ssGBLUP.

The same pattern emerged when we performed LR cross-validation (Mancin et al., 2021b; Cesarani et al., 2021), namely, that higher shrinkages or selected SNPs have high accuracy but carry higher bias and dispersion values. Specifically, (SS)LASSO models showed the best accuracy in all three traits when measured with LR. Conversely, other feature selection models and WssGBLUP presented lower accuracy. Among the variable selection models, we found slightly lower values of accuracy in the XGBoost; however, we suggest that XGBoost could be regarded as the best variable selection model among those tested as it is the only model that presented higher accuracy than ssGBLUP, at a net of better bias and dispersion.

Several questions persist about the use of these models in routine evaluation. One of these issues concerns the implementation of preselected SNPs in multitrait models. However, this is a recurring problem not only when the **G** matrix is built with preselected SNPs but also more in general whenever models take into account the specific genomic architecture of traits, as WssGBLUP does. A



possible solution to bypass this issue might be using multiple **G** matrix prediction models, one for each trait: yet, this is not computationally straightforward. A preliminary selection of SNPs by multiobjective optimization framework algorithms, as in Garcia (2019), could be a more concrete approach for future studies.

Another possible concern about the large-scale use of variable selection ssGBLUP is the fluctuations of SNPs across generations. Similarly to the issue with multitrait models, this regards all genomic selections (Hidalgo et al., 2020); however, it is true that with respect to other methods, such as Bayesian SNP regression, generation-by-generation recalibration of SNP preselection algorithms can be highly computationally demanding, especially when algorithms such as XGBoost are chosen. Finally, SNP preselection could be influenced by variability in SNP frequency across animals or more in general in the presence of population structure, as with subpopulations. Nonetheless, in our study, the PCA plots referring to SIM1 (**Supplementary Materials S2**), where some clusters are present, show that variable selection models overcome this issue quite effectively. It would be interesting to choose one or more variable selection models in future studies and evaluate their impact on more stratified populations.

Besides increasing the EBV accuracies, developing an optimal strategy for SNP variable selection in high-density panels will be particularly useful in local breeds. It would in fact allow the use of informative but lower density and cheaper panels, accounting for the best SNPs suitable for the target trait and population. Furthermore, given that small breeds cannot attract the same

level of technological investment as their cosmopolitan counterparts (e.g., Holstein), decreasing the costs of genomic selection could be critical to help guarantee their selection, and thus their survival.

Aside from the economic factors, the importance of developing *ad hoc* selection methods for small-population cattle, especially for local breeds, is of primary importance for their conservation. Maintaining genetic progress for the productive characters and at the same time keeping intact the genetic variability and the distinct characteristics of the breeds can be guaranteed through breeding plans implementing careful selection (Biscarini et al., 2015). These plans are needed to preserve genetic variability within livestock local populations, a goal which, in the medium term, is critical for the animal husbandry industry to ensure the conservation of native breeds, their productive and reproductive efficiency, health, survival, and overall resilience to future changing environmental pressures (Mastrangelo et al., 2014).

CONCLUSION

Genomic information, especially the single-step GBLUP technique, has brought great improvements to selection and breeding decisions in livestock. However, these methods still present methodological issues when applied to populations with a small size, such as local and endemic cattle breeds. Our rigorous testing of different algorithms for variable selection of

informative SNPs has highlighted that prediction accuracy of variable selection ssGBLUP (especially that of XGBoost) was greater than that of other ssGBLUP methods, without the inflated bias and dispersion that accompany the weighted ssGBLUP. Our use of machine learning models could thus represent a solution to the issue of genomic selection in small populations. Local cattle breeds are an often untapped resource of genetic diversity and have great potential to adapt to varying environmental conditions. The methods presented here might, thus, be used in their conservation, study, and increase their economic competitiveness.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below https://datadryad.org/stash/share/O6ld-ZCZLhAUtXmVpOlXkhkffVagc1_Stfnqtk907w.

ETHICS STATEMENT

Ethical review and approval were not required for the animal study because this study did not require any specific ethics permit. The cattle samples belonged to commercial private herds and were not experimentally manipulated. The samples were collected by technicians from the Breeders Association of Rendena.

REFERENCES

- Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep Learning versus Parametric and Ensemble Methods for Genomic Prediction of Complex Phenotypes. *Genet. Sel. Evol.* 52, 1–15. doi:10.1186/s12711-020-00531-z
- Aguilar, I., Tsuruta, S., Masuda, Y., Lourenco, D. A. L., Legarra, A., and Misztal, I. (2018). “BLUPF90 Suite of Programs for Animal Breeding,” in *The 11th World Congress of Genetics Applied to Livestock Production* (Auckland, New Zealand, 11, 751).
- Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot Topic: A Unified Approach to Utilize Phenotypic, Full Pedigree, and Genomic Information for Genetic Evaluation of Holstein Final Score. *J. Dairy Sci.* 93, 743–752. doi:10.3168/jds.2009-2730
- Akbarzadeh, M., Dehkordi, S. R., Roudbar, M. A., Sargolzaei, M., Guity, K., Sedaghati-khayat, B., et al. (2021). GWAS Findings Improved Genomic Prediction Accuracy of Lipid Profile Traits: Tehran Cardiometabolic Genetic Study. *Sci. Rep.* 11, 1–9. doi:10.1038/s41598-021-85203-8
- Alvarenga, A. B., Veroneze, R., Oliveira, H. R., Marques, D. B. D., Lopes, P. S., Silva, F. F., et al. (2020). Comparing Alternative Single-step GBLUP Approaches and Training Population Designs for Genomic Evaluation of Crossbred Animals. *Front. Genet.* 11, 263. doi:10.3389/fgene.2020.00263
- Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3 Genes, Genomes, Genet.* 9, 3691–3702. doi:10.1534/g3.119.400498
- Bai, R., Ročková, V., and George, E. I. (2021). Spike-and-Slab Meets LASSO: A Review of the Spike-And-Slab LASSO. *Handb. Bayesian Sel.*, 81–108. –108. doi:10.1201/9781003089018-4
- Biscarini, F., Nicolazzi, E. L., Stella, A., Boettcher, P. J., and Gandini, G. (2015). Challenges and Opportunities in Genetic Improvement of Local Livestock Breeds. *Front. Genet.* 6, 33. doi:10.3389/fgene.2015.00033

AUTHOR CONTRIBUTIONS

Idea: EM; conceptualization: EM, CS, BT, LM, and RV; methodology: EM and LM; formal analysis: EM and LM; support to analysis: BT; investigation: BT, CS, RM, and EM; resources: RM and RV; data curation: EM and RM; writing—original draft preparation: EM and BT writing—review and editing: LM, CS, and RM. All authors have read and agreed to the published version of the manuscript.

FUNDING

The study was funded by the DUALBREEDING project (CUP J61J18000030005) and by BIRD183281.

ACKNOWLEDGMENTS

The authors are grateful to the National Breeders Association of Rendena (ANARE) for data support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.814264/full#supplementary-material>

- Blasco, A., and Toro, M. A. (2014). A Short Critical History of the Application of Genomics to Animal Breeding. *Livestock Sci.* 166, 4–9. doi:10.1016/j.livsci.2014.03.015
- Botelho, M. E., Lopes, M. S., Mathur, P. K., Knol, E. F., Guimarães, S. E. F., Marques, D. B. D., et al. (2021). Applying an Association Weight Matrix in Weighted Genomic Prediction of Boar Taint Compounds. *J. Anim. Breed. Genet.* 138, 442–453. doi:10.1111/jbg.12528
- Calus, M. P. L., and Vandenplas, J. (2018). SNPrune: An Efficient Algorithm to Prune Large SNP Array and Sequence Datasets Based on High Linkage Disequilibrium. *Genet. Sel. Evol.* 50, 1–11. doi:10.1186/s12711-018-0404-z
- Cesarani, A., Biffani, S., Garcia, A., Lourenco, D., Bertolini, G., Neglia, G., et al. (2021). Genomic Investigation of Milk Production in Italian buffalo. *Ital. J. Anim. Sci.* 20, 539–547. doi:10.1080/1828051X.2021.1902404
- Cesarani, A., Pocrnic, I., Macciotta, N. P. P., Fragomeni, B. O., Misztal, I., and Lourenco, D. A. L. (2019). Bias in Heritability Estimates from Genomic Restricted Maximum Likelihood Methods under Different Genotyping Strategies. *J. Anim. Breed. Genet.* 136, 40–50. doi:10.1111/jbg.12367
- Chen, T., and Guestrin, C. (2016). “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA (ACM), 785–794. doi:10.1145/2939672.2939785
- Cherkassky, V., and Ma, Y. (2004). Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks* 17, 113–126. doi:10.1016/S0893-6080(03)00169-2
- Christensen, O. F., and Lund, M. S. (2010). Genomic Prediction when Some Animals Are Not Genotyped. *Genet. Sel. Evol.* 42, 2–8. doi:10.1186/1297-9686-42-2
- Evgeniou, T., and Pontil, M. (2005). “Support Vector Machines: Theory and Applications,” in *Machine*. Editor L. Wang (Berlin, Heidelberg: Springer Berlin Heidelberg). doi:10.1007/b95439
- Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4. Harlow, Essex, UK Ed: Longmans Green & Co., 464.

- Fragomeni, B. O., Lourenco, D. A. L., Legarra, A., VanRaden, P. M., and Misztal, I. (2019). Alternative SNP Weighting for Single-step Genomic Best Linear Unbiased Predictor Evaluation of Stature in US Holsteins in the Presence of Selected Sequence Variants. *J. Dairy Sci.* 102, 10012–10019. doi:10.3168/jds.2019-16262
- Fragomeni, B. O., Lourenco, D. A. L., Masuda, Y., Legarra, A., and Misztal, I. (2017). Incorporation of Causative Quantitative Trait Nucleotides in Single-step GBLUP. *Genet. Sel. Evol.* 49, 1–11. doi:10.1186/s12711-017-0335-0
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01
- Friedman, J. H. (2002). Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi:10.1016/S0167-9473(01)00065-2
- Frouin, A., Dandine-Roulland, C., Pierre-Jean, M., Deleuze, J.-F., Ambroise, C., and Le Floch, E. (2020). Exploring the Link between Additive Heritability and Prediction Accuracy from a Ridge Regression Perspective. *Front. Genet.* 11, 1–15. doi:10.3389/fgene.2020.581594
- Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics* 194, 573–596. doi:10.1534/genetics.113.151753
- Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51, 1440–1450. doi:10.2307/2533274
- Gualdrón Duarte, J. L., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., and Steibel, J. P. (2014). Rapid Screening for Phenotype-Genotype Associations by Linear Transformations of Genomic Evaluations. *BMC Bioinformatics* 15, 1–11. doi:10.1186/1471-2105-15-246
- Gualdrón Duarte, J. L., Gori, A.-S., Hubin, X., Lourenco, D., Charlier, C., Misztal, I., et al. (2020). Performances of Adaptive MultiBLUP, Bayesian Regressions, and Weighted-GBLUP Approaches for Genomic Predictions in Belgian Blue Beef Cattle. *BMC Genomics* 21, 1–18. doi:10.1186/s12864-020-06921-3
- Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics* 194, 597–607. doi:10.1534/genetics.113.152207
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer.
- Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31, 423–447. doi:10.2307/2529430
- Karaman, E., Cheng, H., Firat, M. Z., Garrick, D. J., and Fernando, R. L. (2016). An Upper Bound for Accuracy of Prediction Using GBLUP. *PLoS One* 11, e0161054–18. doi:10.1371/journal.pone.0161054
- Legarra, A., Aguilar, I., and Misztal, I. (2009). A Relationship Matrix Including Full Pedigree and Genomic Information. *J. Dairy Sci.* 92, 4656–4663. doi:10.3168/jds.2009-2061
- Legarra, A., and Reverter, A. (2017). Can We Frame and Understand Cross-Validation Results in Animal Breeding? *Proc. Assoc. Advmt. Anim. Breed. Genet.* 22, 73–80.
- Legarra, A., and Reverter, A. (2018). Semi-parametric Estimates of Population Accuracy and Bias of Predictions of Breeding Values and Future Phenotypes Using the LR Method. *Genet. Sel. Evol.* 50, 1–18. doi:10.1186/s12711-018-0426-6
- Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. *Front. Genet.* 9, 1–20. doi:10.3389/fgene.2018.00237
- Liu, Z. (2010). Interbull Validation Test for Genomic Evaluations. *Interbull Bull.* 17.
- Macedo, F. L., Christensen, O. F., Astruc, J.-M., Aguilar, I., Masuda, Y., and Legarra, A. (2020). Bias and Accuracy of Dairy Sheep Evaluations Using BLUP and SSGBLUP with Metafounders and Unknown Parent Groups. *Genet. Sel. Evol.* 52, 1–10. doi:10.1186/s12711-020-00567-1
- Mancin, E., Lourenco, D., Bermann, M., Mantovani, R., and Misztal, I. (2021a). Accounting for Population Structure and Phenotypes from Relatives in Association Mapping for Farm Animals: A Simulation Study. *Front. Genet.* 12. doi:10.3389/fgene.2021.642065
- Mancin, E., Tuliozi, B., Pegolo, S., Sartori, C., and Mantovani, R. (2022). Genome Wide Association Study of Beef Traits in Local Alpine Breed Reveals the Diversity of the Pathways Involved and the Role of Time Stratification. *Front. Genet.* 12, 1–22. doi:10.3389/fgene.2021.746665
- Mancin, E., Tuliozi, B., Sartori, C., Guzzo, N., and Mantovani, R. (2021b). Genomic Prediction in Local Breeds: The Rendena Cattle as a Case Study. *Animals* 11, 1815–1819. doi:10.3390/ani11061815
- Mastrangelo, S., Saura, M., Tolone, M., Salces-Ortiz, J., Di Gerlando, R., Bertolini, F., et al. (2014). The Genome-wide Structure of Two Economically Important Indigenous Sicilian Cattle Breeds. *J. Anim. Sci.* 92, 4833–4842. doi:10.2527/jas.2014-7898
- Masuda, Y., VanRaden, P. M., Misztal, I., and Lawlor, T. J. (2018). Differing Genetic Trend Estimates from Traditional and Genomic Evaluations of Genotyped Animals as Evidence of Preselection Bias in US Holsteins. *J. Dairy Sci.* 101, 5194–5206. doi:10.3168/jds.2017-13310
- Mehrban, H., Naserkheil, M., Lee, D. H., Cho, C., Choi, T., Park, M., et al. (2021). Genomic Prediction Using Alternative Strategies of Weighted Single-step Genomic BLUP for Yearling Weight and Carcass Traits in Hanwoo Beef Cattle. *Genes* 12, 266. doi:10.3390/genes12020266
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2020). *Misc Functions of the Department of Statistics*. Probability Theory Group, e1071. (TU Wien), 1–63.
- Misztal, I., Aggrey, S. E., and Muir, W. M. (2013). Experiences with a Single-step Genome Evaluation. *Poult. Sci.* 92, 2530–2534. doi:10.3382/ps.2012-02739
- Mitchell, R., and Frank, E. (2017). Accelerating the XGBoost Algorithm Using GPU Computing. *PeerJ Comput. Sci.* 3, e127. doi:10.7717/peerj-cs.127
- Natekin, A., and Knoll, A. (2013). Gradient Boosting Machines, a Tutorial. *Front. Neurobot.* 7, 1–21. doi:10.3389/fnbot.2013.00021
- Piles, M., Bergsma, R., Gianola, D., Gilbert, H., and Tusell, L. (2021). Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. *Front. Genet.* 12. doi:10.3389/fgene.2021.611506
- Pocrnic, I., Lourenco, D. A. L., Masuda, Y., and Misztal, I. (2019). Accuracy of Genomic BLUP when Considering a Genomic Relationship Matrix Based on the Number of the Largest Eigenvalues: A Simulation Study. *Genet. Sel. Evol.* 51, 1–10. doi:10.1186/s12711-019-0516-0
- Ren, D., An, L., Li, B., Qiao, L., and Liu, W. (2021). Efficient Weighting Methods for Genomic Best Linear-Unbiased Prediction (BLUP) Adapted to the Genetic Architectures of Quantitative Traits. *Heredity* 126, 320–334. doi:10.1038/s41437-020-00372-y
- Ročková, V., and George, E. I. (2018). The Spike-And-Slab LASSO. *J. Am. Stat. Assoc.* 113, 431–444. doi:10.1080/01621459.2016.1260469
- Sanz, H., Valim, C., Vegas, E., Oller, J. M., and Reverter, F. (2018). SVM-RFE: Selection and Visualization of the Most Relevant Features through Non-linear Kernels. *BMC Bioinformatics* 19, 432. doi:10.1186/s12859-018-2451-4
- Sargolzaei, M., and Schenkel, F. S. (2009). QMSim: a Large-Scale Genome Simulator for Livestock. *Bioinformatics* 25, 680–681. doi:10.1093/bioinformatics/btp045
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
- VanRaden, P. M. (2008a). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980
- VanRaden, P. M. (2008b). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980
- Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. New York, NY: Springer.
- Vitezica, Z. G., Aguilar, I., Misztal, I., and Legarra, A. (2011). Bias in Genomic Predictions for Populations under Selection. *Genet. Res.* 93, 357–366. doi:10.1017/S001667231100022X
- Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., Vitezica, Z., et al. (2014). Genome-wide Association Mapping Including Phenotypes from Relatives without Genotypes in a Single-step (ssGWAS) for 6-week Body

- Weight in Broiler Chickens. *Front. Genet.* 5, 134. doi:10.3389/fgene.2014.00134
- Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., and Misztal, I. (2016). Weighting Strategies for Single-step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Front. Genet.* 7, 151. doi:10.3389/fgene.2016.00151
- Zhu, S., Guo, T., Yuan, C., Liu, J., Li, J., Han, M., et al. (2021). Evaluation of Bayesian Alphabet and GBLUP Based on Different Marker Density for Genomic Prediction in Alpine Merino Sheep. *G3 Genes, Genomes, Genet.* 11. doi:10.1093/g3journal/jkab206

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Mancin, Mota, Tuliozi, Verdiglione, Mantovani and Sartori. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Applications of Omics Technology for Livestock Selection and Improvement

Dibyendu Chakraborty¹, Neelesh Sharma^{2*}, Savleen Kour², Simrinder Singh Sodhi³, Mukesh Kumar Gupta⁴, Sung Jin Lee⁵ and Young Ok Son^{6*}

¹Division of Animal Genetics and Breeding, Faculty of Veterinary Sciences and Animal Husbandry, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Ranbir Singh Pura, India, ²Division of Veterinary Medicine, Faculty of Veterinary Sciences and Animal Husbandry, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Ranbir Singh Pura, India, ³Department of Animal Biotechnology, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, India, ⁴Department of Biotechnology and Medical Engineering, National Institute of Technology, Rourkela, India, ⁵Department of Animal Biotechnology, College of Animal Life Sciences, Kangwon National University, Chuncheon-si, South Korea, ⁶Department of Animal Biotechnology, Faculty of Biotechnology, College of Applied Life Sciences and Interdisciplinary Graduate Program in Advanced Convergence Technology and Science, Jeju National University, Jeju, South Korea

OPEN ACCESS

Edited by:

Sunday O. Peters,
Berry College, United States

Reviewed by:

Suxu Tan,
Michigan State University,
United States
Syarul Nataqain Baharum,
National University of Malaysia,
Malaysia

*Correspondence:

Neelesh Sharma
dneelesh_sharma@yahoo.co.in
Young Ok Son
sounagi@jejunu.ac.kr

Specialty section:

This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 11 September 2021

Accepted: 16 May 2022

Published: 02 June 2022

Citation:

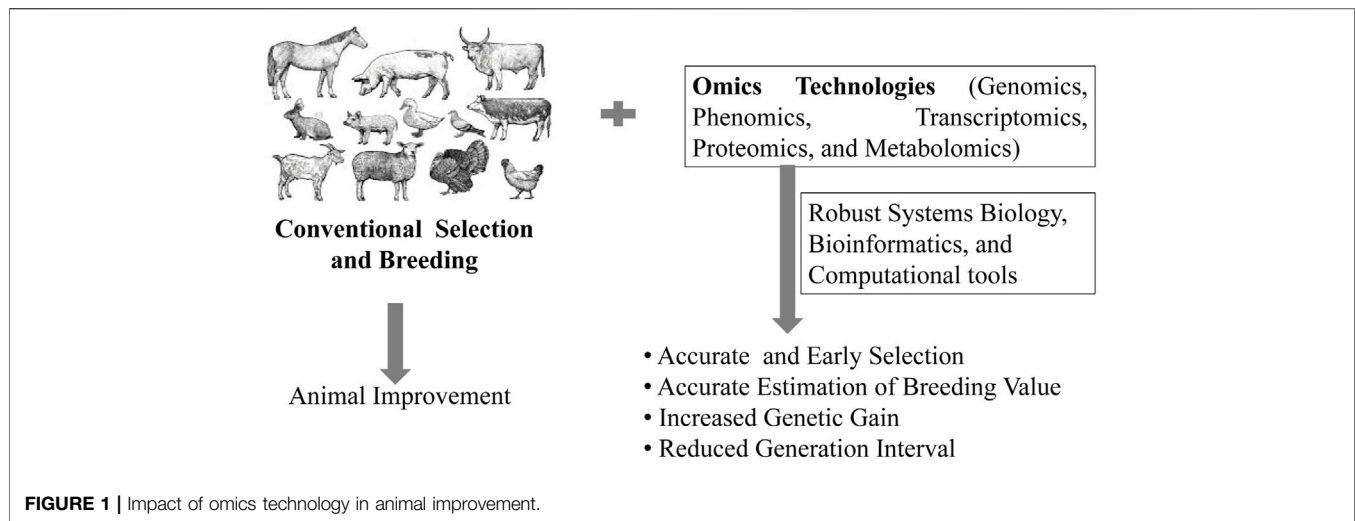
Chakraborty D, Sharma N, Kour S,
Sodhi SS, Gupta MK, Lee SJ and
Son YO (2022) Applications of Omics
Technology for Livestock Selection
and Improvement.
Front. Genet. 13:774113.
doi: 10.3389/fgene.2022.774113

Conventional animal selection and breeding methods were based on the phenotypic performance of the animals. These methods have limitations, particularly for sex-limited traits and traits expressed later in the life cycle (e.g., carcass traits). Consequently, the genetic gain has been slow with high generation intervals. With the advent of high-throughput omics techniques and the availability of multi-omics technologies and sophisticated analytic packages, several promising tools and methods have been developed to estimate the actual genetic potential of the animals. It has now become possible to collect and access large and complex datasets comprising different genomics, transcriptomics, proteomics, metabolomics, and phenomics data as well as animal-level data (such as longevity, behavior, adaptation, etc.), which provides new opportunities to better understand the mechanisms regulating animals' actual performance. The cost of omics technology and expertise of several fields like biology, bioinformatics, statistics, and computational biology make these technology impediments to its use in some cases. The population size and accurate phenotypic data recordings are other significant constraints for appropriate selection and breeding strategies. Nevertheless, omics technologies can estimate more accurate breeding values (BVs) and increase the genetic gain by assisting the selection of genetically superior, disease-free animals at an early stage of life for enhancing animal productivity and profitability. This manuscript provides an overview of various omics technologies and their limitations for animal genetic selection and breeding decisions.

Keywords: omics, selection, animal improvement, phenomics, data analysis

INTRODUCTION

Genetic selection and breeding are crucial tools for livestock improvement. They have resulted in genetically superior and disease-free animals with improved production and efficiency in various livestock species (Rexroad et al., 2019; Erasmus and van Marle-Köster 2021). In earlier days, the genetic selection of animals for breeding was primarily based on their phenotypic characteristics, such as production traits and breeding value (BV) estimation. Later, other economic traits, including



reproduction and longevity traits, animal health, stress tolerance, disease resistance, animal welfare traits, etc., also became vital components of genetic improvement programs (Brito et al., 2020; Brito et al., 2021). Selective breeding of genetically superior animals ensured rapid genetic progress of production efficiency traits to the next generation. Many breeding techniques have thus, evolved to accrue the desired trait in genetically selected animals to meet the market demand for production and animal welfare (Plieschke et al., 2016).

Several selection indices have been developed for the genetic selection of animals for breeding. However, no single trait is ideal for these selection indices in all populations (Cole et al., 2021). Further, while animals for selective breeding can be identified based on phenotypic recordings, traits that are sex-limited, expressed at a later stage of life, difficult to measure, or have low heritability pose difficulties (Calus et al., 2013). The use of complex statistical models, advanced analytic tools, and new molecular methods may divulge newer traits and help identify animals for efficient genetic selection and breeding with greater accuracy (Stock et al., 2020).

The past three decades have seen tremendous advancements in molecular genetics that have provided a better genetic understanding of quantitative economic traits (Dekkers and Hospital 2002). A number of genes and gene combinations have been found to directly correlate with animal performance and production efficiency (Rexroad et al., 2019; Ruan et al., 2021). Many quantitative trait loci (QTL)—gene loci responsible for trait diversity—have been identified for various production and reproductive traits and used for selection and breeding decisions (Zhang et al., 2021; Al-Sharif et al., 2022). Several genetic markers have also been discovered for use in marker-assisted selection (MAS) of breeding stock (Ma et al., 2021; Raza et al., 2021). More recently, with advancements in high throughput *omics* technologies, genome selection is becoming widely accepted for the selection of animals for breeding (Tan et al., 2017; Yang et al., 2020). The application of *omics* tools in livestock improvement may provide a more accurate technology for animal selection and breeding and therefore has become a hot

spot of research (Pedrosa et al., 2021; Ruan et al., 2021). This manuscript provides an overview of various *omics* tools and technologies for their application in livestock selection and improvement programs.

OMICS TECHNOLOGY

Omics technologies such as genomics, metagenomics, metabolomics, proteomics, transcriptomics, epigenomics, translomics, etc., can allow rapid and effective detection of subtle phenotypic changes, dietary responses, and innate phenotypic propensities in animals (Mu et al., 2022; Wang et al., 2022). The utilization of *omics* tools in animal selection and breeding programs is thus, expected to provide an accurate estimation of BV for early selection, reduce generation interval and increase the rate of genetic gain (Figure 1). The word '*omics*' originates from the suffix '-ome', derived from a Greek word that means "whole", "all" or "complete". The suffix "-omics" is frequently used to refer to a field of study in life sciences that emphasizes large-scale high throughput data/information to understand life summed up in "omes" (Yadav, 2007). Several *omics* tools have been developed in the last two decades to collect and analyze high-throughput data on proteins (proteomics), mRNA transcripts (transcriptomics), gene sequences (genomics), microbial diversity (metagenomics), epigenetic regulation of gene expression (epigenomics), metabolic profile (metabolomics), lipid profile (lipidomics), etc., of a particular cell, tissue, organ or whole organism at a specific time point. The time (temporal) and space (spatial) level information from *omics* data can be integrated through robust bioinformatics and computational tools to the systems biology level (Odom et al., 2021; Velten et al., 2022). Network modeling of *omics* data can be used to study the mechanism, relationship, interaction, and function of cells, tissues, organs, and the whole organism at a molecular level in an unbiased manner (Aardema and MacGregor, 2003). More recently, *multi-omics* has emerged as high-dimensional biology (HBD) for simultaneous study of genetic variations in biological

systems at the genes, transcripts, proteins, and metabolites level (Romero et al., 2006; Krassowski et al., 2020).

In the last two decades, a few landmark technological revolutions took place in *omics* technologies that revolutionized their application in genetic selection for animal breeding. In 2007, the first “high density” panel of bovine genetic markers was released commercially with a set of 54,001 single nucleotide polymorphisms (SNPs). Using such high-density SNP chips, genome-wide association studies (GWAS) demonstrated the link between SNPs and QTLs, such as coat color and presence or absence of horns (Matukumalli et al., 2009). In other studies, high-density SNP chips were shown to be useful in the genetic characterization of pig breeds for preserving their genomic variability (Muñoz et al., 2019). Whole-genome sequencing (WGS) by massively parallel sequencing was yet another breakthrough for detecting molecular signatures for the selection and breeding of animals (Elsik et al., 2009; Bovo et al., 2020). The Bovine Genome Sequencing and Assembly project (Elsik et al., 2009; The Bovine Genome Sequencing and Analysis Consortium et al., 2009; The Bovine HapMap Consortium, 2009) provided a landscape of genome sequence that subsequently led to a paradigm shift in QTL- and candidate gene-based approaches for genetic selection.

Molecular databases of NCBI (United States), EMBL (Europe), and DDBJ (Japan) provide vast information on nucleotide and protein sequences. These databases have been utilized in *omics* technology for understanding the genomic variability and molecular and physiological basis of economic traits (Wu et al., 2018; Ng et al., 2021). Unfortunately, however, very scant information is available on the precise regulatory networks through which these genes and proteins determine the phenotypic expression of economic traits. Further, a significant unexplained source of variation among phenotypes of various economic traits remains a matter of concern in livestock. Newer machine learning (ML) tools have been developed recently that can be exploited to analyze high throughput *omics* data, available in databases, for a greater understanding of gene regulatory networks (GRNs) and identification of functional genes by a systems biology approach (Guttula et al., 2020; Guttula et al., 2021; Ng et al., 2021).

Omics technologies can help identify functional SNPs and their prioritizing to increase the accuracy of genetic selection (Chang et al., 2019). They can also be used for selecting animals resistant to production diseases such as mastitis and thereby enhance their productivity (Russell et al., 2012; Bhattarai et al., 2017; Jaiswal et al., 2021). Further, population-level *omics* (e.g., population genomics) hold tremendous potential for classifying individuals based on allelic diversity and identifying genetically-related individuals (Lippert et al., 2017). Such strategies can help calculate homozygosity and inbreeding coefficients (Ghoreishifar et al., 2020; Sumreddee et al., 2021) for designing appropriate breeding programs to maintain genetic diversity and avoid inbreeding depression (Alemu et al., 2021; Bu et al., 2021). However, while many WGS databases and consortiums have been formed in humans (Zhang et al., 2018) (GenomeAsia 100K Consortium, 2019), no

high-resolution database of population-level genetic variants is available for animals.

GENOMICS

The genome is defined as the complete set of genetic material present in an organism. The term “genomics” was coined in 1986 by scientists who were naming a new journal (Kuska, 1998), and thus, the era of *omics* began. The major developments in genomics are the discovery of the genes and genetic codes, polymerase chain reaction (PCR), Genome sequencing by Sanger sequencing or Next Generation Sequencing (NGS), and genome editing tools such as Transcription activator-like effector nuclease (TALEN), Zinc-finger nuclease (ZNF), an Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (CRISPR/Cas9) technologies. The field of genomics started gaining popularity after the invention of PCR in the year 1985. Thereafter, MAS- and candidate gene-based approaches for selecting genetically superior animals became very popular and were found to be better than conventional phenotype-based selection and breeding (Williams 2005). Gene map construction was also used for genome sequencing. The gene map construction was initially based on the segregation of enzyme markers across panels of hybrid cell lines. However, with advancements in recombinant DNA (rDNA) technology, denser physical and genetic maps formed an important framework for genome sequencing (Riggs and Gill, 2009). During the early 2000s, most livestock genome sequencing was based on linkage maps using single markers and quantifying one or a few genes by real-time quantitative PCR (qPCR). Elisk and associates, in 2009, published the first bovine genome assembly (Elsik et al., 2009). Since then, rapid progress has been made in developing and using several whole genome-*omics* tools that have accelerated cattle genetics research (Reverter et al., 2013; Snelling et al., 2013).

The concept of “Genetical Genomics”, which integrates structural and functional genomic data, has evolved with the development of microarray technology for gene expression analysis, which divulged marker genotypes across whole genome. The field of Genetical Genomics has expanded with the availability of high throughput tools for genomic analysis such as high-density (HD) genotyping-chips (Illumina, San Diego, CA), WGS, genotyping by sequencing, and RNA-sequencing (RNAseq) to measure the gene expression in the entire transcriptome (Wickramasinghe et al., 2014). Several SNP chips of 60 K for pig and chicken, 50 K for sheep, and 77 K for cattle have also been developed (Suravajhala et al., 2016). The GWAS studies have become very popular among different livestock species focusing on production and health traits (Shamra et al., 2015). For example, GWAS on female reproduction traits in tropically adapted beef cattle (Hawken et al., 2012), feed efficiency traits in pigs (Do et al., 2013 and 2014), body weight in broilers (Wang et al., 2014), and obesity and metabolic diseases using the pig as a model (Kogelman et al., 2014) have been conducted. The genomic selection is particularly advantageous as it can be used for selecting animals for breeding

TABLE 1 | Impact of genomic selection^a.

Animals	Added genetic Gain	References
Dairy cattle	60–120%	Pryce and Daetwyler, (2011)
Beef cattle	15–44%	Pimentel and König, (2012)
Dairy goat	26.2%	Shumbusho et al. (2013)
Dairy sheep	51.7%	Shumbusho et al. (2013)
Meat sheep	17.9%	Shumbusho et al. (2013)
Pig	23–91%	Lillehammer et al. (2011)
Layers	60%	Sitzenstock et al. (2013)
Broilers	20%	Dekkers et al. (2009)
Dairy Bulls	30–71%	Doublet et al. (2019)

^aSource: Modified from Ibisham et al., 2017.

at an early stage of life without having reference to their own breeding or production records. Studies have shown that genomic selection improved the genetic gain as much as 60–120% in dairy cattle by decreasing genetic interval by 2 years, although the extent of added genetic gain was lower in other livestock species (**Table 1**). The term “systems genetics” or “systems genomics” was also proposed by Kadarmideen (2014). This branch has a wide range of approaches ranging from relating the individual's *omics* levels data to their functional annotation and analysis of signaling pathways by integrating different multi-omic levels data to phenotypes.

TRANSCRIPTOMICS

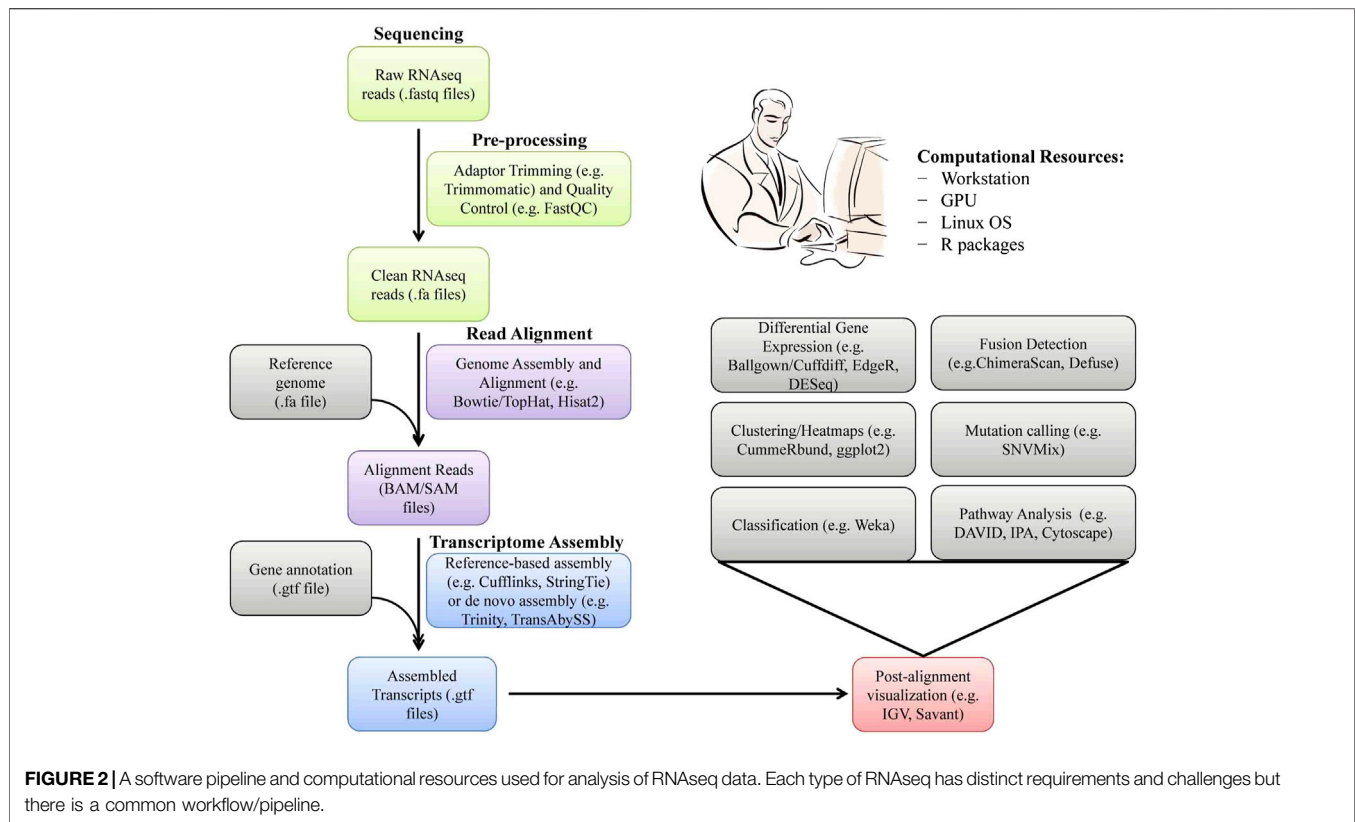
Transcriptomic methods can be used to compare a biological response to different conditions or treatments or to assess physiological responses to external stimuli (Brannan et al., 2014). Whole transcriptome sequencing is the most widely used method for studying RNA functions, exploring and analyzing the gene structure and function, and revealing intrinsic links between gene expression and life phenomena (Shi and Zhang, 2019). To date, extensive research has been carried out in different livestock species using high-throughput RNAseq technology that has replaced the earlier used Maxam and Gilbert chemical degradation sequencing method. The NGS technologies generate sequence data by producing millions of short DNA fragments in parallel. The template is broken into many smaller fragments by sheering, which are then ligated to adapters to create cDNA libraries by the bridge (e.g., Illumina sequencing) or emulsion (e.g., pyrosequencing) PCR. The clones of cDNA fragments of each library are then sequenced to obtain short *reads*; the length and number of the reads vary with the specific technology but generally range between 30 and 300 bases, which is shorter than those obtained from Sanger sequencing (Ghaffari et al., 2013). NGS has led to the characterization and quantification of many omics, including genomics (DNA sequencing), transcriptomics (RNA and cDNA sequencing), and epigenomics (ChIP-seq and DNA methylation analysis). More recently, third-generation sequencing methods involving single-molecule real-time (SMRT) sequencing have emerged (Sahoo et al., 2021a; Sahoo et al., 2021b). These newer SMRT sequencing methods do not require PCR amplification of

templates and hence are devoid of PCR biases. Moreover, the SMRT sequencing approaches generally produce longer reads for better genome assembly and identification of indels (Athanasopoulou et al., 2021). However, SMRT methods such as NanoporeTM and PacBioTM sequencing also offer versatility in terms of rapid time and the transportability of the equipment. Newer techniques such as tunneling currents DNA sequencing, sequencing with mass spectrometry, microscopy-based sequencing, etc., are under development.

The RNAseq has made a revolutionary impact on transcriptome analysis (Mortazavi et al., 2008). RNAseq has major advantages such as a large dynamic range and sensitivity, precise, unbiased quantification of transcripts, and comprehensive coverage of all expressed sequences in a given tissue sample. The direct RNAseq is vital for functional studies to capture the dynamic RNA population under different environmental conditions (Athanasopoulou et al., 2021). It has revolutionized gene annotation, which was hitherto very difficult with genome sequencing. The RNAseq also finds application in analyzing molecular features such as alternate isoforms, splice variants, fusion transcripts, RNA editing, etc. (Li and Wang 2021). Combination of genome sequencing with RNAseq can be utilized to interpret mutations on regulatory regions of genes, which do not produce an obvious effect on the protein sequence (Cohen et al., 2020).

Today, very accurate and efficient sequencing platforms are available, which can distinguish closely related transcripts from each other (Marguerat and Braga-Neto, 2015). Therefore, RNAseq has become very popular for the identification and quantification of splice variants, fused transcripts, and mutants. In RNAseq technology, messenger RNAs are first randomly fragmented into small pieces by sheering and converted to a library of complementary DNA (cDNA) fragments. These cDNA fragments are then amplified and sequenced in parallel and mapped to a given region of the target genome. PCR-free cDNA sequencing and direct RNAseq without first-strand cDNA synthesis have also become possible with SMRT technology such as NanoporeTM sequencing. In expression quantification, a count, which is determined by the number of reads mapping to each gene (FPKM or TPM; fragments per kilobase of transcripts per million mapped reads or transcripts per million), is a discrete measure of the corresponding gene expression level (Ghaffari et al., 2013, Li et al., 2012) (**Figure 2**). The differentially expressed genes (DEGs) between two samples can be obtained by transcript compilation with gene annotation file followed by gene identification and differential expression analysis based on FPKM or TPM values (Alessandri et al., 2019). Functional analysis of DEGs by bioinformatics tools revealed that the immune and inflammatory responses were the most impacted pathways between purebred and crossbred cattle populations (Moridi et al., 2019). Such type of RNAseq-based transcriptomic studies on animals of high- and low-genetic merit may be helpful for the selection and breeding of elite animals in the future to enhance health, productivity, and profitability.

Canovas et al. (2014) integrated the RNAseq data with GWAS and bovine transcriptional factors in multiple tissues from pre-



and post-pubertal cattle and constructed co-expression GRNs, which revealed genes and their complex interactions during puberty in cattle. Therefore, early selection of individuals based on *multi-omics* data from early sexual maturity may help increase the genetic gain by reducing generation intervals. In another study, the resistance or susceptibility of Creole goats to gastrointestinal nematodes was studied by RNAseq (Aboshady et al., 2021). Aboshady et al. (2021) reported that the T-cell receptor signaling pathway was one of the top significant pathways that distinguish the resistant from the susceptible genotype, with 78% of the genes involved in this pathway showing genomic variants in Creole goats. This shows another important example of applying *omics* for selecting disease-resistant animals.

PROTEOMICS

Wilkins and Williams first described proteomics in the mid-1990s (Speicher, 2004). Proteomics allows analysis of all proteins, including their isoforms, in a particular cell, tissue, or organ at a specific time in a single experiment. Advance proteomic tools can also provide information on various protein isoforms, their quantification, and protein-protein interaction. However, the application of proteomics in livestock research has been limited in the past due to its high cost and lack of optimized protocols for various cell types in different species (Baykalir et al., 2018). Nevertheless, with advancements in new analytical methods and computational tools for the analysis of

proteomic data, reports on proteomics studies in animal science are increasing for understanding the animal health status and production and reproduction efficiency (Zhao et al., 2021; Kaya et al., 2022; Ye et al., 2022).

Proteomics techniques range from one-dimensional (1D) gel electrophoresis, two-dimensional (2D) gel electrophoresis, Chromatography (liquid and gas) methods to sophisticated mass spectrometry (MALDI-MS, ESI-MS, LC-MS/MS, MALDI-TOF MS, etc.), which measures mass-to-charge (m/z ratio) of ionized peptides to identify proteins (Gupta et al., 2009a). In a typical MS experiment, the proteins are isolated from target cells/tissue/organ/biological fluid, separated by 1D or 2D gel electrophoresis or liquid chromatography, and digested by a sequence-specific protease such as trypsin (Figure 3). The trypsin digests are then purified by affinity chromatography or biochemical fractionation and ionized by electrospray ion (ESI), matrix-assisted laser desorption ionization (MALDI), or surface-enhanced laser desorption ionization (SELDI) before being pushed into the mass spectrometer to measure the m/z ratio. The m/z ratio can be measured in quadrupole (Q), ion trap (e.g., quadrupole ion trap or QIT and linear ion trap or LIT), time-of-flight (TOF), quadrupole mass filters (QMF), ion cyclotron resonance (ICR), high-resolution orbitraps, or a hybrid of mass spectrometers. The MS spectra are then matched with protein databases to identify proteins using a variety of algorithms that usually come in-built with the MS machines (e.g., SpectraMill™) (Gupta et al., 2009b). Several methods have also been developed for relative or absolute quantification (AQUA) of proteins and identification of post-

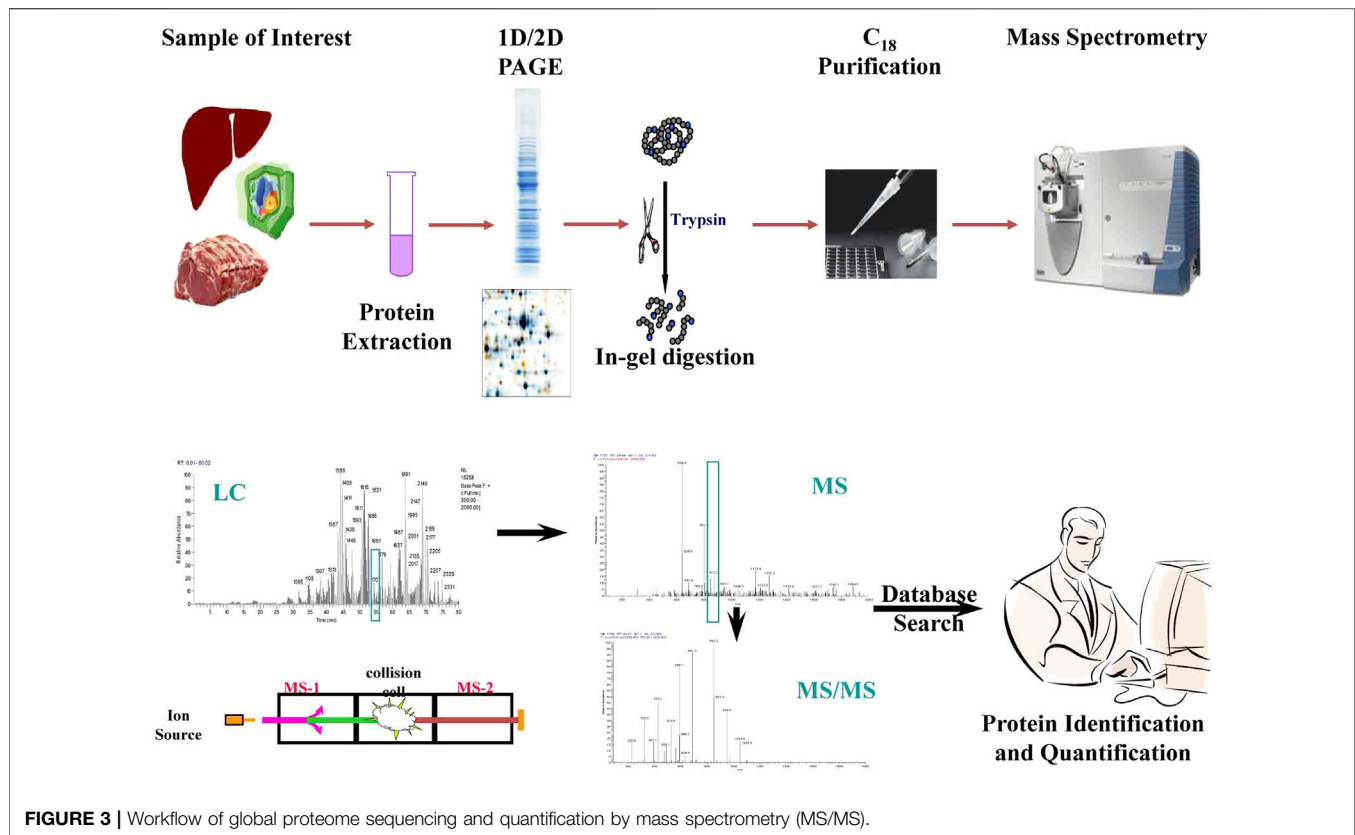


FIGURE 3 | Workflow of global proteome sequencing and quantification by mass spectrometry (MS/MS).

translational modifications by modified MS such as selected reaction monitoring (SRM), isotope labeling of amino acids in cell culture (SILAC), isotope-coded affinity tags (ICAT), isobaric mass tagging (iTRAQ), etc.

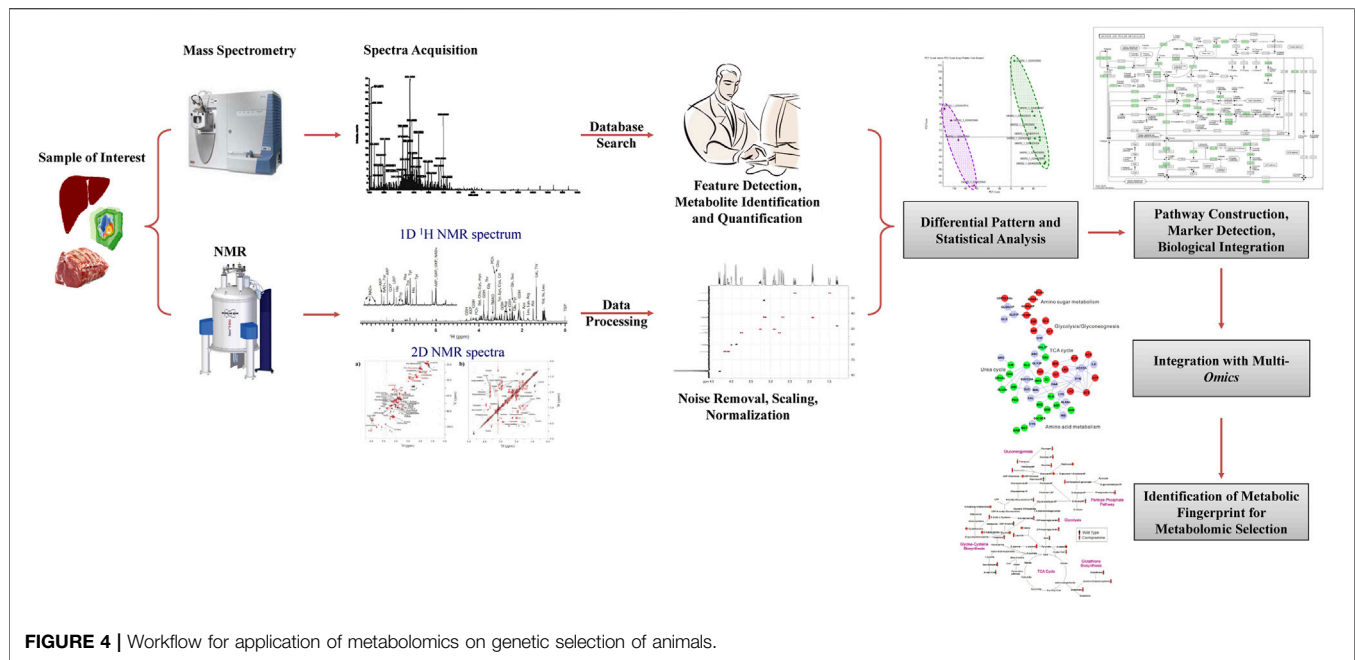
The number of proteomics studies associated with reproductive problems has been increased dramatically in the last decade (Peddinti et al., 2011). The research applications in proteomics range from early growth and development to postmortem events important for meat quality (Yarmush and Jayaraman, 2002; Bendixen, 2005). One of the major areas of interest in proteomics is finding out robust protein biomarkers that could be useful in disease surveillance, monitoring the health and wellbeing of animals, elucidating disease mechanisms, and assessing pharmacologic response to therapeutics (Oskoueian et al., 2016) (Muhanguzi et al., 2022). Proteomics can also be applied for different animal products post-harvest like meat, milk, cheese, etc., to identify genetic variants with desirable traits for selection and breeding (Almeida et al., 2015).

Proteomics has also enabled the identification of candidate protein markers of fertility for molecular breeding. The LC-MS/MS analysis of pig sperm revealed eight fertility-related proteins over-represented in Tibetan pigs having heritable adaptation to hypoxic environments (Zhao et al., 2021). In another study, analysis of seminal plasma proteins by LC-MS/MS found a consistent correlation of 1,343 proteins with fertility (Willfors et al., 2021). Thus, identifying fertility markers by proteomics can help identify fertile bulls to reduce non-return rates (NRR) and increase productivity. Proteomic tools can also be harnessed to

identify superior genetic variants to dietary response and muscle growth for selective breeding. By a newly described transcriptome-assisted label-free shotgun proteomics method, Mullins et al. (2021) identified 24 differentially abundant proteins in liver tissues from cattle that were fed *ad libitum* or restricted diet. Identifying protein markers by proteomics could help the selection of genetic variants for compensatory growth upon undernutrition, which may accelerate genetic gain and increase profitability (Mullins et al., 2021).

METABOLOMICS

An emerging area in the application of *omics* tools is the interrogation of the metabolome. Metabolomics is a comprehensive, qualitative and quantitative study of all the small molecules in an organism (Kalaiselvi et al., 2019; Lipka et al., 2022). Metabolomic tools are being increasingly used to generate an unbiased global profile of metabolites in samples (i.e., untargeted analysis) or to quantify with high sensitivity a small panel of metabolites (targeted analysis) (Riggs et al., 2017; Evans et al., 2020). In dairy cattle, many potential biomarkers of milk yield and quality have been detected by studying the metabolome of different body fluids (Sun et al., 2015). One advantage of profiling metabolites is exploring the impact of metabolism on systemic health by monitoring the production and further metabolism of compounds present in the diet, digesta, and plasma (Seidel et al., 2014). It can also be used to evaluate feed



conversion efficiency, metabolic response of animals to environmental conditions, and estimation of production efficiency and carcass quality traits (Jorge-Smeding et al., 2021; Martin et al., 2021; Artegoitia et al., 2022). Studies have also shown that the heritability of water-soluble compounds such as free amino acids, nucleotides, and sugars in beef was less than 0.30 and varied with animal age (Sakuma et al., 2016). However, these water-soluble compounds were negatively correlated with carcass weight and beef marbling standard at the genetic level. Thus, metabolomics may help identify animals with high carcass quality for breeding. In another study, metabolic profiling of muscle by GC-MS and LC-MS could distinguish grass- and grain-fed cattle with 100% predictive accuracy (Carrillo et al., 2016). These results suggest that metabolic signatures could be a good indicator of animals' feeding habits and carcass quality and, therefore, could be utilized to select animals with desired traits (Figure 4).

Metabolomic selection is an emerging breeding technology based on nuclear magnetic resonance (NMR) or LC-MS metabolomics (Evans et al., 2020; Lipa et al., 2022). The NMR spectra of biological samples can be analyzed for chemical shifts, peak intensities, and coupling patterns to identify and quantify various metabolites and generate NMR fingerprints of the sample (Figure 4). Metabolomic studies on muscle and fat from cattle, pigs, and poultry have shown tissue- and species-specific differences in metabolites with specific compounds detected in each species (Ueda et al., 2018). The GC-MS could also distinguish between cattle breeds (Ueda et al., 2018). Thus, comparing NMR spectra from different animals such as those from low- and high-performing animals may help identify NMR fingerprints in high-performing animals. Such NMR fingerprints can then be used for the genetic selection of animals for breeding purposes (Figure 4). Metabolomic analysis of muscle from Nellore cattle having high- or low-growth traits

revealed that high growth animals had a distinct metabolic profile with a higher concentration of specific metabolites affecting protein and fatty acid metabolism (Cansolo et al., 2020) that can be harnessed for selection of animals for growth.

High-resolution MS (HRMS) can detect metabolites at nano- to the pico-molar concentration of metabolome and, therefore, can provide a better landscape of metabolites than NMR (Goldansaz et al., 2017). The MS is usually combined with separation techniques such as capillary, liquid, or gas chromatography, depending on the polarity and lipophilicity of the metabolites of interest. The separated molecules are ionized by ESI, electron ionization (EI), chemical ionization (CI), or atmospheric pressure CI (APCI) and evaluated for m/z ratio in the mass spectrometer based on TOF, Fourier transformation ion cyclotron resonance (FT-ICR) or orbitrap to obtain structural information for identifying the metabolites. Several databases such as METLIN, Human Metabolome Database (HMDB), and MassBank are available that can be used to match the MS spectra of metabolites for their identification. A number of statistical and bioinformatic tools can then be applied to discover molecular pathways involved in the generation of critical metabolites. Software such as MetaboAnalyst and Kyoto Encyclopedia of Genes and Genomes (KEGG) can be used for multivariate analysis and visualization of metabolic pathways. The correlation analysis between animal performance parameters and metabolic profiles may help identify key metabolic markers of animals' performance for genetic selection. Metabolomics has also been used in GWAS for metabolite-featured phenotyping and animal breeding (Fontanesi, 2016).

The metabolomic tools can also be combined with molecular breeding tools such as WGS and high-density SNP chips to increase the accuracy of genetic selection and livestock breeding (Wang and Kadarmideen 2020) (Ehret et al., 2015).

Given that genomic prediction to predict breeding values based on phenotypic, pedigree, and genomic data is insufficient to describe the genetic potential of animals, incorporating the whole-metabolomic data in the genomic prediction equation may play a crucial role in increasing the genetic gain by increasing the accuracy of selection. The latter is further substantiated by the fact that metabolites represent cells' ultimate physiological response and thereby represent a link between genotype and phenotype (Wang and Kadarmideen 2020). GWAS-based studies, using SNP chips and LC-MS metabolomics, the identified mechanism underlying the genetic variation in pigs for feed efficiency (Banerjee et al., 2020; Wang and Kadarmideen 2020). Integrating high-density SNP data and metabolite information with predictive value was also found to help improve the accuracy of genetic selection in cattle (Ehret et al., 2015). The power of metabolomics is that it non-invasively detects subtle phenotypic changes, innate phenotypic propensities, and dietary responses in livestock research, breeding, and assessment through new varieties of bio-samples such as semen, amniotic fluid, saliva, and urine.

Wu et al. (2021) found that small metabolite profiling of pig feces by LC-MS metabolomics correlated with their feed efficiency and can be used as a reference for selecting animals with high feed conversion efficiency and responsiveness to new feed additives (Wu et al., 2021). Given that fecal metabolome are reflections of intestinal microbiota, cellular metabolism and digestion/absorption of nutrients in the gut (Zierer et al., 2018; Malheiros et al., 2021), the metabolites present in the feces could indicate their feed conversion efficiency (Wu et al., 2021). Fecal metabolome was also shown to change as a function of stress in beef cattle (Valerio et al., 2020). Thus, metabolic profiling of fecal matter may be used to identify animals with "metabolic fingerprints" that are known to exist in animals of high feed conversion efficiency or tolerance to stress. Such animal can then be selected for breeding purposes. Metabolomics has also been used to study the effect of genetic selection on indirect genetic effects (IGE) in breeding programs (Dervishi et al., 2021). Future metabolomics research may be integrated with *multi-omics* experiments using various analytical platforms/techniques (e.g., ICP-MS, MSI, and fluxomics) by using more sensitive platforms, such as ESI-MS, to get more accurate information.

METAGENOMICS

Metagenomics is the collection and analysis of genetic material (genomes) from a mixed community of organisms. Metagenomics is an area of considerable research interest, particularly in ruminant animals to study microbial communities in rumen and milk. In metagenomics, genomic sequencing tools are used to identify the complex structure of the rumen microbiota and their changes in response to diet in concert with the host ruminant genome. These rumen microbiotas may influence a range of phenotypes in the host, including feed efficiency, the inflammatory state in the digestive tract, and volume of methane production in the rumen (Morgan et al., 2014; Ritchie et al., 2015). Metagenomics is also the best way to

reveal modern species' phylogenetic and evolutionary relationships with the natives and ancestors of livestock and poultry (Sahu et al., 2017). Other important applications of metagenomics in livestock improvement are to identify the disease-resistant strains for drug of choice and information generation for genotype and environmental interactions for better control over management (Sahu et al., 2017).

A typical metagenomic experiment involves isolation of genomic DNA from microbial population and amplicon sequencing of 16 rRNA hypervariable V3-V4 region of bacteria and/or WGS by NGS (e.g., IlluminaTM sequencing) or the third-generation sequencing [e.g., Oxford Nanopore TechnologyTM (ONT) and PacBioTM] (Figures 5, 6). The DNA reads obtained from WGS data are assembled computationally to obtain larger DNA sequences and identify the operational taxonomic units (OUTs) of the microbes. Statistical tools can then be utilized to estimate richness (number of taxonomic groups) and evenness (distribution of abundances of the groups) of various microbial populations by computing alpha and beta-diversities. A number of tools such as Mothur, QIIME2 (Quantitative Insights Into Microbial Ecology), DADA2 (Divisive Amplicon Denoising Algorithm), Usearch etc. are available for amplicon sequencing of 16 rDNA in bacteria. A typical bioinformatics pipeline and relevant tools for analysis of amplicon sequencing is shown in Figure 5. On the other hand, while WGS allows high analysis of entire community of microbes, including viruses and fungi, they are relatively expensive, time consuming and computationally demanding. A bioinformatic pipeline and tools for WGS analysis of WGS is shown in Figure 6. The details of various metagenomic pipelines for amplicon sequencing and WGS can be seen elsewhere (Florian et al., 2019).

Metagenomic studies have been used extensively in cattle, pigs, and horses to understand the importance of the microbiome in the gut and mammary microbiome and their relation to feeding efficiency, immunity, and mastitis (Chen C. et al., 2021; Gomez et al., 2021). Metagenomics has shown that gut microbiota can affect feed intake, feed conversion ratio, and production traits such as daily weight gain and back-fat thickness in pigs (Aliakbari et al., 2021; Jiang et al., 2021; Tiezzi et al., 2021). It can also relieve immune stress and help maintain homeostasis in the intestine (Sun et al., 2021). Similarly, parasitic infestations of tapeworm in horses were also found to alter the colonic microbiome (Slater et al., 2021). Such changes in gut microbiota showed implications in individual animals' performance and metabolic health. An improved understanding of gut microbiota by metagenomics can, therefore, help to genetic selection of animals for better animal health and productivity. The metagenomic profile of fecal matter or oral swab may be used to identify animals with "microbial fingerprints" that are known to exist in animals of high genetic merit and can subsequently be used for breeding purposes. Thus, metagenomics may offer a non-invasive means of animal selection and breeding. Metagenomic studies have also revealed that the milk microbiome varies with health status (e.g., mastitis, endometritis, bacteremia, etc.), age, parity, lactation duration, and feed composition (Bach et al., 2021). Consequently, characterization of milk microbiomes may help identify novel "microbial fingerprints" for healthy mammary

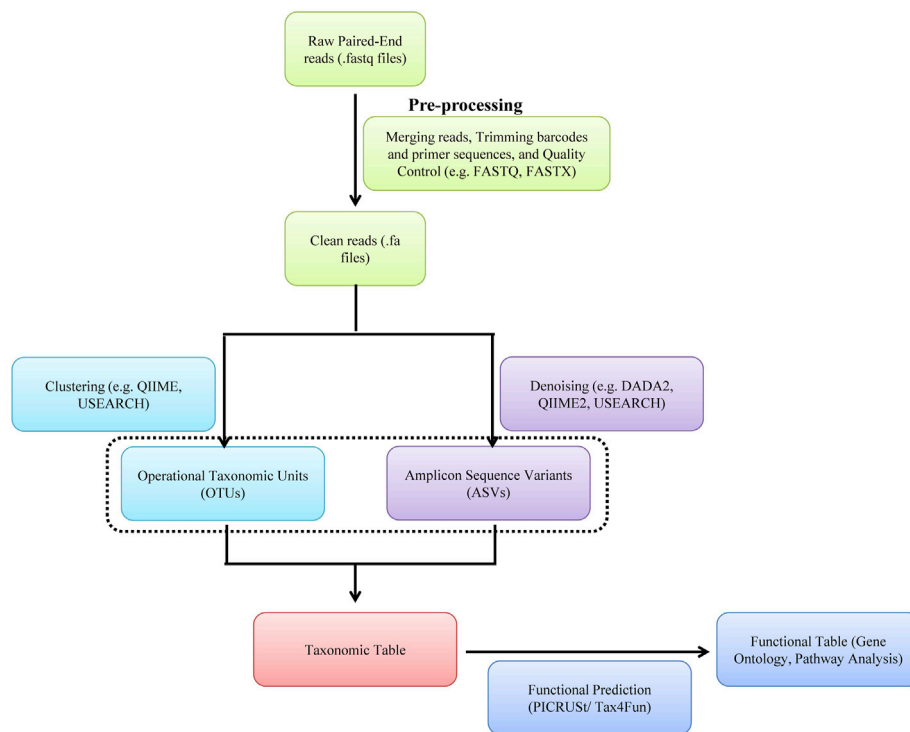


FIGURE 5 | A software pipeline for analysis of amplicon sequencing of bacteria. Each type of experiment has distinct requirements and challenges but there is a common workflow/pipeline.

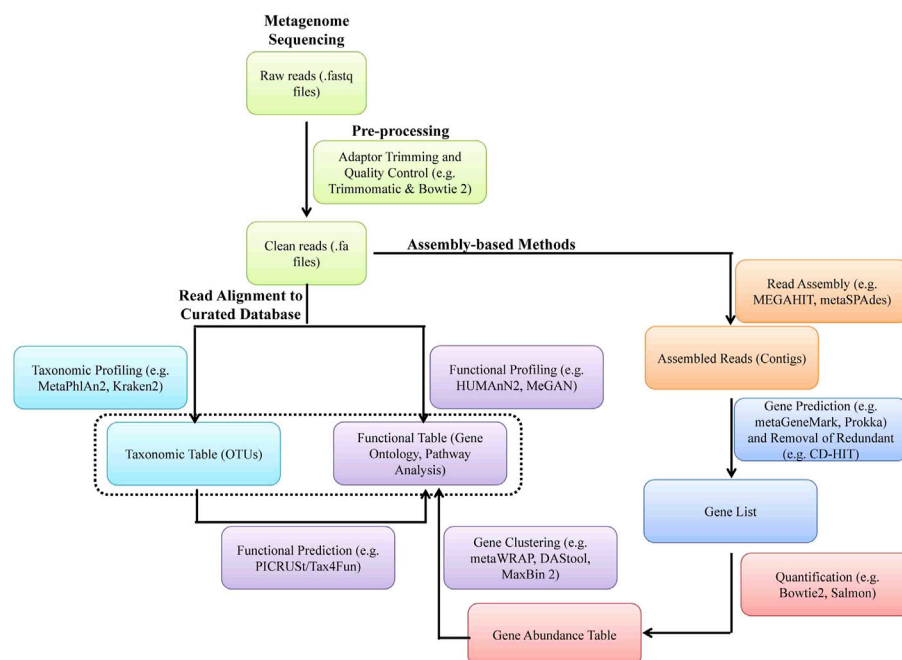


FIGURE 6 | A software pipeline for analysis of whole genome metagenomic sequencing data. Each type of experiment has distinct requirements and challenges but there is a common workflow/pipeline.

TABLE 2 | Overview of some free bioinformatics software for integrating information across several *omics* techniques.

Name	Integration of types of omics	References and URL
Cytoscape	Mainly protein-protein, protein–DNA, and DNA–DNA interactions, but plug-ins (apps) are available for all types of omics	https://cytoscape.org/ ; Shannon et al. (2003)
MOFA	All types (multi-omics)	https://github.com/bioFAM/MOFA ; Argelaguet et al. (2018)
LUCID	Mainly genomics and metabolomics; integration of phenotypic data	https://github.com/USCbiostats/LUCIDus ; Peng et al. (2020)
MultiDataSet	Epigenomics, transcriptomics, assay data, feature data, phenotypic data stored in a single object	https://bioconductor.org/packages/release/bioc/html/MultiDataSet.html ; Hernandez-Ferrer et al. (2017)
Logicome Profiler	Applied to genomics and metagenomics, but applicable to any omics data	https://github.com/fukunagatsu/LogicomeProfiler ; Fukunaga and Iwasaki, (2020)
CoCoNet	Integration of GWAS and gene expression data	http://www.xzlab.org/software.html ; Shang et al. (2020)
NEO	Integration of GWAS and gene expression data	https://horvath.genetics.ucla.edu/html/aten/NEO/ ; Aten et al. (2008)
WGCNA	Mainly gene-expression data, but can be applied to other omics	https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA Zhang and Horvath, (2005); Langfelder and Horvath, (2008)
DIABLO in mixOmics	All types (multi-omics)	http://mixomics.org/mixdiablo Tenenhaus et al., (2014); Rohart et al., (2017); Singh et al. (2019)
The Cancer Genome Atlas (TCGA)	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA	https://cancergenome.nih.gov/
Omics Discovery Index	Genomics, transcriptomics, proteomics, and metabolomics	https://www.omicsdi.org/ Perez-Riverol et al. (2017)
OMICtools	NGS, microarray, polymerase chain reaction (PCR), MS and NMR technologies	http://omictools.com/
NGOMICS-WF	Metagenomic, metatranscriptomic, RNA-seq and 16S data	https://github.com/weizhongli/ngomicswf
Paintomics	Integrated visual analysis of transcriptomics and metabolomics data	http://www.paintomics.org
GalaxyP, GalaxyM	Integrated omics analysis, proteomics informed by transcriptomics analysis	https://usegalaxy.org/
Omics Integrator	Integrate proteomic data, gene expression data and/or epigenetic data using a protein-protein interaction network	http://fraenkel.mit.edu/omicsintegrator , https://github.com/fraenkel-lab/OmicsIntegrator
IMPALA	Joint pathway analysis of transcriptomics or proteomics and metabolomics data	http://impala.molgen.mpg.de

glands and genetic selection of healthy dairy animals (Andrews et al., 2019).

EPIGENOMICS

Epigenomics is another branch of *omics* technology that deals with studying epigenetic changes in a cell. Epigenetic changes regulate gene expression without changing the actual DNA sequence. Epigenetic modifications can be altered by external or internal environmental factors such as diet, exercise, drugs, and chemicals and can change gene expression and define specific phenotypes. Mapping epigenomics components in many cell types helped identify millions of putative regulatory elements (Zentner and Henikoff, 2015). A whole-genomic bisulfate sequencing identified breed-specific hypomethylated regions that were associated with male fertility (Chen S. et al., 2021). Epigenomic biomarkers of male fertility were also identified in the genome-wide DNA methylation map of pig testis (Wang and Kadarmideen 2019). However, studies on global-level epigenomics for genetic selection and breeding of animals are very limited.

ANALYSIS OF OMICS DATA

Omics technologies generate a voluminous amount of complex data in gigabyte to terabyte range (hence, called Big Data) that are difficult to handle by traditional data management tools (Angerer

et al., 2017). Expertise from different biological fields, skilled and knowledgeable bioinformaticians, statisticians, and computer scientists are required to analyze and interpret these data (Riggs et al., 2017). The tremendous high-dimensional data resulting from a large number of experimental variables (e.g., physiological state, age, sex, parity, nutritional status, experimental design, etc.) and simultaneous evaluation of multiple genes/proteins/metabolites/transcripts, etc. requires implementation of complex statistical techniques and models to avoid spurious results and misinterpretation of research data. Various open-source and commercial bioinformatics softwares are now available in online and offline modes in R packages of Bioconductor, EMBOSS, Galaxy, Staden, Biophyhton, Bioconda, Linux, etc., for various *omics* data analysis (Table 2).

Data handling is a vital component of analyzing raw data from *omics* experiments for their correct biological interpretation. Data handling must address issues related to data filtering, imputation, transformation, normalization, quality control, and scaling (Li et al., 2022). Several algorithms and pipelines are now available for the analysis of various *omics* data, including transcriptomics (Figure 3), proteomics (Figure 4), and metagenomics (Figure 4). However, using one pipeline or tool may yield different results from other pipelines. One approach to avoid this problem is to use multiple well-documented analysis pipelines for each step in the pipeline (Misra et al., 2019). The detailed discussion on various *omics* and *multi-omics* pipelines is beyond the manuscript's scope. There are excellent reviews available on

transcriptomic (Wadapurkar et al., 2021), proteomic (Halder et al., 2021), metagenomic (Yang et al., 2021), and metabolomic (Du et al., 2022) pipelines and their integration for *multi-omics* (Subramanian et al., 2020; Reel et al., 2021), which can be referred. GitHub (<https://github.com/danielecook/Awesome-Bioinformatics>) and Biostars (<https://www.biostars.org/>) are also good sources of various updates on *omics*-related softwares and data analysis, respectively.

CHALLENGES IN APPLICATIONS OF OMICS STRATEGIES IN LIVESTOCK SELECTION AND BREEDING

Applications of omics technology to explore the full potential of livestock face many practical challenges. Some of those challenges are as follows:

Proper Maintenance of Data

Data recording and handling of raw data is a big challenge for breeders. To avoid errors and bias in data processing and analysis, suitable cutoffs (e.g., microbial relative abundance, gene expression threshold, metabolite similarity, differential expression cutoff, enriched function cutoff, significant impact value of pathways), data preprocessing options (e.g., data baseline filtering and calibration, peak alignment, deconvolution analysis, peak identification), data normalization, data transformation, and data scaling methods should be carefully considered and addressed within each study. Database construction is an important way for data storage and data maintenance. One such database is ASLive, which has been designed for livestock to capture alternative splicing events in heterogeneous samples from a wide range of tissues, cell types, and biological conditions (Liu et al., 2020). More such databases will accelerate the study and applications of *omics* technology for animal improvement.

Lack of Phenomics Data

Most organized animal farms maintain the performance records of various economic traits, including production traits, reproduction traits, and growth traits. However, organism-wide phenotypic data of animals during different growth phases, various physiological or production stages, in response to dietary changes or upon their selective breeding (i.e., phenome-level data) are challenging to maintain and are generally missing (Pérez-Enciso and Steibel 2021). Accurate phenomics data on adaptability, fitness, body conformation, disease resistance/susceptibility, production performance, reproduction, and growth characteristics will help better estimate accurate BV and selection of genetically superior animals (Juárez et al., 2021; Pérez-Enciso and Steibel 2021). Particular emphasis should be given for *multi-omics* with other “big data”; for example, those detected by advanced management technologies (e.g., using remote sensors communicating with the Internet of Things to measure physiological and behavioral data, which can be applied to monitor estrus, lameness, or rumination) to have complete data set (Sun et al., 2019). The systematic collection of large data sets from different biological layers will help generate a more holistic understanding of the biological factors affecting the performances of the animals.

Expertise

Omics technology generates enormous amounts of data at the genome, transcriptome, proteome, or metabolome levels. Proper handling of *omics* data and advanced knowledge of statistics and bioinformatics are prerequisites for the adequate utilization of *omics* technology. One should have good knowledge in the above-mentioned fields and computer knowledge to interpret the data generated through *omics* technology. Lack of good expertise may mislead for selection of suitable animals. There are many challenges associated with proper data recording, processing, quality control, normalization, and genetic prediction (Yamada et al., 2021). Breeders, biological scientists, veterinarians, statisticians, and computer scientists should be trained to overcome these problems. All should collaborate to interpret and adequately utilize *omics* data, including phenomics data for animal improvement.

CONCLUSION

The goal of animal production is to achieve increased productivity to fulfill human demand while enhancing the health and wellbeing of animals. Population growth, climate change, resource depletion, human health and nutrition, and sustainability are all issues with which the world is grappling. Different new breeding technologies and molecular technologies such as genomic selection, WGS, and gene editing contribute tremendously to the selection and breeding of livestock species for sustainable improvement in productivity and profitability. *Omics* technologies such as genomics, proteomics, transcriptomics, metagenomics, and metabolomics offer powerful analytical tools that can be combined with molecular breeding for the accurate selection of animals for improved productivity. Genomics, in particular, can make conventional breeding and advanced breeding techniques more efficient and precisely targeted by increasing consistency and predictability. Integration of multi-layers of *omics* technology, including phenomics, into the breeding models, will be helpful for proper selection and breeding for animal improvement in the near future.

AUTHOR CONTRIBUTIONS

DC, NS, SK, and MG conceptualized the idea, collected literature, and drafted the manuscript; SS, SL, NS, and MG reviewed, edited, and finalized the manuscript.

FUNDING

This study was supported by the Department of Biotechnology, Government of India (No. BT/PR26321/SPD/9/1307/2017), the National Research Foundation of Korea grant funded by Korean Government (MSIT) (No. 2020R1A2C2004128) and DBT, Government of India (No. BT/PR40124/BTIS/137/14/2021).

REFERENCES

- Aardema, M. J., and MacGregor, J. T. (2003). Toxicology and Genetic Toxicology in the New Era of "Toxicogenomics": Impact of "-omics" Technologies. *Mutat. Res.* 499, 171–193. doi:10.1007/978-4-431-66999-9_22
- Aboshady, H. M., Mandonnet, N., Johansson, A. M., Jonas, E., and Bambou, J. C. (2021). Genomic Variants from RNA-Seq for Goats Resistant or Susceptible to Gastrointestinal Nematode Infection. *PLoS One* 16 (3), e0248405. doi:10.1371/journal.pone.0248405
- Al-Sharif, M., Radwan, H., Hendam, B., and Ateya, A. (2022). DNA Polymorphisms of FGF21, Leptin, κ -casein, and α s1-casein Genes and Their Association with Reproductive Performance in Dromedary She-Camels. *Theriogenology* 178, 18–29. doi:10.1016/j.theriogenology.2021.11.001
- Alemu, S. W., Kadri, N. K., Harland, C., Faux, P., Charlier, C., Caballero, A., et al. (2021). An Evaluation of Inbreeding Measures Using a Whole-Genome Sequenced Cattle Pedigree. *Heredity* 126 (3), 410–423. doi:10.1038/s41437-020-00383-9
- Alessandri, L., Arigoni, M., and Calogero, R. (2019). Differential Expression Analysis in Single-Cell Transcriptomics. *Methods Mol. Biol. Clift. NJ* 1979, 425–432.
- Aliakbari, A., Zemb, O., Billon, Y., Barilly, C., Ahn, I., Riquet, J., et al. (2021). Genetic Relationships between Feed Efficiency and Gut Microbiome in Pig Lines Selected for Residual Feed Intake. *J. animal Breed. Genet.* 138 (4), 491–507. doi:10.1111/jbg.12539
- Almeida, A. M., Bassols, A., Bendixen, E., Bhide, M., Cecilian, F., Cristobal, S., et al. (2015). Animal Board Invited Review: Advances in Proteomics for Animal and Food Sciences. *Animal* 9 (1), 1–17. doi:10.1017/s1751731114002602
- Andrews, T., Neher, D. A., Weicht, T. R., and Barlow, J. W. (2019). Mammary Microbiome of Lactating Organic Dairy Cows Varies by Time, Tissue Site, and Infection Status. *PLoS one* 14 (11), e0225001. doi:10.1371/journal.pone.0225001
- Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics. *Curr. Opin. Syst. Biol.* 4, 85–91. doi:10.1016/j.coisb.2017.07.004
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-Omics Factor Analysis—A Framework for Unsupervised Integration of Multi-omics Data Sets. *Mol. Syst. Biol.* 14 (6), 8124. doi:10.15252/msb.20178124
- Artegoitia, V. M., Newman, J. W., Foote, A. P., Shackelford, S. D., King, D. A., Wheeler, T. L., et al. (2022). Non-invasive Metabolomics Biomarkers of Production Efficiency and Beef Carcass Quality Traits. *Sci. Rep.* 12 (1), 231. doi:10.1038/s41598-021-04049-2
- Aten, J. E., Fuller, T. F., Lusi, A. J., and Horvath, S. (2008). Using Genetic Markers to Orient the Edges in Quantitative Trait Networks: the NEO Software. *BMC Syst. Biol.* 2 (1), 1–21. doi:10.1186/1752-0509-2-34
- Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., and Scorilas, A. (2021). Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life (Basel, Switz.)* 12 (1). doi:10.3390/life12010030
- Bach, A., Joulie, I., Chevaux, E., Elcoso, G., and Ragués, J. (2021). Milk Performance and Rumen Microbiome of Dairy Cows as Affected by the Inclusion of Corn Silage or Corn Shredlage in a Total Mixed Ration. *Animal Int. J. animal Biosci.* 15 (1), 100014. doi:10.1016/j.animal.2020.100014
- Banerjee, P., Carmelo, V. A. O., and Kadarmideen, H. N. (2020). Genome-Wide Epistatic Interaction Networks Affecting Feed Efficiency in Duroc and Landrace Pigs. *Front. Genet.* 11, 121. doi:10.3389/fgene.2020.00121
- Baykalir, Y., Baykalir, B. G., and Simsek, U. G. (2018). "Application of Some Proteome Analysis Techniques in Animal Reproduction." in *New Insights into Theriogenology*. Editor Payan-Carreira, R. (London, UK: IntechOpen), 63–76. doi:10.5772/intechopen.80521
- Bendixen, E. (2005). The Use of Proteomics in Meat Science. *Meat Sci.* 71, 138–149. doi:10.1016/j.meatsci.2005.03.013
- Bhattarai, D., Chen, X., Ur Rehman, Z., Hao, X., Ullah, F., Dad, R., et al. (2017). Association of MAP4K4 Gene Single Nucleotide Polymorphism with Mastitis and Milk Traits in Chinese Holstein Cattle. *J. dairy Res.* 84 (1), 76–79. doi:10.1017/s0022029916000832
- Bovo, S., Ribani, A., Muñoz, M., Alves, E., Araujo, J. P., Bozzi, R., et al. (2020). Whole-genome Sequencing of European Autochthonous and Commercial Pig Breeds Allows the Detection of Signatures of Selection for Adaptation of Genetic Resources to Different Breeding and Production Systems. *Genet. Sel. Evol. GSE* 52 (1), 33. doi:10.1186/s12711-020-00553-7
- Brito, L. F., Bedere, N., Douhard, F., Oliveira, H. R., Arnal, M., Peñagaricano, F., et al. (2021). Review: Genetic Selection of High-Yielding Dairy Cattle toward Sustainable Farming Systems in a Rapidly Changing World. *Animal Int. J. animal Biosci.* 15 (Suppl. 1), 100292. doi:10.1016/j.animal.2021.100292
- Brito, L. F., Oliveira, H. R., McConn, B. R., Schinckel, A. P., Arrazola, A., Marchant-Forde, J. N., et al. (2020). Large-Scale Phenotyping of Livestock Welfare in Commercial Production Systems: A New Frontier in Animal Breeding. *Front. Genet.* 11, 793. doi:10.3389/fgene.2020.00793
- Bu, D., Wang, X., and Tang, H. (2021). Haplotype-based Membership Inference from Summary Genomic Data. *Bioinforma. Oxf. Engl.* 37 (Suppl. 1_1), i161–i168. doi:10.1093/bioinformatics/btab305
- Calus, M. P. L., De Haas, Y., Psczola, M., and Veerkamp, R. F. (2013). Predicted Accuracy of and Response to Genomic Selection for New Traits in Dairy Cattle. *Animal* 7 (2), 183–191. doi:10.1017/s1751731112001450
- Cánovas, A. (2016). Looking Ahead: Applying New Genomic Technologies to Accelerate Genetic Improvement in Beef Cattle. *Ceiba* 54 (1), 41–49.
- Cánovas, A., Reverter, A., DeAtley, K. L., Ashley, R. L., Colgrave, M. L., Fortes, M. R., et al. (2014). Multi-tissue Omics Analyses Reveal Molecular Regulatory Networks for Puberty in Composite Beef Cattle. *PLoS One* 9 (7), 102551. doi:10.1371/journal.pone.0102551
- Cansolo, N. R. B., Silva, J. D., Buarque, V. L. M., Higuera-Padilla, A., Barbosa, L., Zawadzki, A., et al. (2020). Selection for Growth and Precocity Alters Muscle Metabolism in Nellore Cattle. *Metabolites* 10 (2). doi:10.3390/metabo10020058
- Carrillo, J. A., He, Y., Li, Y., Liu, J., Erdman, R. A., Sonstegard, T. S., et al. (2016). Integrated Metabolomic and Transcriptome Analyses Reveal Finishing Forage Affects Metabolic Pathways Related to Beef Quality and Animal Welfare. *Sci. Rep.* 6, 25948. doi:10.1038/srep25948
- Chang, L. Y., Toghiani, S., Aggrey, S. E., and Rekaya, R. (2019). Increasing Accuracy of Genomic Selection in Presence of High Density Marker Panels through the Prioritization of Relevant Polymorphisms. *BMC Genet.* 20 (1), 21. doi:10.1186/s12863-019-0720-5
- Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., et al. (2021a). Expanded Catalog of Microbial Genes and Metagenome-Assembled Genomes from the Pig Gut Microbiome. *Nat. Commun.* 12 (1), 1106. doi:10.1038/s41467-021-21295-0
- Chen, S., Liu, S., Mi, S., Li, W., Zhang, S., Ding, X., et al. (2021b). Comparative Analyses of Sperm DNA Methylomes Among Three Commercial Pig Breeds Reveal Vital Hypomethylated Regions Associated with Spermatogenesis and Embryonic Development. *Front. Genet.* 12, 740036. doi:10.3389/fgene.2021.740036
- Cohen, D., Hondelink, L. M., Solleveld-Westerink, N., Uljee, S. M., Ruano, D., Cleton-Jansen, A. M., et al. (2020). Optimizing Mutation and Fusion Detection in NSCLC by Sequential DNA and RNA Sequencing. *J. Thorac. Oncol. official Publ. Int. Assoc. Study Lung Cancer* 15 (6), 1000–1014. doi:10.1016/j.jtho.2020.01.019
- Cole, J. B., Dürr, J. W., and Nicolazzi, E. L. (2021). Invited Review: The Future of Selection Decisions and Breeding Programs: What Are We Breeding for, and Who Decides? *J. dairy Sci.* 104 (5), 5111–5124. doi:10.3168/jds.2020-19777
- Dekkers, J. C., and Hospital, F. (2002). The Use of Molecular Genetics in the Improvement of Agricultural Populations. *Nat. Rev. Genet.* 3 (1), 22–32. doi:10.1038/nrg701
- Dekkers, J. C. M., Zhao, H. H., Habier, D., and Fernando, R. L. (2009). Opportunities for Genomic Selection with Redesign of Breeding Programs. *J. Animal Sci.* 87 (Suppl. E), 275.
- Dervishi, E., Reimert, I., van der Zande, L. E., Mathur, P., Knol, E. F., and Plastow, G. S. (2021). Relationship between Indirect Genetic Effects for Growth, Environmental Enrichment, Coping Style and Sex with the Serum Metabolome Profile of Pigs. *Sci. Rep.* 11 (1), 23377. doi:10.1038/s41598-021-02814-x
- Do, D. N., Strathe, A. B., Ostensen, T., Pant, S. D., and Kadarmideen, H. N. (2014). Genome-wide Association and Pathway Analysis of Feed Efficiency in Pigs Reveal Candidategenes and Pathways for Residual Feed Intake. *Front. Genet.* 5, 307. doi:10.3389/fgene.2014.00307
- Do, D. N., Strathe, A. B., Ostensen, T., Jensen, J., Mark, T., and Kadarmideen, H. N. (2013). Genome-wide Association Study Reveals Genetic Architecture of Eating

- Behavior in Pigs and its Implications for Humans Obesity by Comparative Mapping. *PLoS One* 8 (8), 71509. doi:10.1371/journal.pone.0071509
- Doublet, A. C., Croiseau, P., Fritz, S., Michenet, A., Hozé, C., Danchin-Burge, C., et al. (2019). The Impact of Genomic Selection on Genetic Diversity and Genetic Gain in Three French Dairy Cattle Breeds. *Genet. Sel. Evol.* 51 (1), 1–13. doi:10.1186/s12711-019-0495-1
- Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward Genomic Prediction from Whole-Genome Sequence Data: Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions. *Heredity* 112 (1), 39–47. doi:10.1038/hdy.2013.13
- Du, X., Aristizabal-Henao, J. J., Garrett, T. J., Brochhausen, M., Hogan, W. R., and Lemas, D. J. (2022). A Checklist for Reproducible Computational Analysis in Clinical Metabolomics Research. *Metabolites* 12 (1). doi:10.3390/metabo12010087
- Ehret, A., Hochstuhl, D., Krattenmacher, N., Tetens, J., Klein, M. S., Gronwald, W., et al. (2015). Use of Genomic and Metabolic Information as Well as Milk Performance Records for Prediction of Subclinical Ketosis Risk via Artificial Neural Networks. *J. Dairy Sci.* 98 (1), 322–329. doi:10.3168/jds.2014-8602
- Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstein, G. M., et al. (2009). Bovine Genome Sequencing and Analysis Consortium. The Genome Sequence of Taurine Cattle: a Window to Ruminant Biology and Evolution. *Science* 324, 522–528. doi:10.1126/science.1169588
- Erasmus, L. M., and van Marle-Köster, E. (2021). Moving towards Sustainable Breeding Objectives and Cow Welfare in Dairy Production: a South African Perspective. *Trop. animal health Prod.* 53 (5), 470. doi:10.1007/s11250-021-02914-w
- Evans, A. M., O'Donovan, C., Playdon, M., Beecher, C., Beger, R. D., Bowden, J. A., et al. (2020). Dissemination and Analysis of the Quality Assurance (QA) and Quality Control (QC) Practices of LC-MS Based Untargeted Metabolomics Practitioners. *Metabolomics* 16 (10), 113. doi:10.1007/s11306-020-01728-5
- Florian, P. B., Lu, J., and Salzberg, S. L. (2019). A Review of Methods and Databases for Metagenomic Classification and Assembly. *Briefings Bioinforma.* 20 (4), 1125–1136.
- Fontanesi, L. (2016). Metabolomics and Livestock Genomics: Insights into a Phenotyping Frontier and its Applications in Animal Breeding. *Anim. Front.* 6, 73–79. doi:10.2527/af.2016-0011
- Fukunaga, T., and Iwasaki, W. (2020). Logicome Profiler: Exhaustive Detection of Statistically Significant Logic Relationships from Comparative Omics Data. *PLoS One* 15 (5), 0232106. doi:10.1371/journal.pone.0232106
- Ghaffari, N., Yousefi, M. R., Johnson, C. D., Ivanov, I., and Dougherty, E. R. (2013). Modeling the Next Generation Sequencing Sample Processing Pipeline for the Purposes of Classification. *BMC Bioinforma.* 14 (1), 1–14. doi:10.1186/1471-2105-14-307
- Ghoreishifar, S. M., Moradi-Shahrababak, H., Fallahi, M. H., Jalil Sarghale, A., Moradi-Shahrababak, M., Abdollahi-Arpanahi, R., et al. (2020). Genomic Measures of Inbreeding Coefficients and Genome-wide Scan for Runs of Homozygosity Islands in Iranian River Buffalo, Bubalus Bubalis. *BMC Genet.* 21 (1), 16. doi:10.1186/s12863-020-0824-y
- Goldansaz, S. A., Guo, A. C., Sajed, T., Steele, M. A., Plastow, G. S., and Wishart, D. S. (2017). Livestock Metabolomics and the Livestock Metabolome: A Systematic Review. *PLoS One* 12 (5), e0177675. doi:10.1371/journal.pone.0177675
- Gomez, A., Sharma, A. K., Grev, A., Sheaffer, C., and Martinson, K. (2021). The Horse Gut Microbiome Responds in a Highly Individualized Manner to Forage Lignification. *J. equine veterinary Sci.* 96, 103306. doi:10.1016/j.jevs.2020.103306
- Gupta, M. K., Jang, J. M., Jung, J. W., Uhm, S. J., Kim, K. P., and Lee, H. T. (2009a). Proteomic Analysis of Parthenogenetic and *In Vitro* Fertilized Porcine Embryos. *Proteomics* 9 (10), 2846–2860. doi:10.1002/pmic.200800700
- Gupta, M. K., Jung, J. W., Uhm, S. J., Lee, H., Lee, H. T., and Kim, K. P. (2009b). Combining Selected Reaction Monitoring with Discovery Proteomics in Limited Biological Samples. *Proteomics* 9 (21), 4834–4836. doi:10.1002/pmic.200900310
- Guttula, P. K., Monteiro, P. T., and Gupta, M. K. (2020). A Boolean Logical Model for Reprogramming of Testes-Derived Male Germline Stem Cells into Germline Pluripotent Stem Cells. *Comput. methods programs Biomed.* 192, 105473. doi:10.1016/j.cmpb.2020.105473
- Guttula, P. K., Monteiro, P. T., and Gupta, M. K. (2021). Prediction and Boolean Logical Modelling of Synergistic microRNA Regulatory Networks during Reprogramming of Male Germline Pluripotent Stem Cells. *Bio Syst.* 207, 104453. doi:10.1016/j.biosystems.2021.104453
- Halder, A., Verma, A., Biswas, D., and Srivastava, S. (2021). Recent Advances in Mass-Spectrometry Based Proteomics Software, Tools and Databases. *Drug Discov. today Technol.* 39, 69–79. doi:10.1016/j.ddtec.2021.06.007
- Hawken, R. J., Zhang, Y. D., Fortes, M. R. S., Collis, E., Barris, W. C., Corbet, N. J., et al. (2012). Genome-wide Association Studies of Female Reproduction in Tropically Adapted Beef Cattle. *J. Animal Sci.* 90 (5), 1398–1410. doi:10.2527/jas.2011-4410
- Hernandez-Ferrer, C., Ruiz-Arenas, C., Beltran-Gomila, A., and González, J. R. (2017). MultiDataSet: an R Package for Encapsulating Multiple Data Sets with Application to Omic Data Integration. *BMC Bioinforma.* 18 (1), 1–7. doi:10.1186/s12859-016-1455-1
- Ibtisham, F., Zhang, L., Xiao, M., An, L., Ramzan, M. B., Nawab, A., et al. (2017). Genomic Selection and its Application in Animal Breeding. *Thai J. Veterinary Med.* 47 (3), 301.
- Jaiswal, S., Jagannadham, J., Kumari, J., Iquebal, M. A., Gurjar, A. K. S., Nayan, V., et al. (2021). Genome Wide Prediction, Mapping and Development of Genomic Resources of Mastitis Associated Genes in Water Buffalo. *Front. veterinary Sci.* 8, 593871. doi:10.3389/fvets.2021.593871
- Jiang, H., Fang, S., Yang, H., and Chen, C. (2021). Identification of the Relationship between the Gut Microbiome and Feed Efficiency in a Commercial Pig Cohort. *J. Anim. Sci.* 99 (3). doi:10.1093/jas/skab045
- Jorge-Smeding, E., Bonnet, M., Renand, G., Taussat, S., Graulet, B., Ortigues-Marty, I., et al. (2021). Common and Diet-specific Metabolic Pathways Underlying Residual Feed Intake in Fattening Charolais Yearling Bulls. *Sci. Rep.* 11 (1), 24346. doi:10.1038/s41598-021-03678-x
- Juárez, M., Lam, S., Bohrer, B. M., Dugan, M. E. R., Vahmani, P., Aalhus, J., et al. (2021). Enhancing the Nutritional Value of Red Meat through Genetic and Feeding Strategies. *Foods (Basel, Switz.)* 10 (4).
- Kadarmideen, H. N. (2014). Genomics to Systems Biology in Animal and Veterinary Sciences: Progress, Lessons and Opportunities. *Livest. Sci.* 166, 232–248. doi:10.1016/j.livsci.2014.04.028
- Kalaiselvi, G., Tirumurugan, K. G., Vijayarani, K., Dhinakar Raj, G., Baranidharan, G. R., and Bobade, S. (2019). Livestock Metabolomics-An Overview. *Int. J. Curr. Res.* 11 (3), 1972–1980.
- Kasper, C., Ribeiro, D., Almeida, A. M. D., Larzul, C., Liaubet, L., and Murani, E. (2020). Omics Application in Animal Science—A Special Emphasis on Stress Response and Damaging Behaviour in Pigs. *Genes* 11 (8), 920. doi:10.3390/genes11080920
- Kaya, A., Dogan, S., Vargovic, P., Kutchy, N. A., Ross, P., Topper, E., et al. (2022). Sperm Proteins ODF2 and PAWP as Markers of Fertility in Breeding Bulls. *Cell. tissue Res.* 387 (1), 159–171. doi:10.1007/s00441-021-03529-1
- Kogelman, L. J., Pant, S. D., Fredholm, M., and Kadarmideen, H. N. (2014). Systems Genetics of Obesity in an F2 Pig Model by Genome-wide Association, Genetic Network, and Pathway Analyses. *Front. Genet.* 5, 214. doi:10.3389/fgene.2014.00214
- Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front. Genet.* 11, 610798. doi:10.3389/fgene.2020.610798
- Kuska, B. (1998). Beer, Bethesda, and Biology: How “Genomics” Came into Being. *J. Natl. Cancer Inst.* 90, 93. doi:10.1093/jnci/90.2.93
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinforma.* 9 (1), 1–13. doi:10.1186/1471-2105-9-559
- Li, C., Gao, Z., Su, B., Xu, G., and Lin, X. (2022). Data Analysis Methods for Defining Biomarkers from Omics Data. *Anal. Bioanal. Chem.* 414 (1), 235–250. doi:10.1007/s00216-021-03813-7
- Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data. *Biostatistics* 13 (3), 523–538. doi:10.1093/biostatistics/kxr031
- Li, X., and Wang, C. Y. (2021). From Bulk, Single-Cell to Spatial RNA Sequencing. *Int. J. oral Sci.* 13 (1), 36. doi:10.1038/s41368-021-00146-0
- Lillehammer, M., Meuwissen, T. H. E., and Sonesson, A. K. (2011). Genomic Selection for Maternal Traits in Pigs. *J. animal Sci.* 89 (12), 3908–3916. doi:10.2527/jas.2011-4044
- Lippa, K. A., Aristizabal-Henao, J. J., Beger, R. D., Bowden, J. A., Broeckling, C., Beecher, C., et al. (2022). Reference Materials for MS-based Untargeted

- Metabolomics and Lipidomics: a Review by the Metabolomics Quality Assurance and Quality Control Consortium (mQACC). *Metabolomics* 18 (4), 24. doi:10.1007/s11306-021-01848-6
- Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., et al. (2017). Identification of Individuals by Trait Prediction Using Whole-Genome Sequencing Data. *Proc. Natl. Acad. Sci. U. S. A.* 114 (38), 10166–10171. doi:10.1073/pnas.1711125114
- Liu, J., Tan, S., Huang, S., and Huang, W. (2020). ASlive: a Database for Alternative Splicing Atlas in Livestock Animals. *BMC Genomics* 21, 97. doi:10.1186/s12864-020-6472-9
- Ma, B., Khan, R., Raza, S. H. A., Gao, Z., Hou, S., Ullah, F., et al. (2021). Determination of the Relationship between Class IV Sirtuin Genes and Growth Traits in Chinese Black Tibetan Sheep. *Anim. Biotechnol.* 2021, 1–7. doi:10.1080/10495398.2021.2016434
- Malheiros, J. M., Correia, B. S. B., Ceribeli, C., Cardoso, D. R., Colnago, L. A., Junior, S. B., et al. (2021). Comparative Untargeted Metabolome Analysis of Ruminal Fluid and Feces of Nelore Steers (*Bos indicus*). *Sci. Rep.* 11 (1), 12752. doi:10.1038/s41598-021-92179-y
- Martin, M. J., Pralle, R. S., Bernstein, I. R., VandeHaar, M. J., Weigel, K. A., Zhou, Z., et al. (2021). Circulating Metabolites Indicate Differences in High and Low Residual Feed Intake Holstein Dairy Cows. *Metabolites* 11 (12). doi:10.3390/metabo11120868
- Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., et al. (2009). Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS One* 4 (4), 5350. doi:10.1371/journal.pone.0005350
- Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi:10.1530/jme-18-0055
- Morgan, J. L., Ritchie, L. E., Crucian, B. E., Theriot, C., Wu, H., Sams, C., et al. (2014). Increased Dietary Iron and Radiation in Rats Promote Oxidative Stress, Induce Localized and Systemic Immune System Responses, and Alter Colon Mucosal Environment. *FASEB J.* 28 (3), 1486–1498. doi:10.1096/fj.13-239418
- Moridi, M., Moghaddam, S. H. H., Mirhoseini, S. Z., and Bionaz, M. (2019). Transcriptome Analysis Showed Differences of Two Purebred Cattle and Their Crossbreds. *Italian J. Animal Sci.* 18 (1), 70–79. doi:10.1080/1828051x.2018.1482800
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat. methods* 5 (7), 621–628. doi:10.1038/nmeth.1226
- Mu, Y., Qi, W., Zhang, T., Zhang, J., and Mao, S. (2022). Multi-omics Analysis Revealed Coordinated Responses of Rumen Microbiome and Epithelium to High-Grain-Induced Subacute Rumen Acidosis in Lactating Dairy Cows. *mSystems* 7 (1), e0149021. doi:10.1128/msystems.01490-21
- Muhanguzi, D., Ndekezi, C., Nkamwesiga, J., Kalayou, S., Ochwo, S., Vuyani, M., et al. (2022). Anti-Tick Vaccines: Current Advances and Future Prospects. *Methods Mol. Biol. Clift. NJ* 2411, 253–267. doi:10.1007/978-1-0716-1888-2_15
- Mullins, Y., Keogh, K., Blackshields, G., Kenny, D. A., Kelly, A. K., and Waters, S. M. (2021). Transcriptome Assisted Label Free Proteomics of Hepatic Tissue in Response to Both Dietary Restriction and Compensatory Growth in Cattle. *J. proteomics* 232, 104048. doi:10.1016/j.jpro.2020.104048
- Muñoz, M., Bozzi, R., García-Casco, J., Núñez, Y., Ribani, A., Franci, O., et al. (2019). Genomic Diversity, Linkage Disequilibrium and Selection Signatures in European Local Pig Breeds Assessed with a High Density SNP Chip. *Sci. Rep.* 9 (1), 13546.
- Ng, B., Casazza, W., Kim, N. H., Wang, C., Farhadi, F., Tasaki, S., et al. (2021). Cascading Epigenomic Analysis for Identifying Disease Genes from the Regulatory Landscape of GWAS Variants. *PLoS Genet.* 17 (11), e1009918. doi:10.1371/journal.pgen.1009918
- Odom, G. J., Colaprico, A., Silva, T. C., Chen, X. S., and Wang, L. (2021). PathwayMultiomics: An R Package for Efficient Integrative Analysis of Multi-Omics Datasets with Matched or Un-matched Samples. *Front. Genet.* 12, 783713. doi:10.3389/fgene.2021.783713
- Oskoueian, E., Eckersall, P. D., Bencurova, E., and Dandekar, T. (2016). “Application of Proteomic Biomarkers in Livestock Disease Management,” in *Agricultural Proteomics- Volume 2*. Editor G. H. Salekdeh (Philippines: Springer), 299–310. doi:10.1007/978-3-319-43278-6_14
- Peddinti, D., Memili, E., and Burgess, S. C. (2011). 13 Proteomics in Animal Reproduction and Breeding. *Methods Animal Proteomics* 369.
- Pedrosa, V. B., Schenkel, F. S., Chen, S. Y., Oliveira, H. R., Casey, T. M., Melka, M. G., et al. (2021). Genomewide Association Analyses of Lactation Persistency and Milk Production Traits in Holstein Cattle Based on Imputed Whole-Genome Sequence Data. *Genes* 12 (11). doi:10.3390/genes12111830
- Peng, C., Wang, J., Asante, I., Louie, S., Jin, R., Chatzi, L., et al. (2020). A Latent Unknown Clustering Integrating Multi-Omics Data (LUCID) with Phenotypic Traits. *Bioinformatics* 36 (3), 842–850. doi:10.1093/bioinformatics/btz667
- Pérez-Enciso, M., and Steibel, J. P. (2021). Phenomes: the Current Frontier in Animal Breeding. *Genet. Sel. Evol. GSE* 53 (1), 22.
- Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., et al. (2017). Discovering and Linking Public Omics Data Sets Using the Omics Discovery Index. *Nat. Biotechnol.* 35 (5), 406–409. doi:10.1038/nbt.3790
- Pimentel, E. C. G., and König, S. (2012). Genomic Selection for the Improvement of Meat Quality in Beef. *J. Animal Sci.* 90 (10), 3418–3426. doi:10.2527/jas.2011-5005
- Plieschke, L., Edel, C., Pimentel, E. C., Emmerling, R., Bennewitz, J., and Götz, K. U. (2016). Systematic Genotyping of Groups of Cows to Improve Genomic Estimated Breeding Values of Selection Candidates. *Genet. Sel. Evol.* 48 (1), 1–11. doi:10.1186/s12711-016-0250-9
- Pryce, J. E., and Daetwyler, H. D. (2011). Designing Dairy Cattle Breeding Schemes under Genomic Selection: a Review of International Research. *Animal Prod. Sci.* 52 (3), 107–114.
- Raza, S. H. A., Khan, R., Pant, S. D., Shah, M. A., Quan, G., Feng, L., et al. (2021). Genetic Variation in the OPN Gene Affects Milk Composition in Chinese Holstein Cows. *Anim. Biotechnol.* 2021, 1–7. doi:10.1080/10495398.2021.2001343
- Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739
- Reverter, A., and Fortes, M. R. S. (2013). Breeding and Genetics Symposium: Building Single Nucleotide Polymorphism-Derived Gene Regulatory Networks: Towards Functional Genome-wide Association Studies. *J. Animal Sci.* 91 (2), 530–536. doi:10.2527/jas.2012-5780
- Rexroad, C., Vallet, J., Matukumalli, L. K., Reecy, J., Bickhart, D., Blackburn, H., et al. (2019). Genome to Phenome: Improving Animal Health, Production, and Well-Being - A New USDA Blueprint for Animal Genome Research 2018-2027. *Front. Genet.* 10, 327. doi:10.3389/fgene.2019.00327
- Riggs, P. K., and Gill, C. A. (2009). “Molecular Mapping and Marker Assisted Breeding for Meat Quality,” in *Applied Muscle Biology and Meat Science*. Editor M. Du (Boca Raton: CRC Press), 288–310.
- Riggs, P. K., Tedeschi, L. O., Turner, N. D., Braga-Neto, U., and Jayaraman, A. (2017). The Role of “Omics” Technologies for Livestock Sustainability. *Arch. Prod. Anim.* 25 (3-4).
- Ritchie, L. E., Sturino, J. M., Carroll, R. J., Rooney, L. W., Azcarate-Peril, M. A., and Turner, N. D. (2015). Polyphenol-rich Sorghum Brans Alter Colon Microbiota and Impact Species Diversity and Species Richness after Multiple Bouts of Dextran Sodium Sulfate-Induced Colitis. *FEMS Microbiol. Ecol.* 91 (3). doi:10.1093/femsec/fiv008
- Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017). mixOmics: An R Package for ‘omics Feature Selection and Multiple Data Integration. *PLoS Comput. Biol.* 13 (11), 1005752. doi:10.1371/journal.pcbi.1005752
- Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J. P., Friel, L. A., Erez, O., et al. (2006). The Use of High-Dimensional Biology (Genomics, Transcriptomics, Proteomics, and Metabolomics) to Understand the Preterm Parturition Syndrome. *BJOG Int. J. obstetrics Gynaecol.* 113 (Suppl. 3), 118–135. doi:10.1111/j.1471-0528.2006.01150.x
- Ruan, D., Zhuang, Z., Ding, R., Qiu, Y., Zhou, S., Wu, J., et al. (2021). Weighted Single-step GWAS Identified Candidate Genes Associated with Growth Traits in a Duroc Pig Population. *Genes* 12 (1). doi:10.3390/genes12010117
- Russell, C. D., Widdison, S., Leigh, J. A., and Coffey, T. J. (2012). Identification of Single Nucleotide Polymorphisms in the Bovine Toll-like Receptor 1 Gene and Association with Health Traits in Cattle. *Veterinary Res.* 43 (1), 17. doi:10.1186/1297-9716-43-17
- Sahoo, B., Choudhary, R. K., Sharma, P., Choudhary, S., and Gupta, M. K. (2021a). Significance and Relevance of Spermatozoal RNAs to Male Fertility in Livestock. *Front. Genet.* 12, 768196. doi:10.3389/fgene.2021.768196

- Sahoo, B., Guttula, P. K., and Gupta, M. K. (2021b). Comparison of Spermatozoal RNA Extraction Methods in Goats. *Anal. Biochem.* 614, 114059. doi:10.1016/j.ab.2020.114059
- Sahu, A., Nayak, N., Sahu, R., and Kumar, J. (2017). Application of Metagenomics in Livestock Improvement. *Int. J. Livest. Res.* 7, 30–38. doi:10.5455/ijlr.20170423033727
- Sakuma, H., Saito, K., Kohira, K., Ohhashi, F., Shoji, N., and Uemoto, Y. (2016). Estimates of Genetic Parameters for Chemical Traits of Meat Quality in Japanese Black Cattle. *Anim. Sci. J.* 88 (2), 203–212. doi:10.1111/asj.12622
- Seidel, D., Martínez, I., Taddeo, S., Joseph, M., Carroll, R., Haub, M., et al. (2014). A Polyphenol-rich Sorghum Cereal Alters Colon Microbiota and Plasma Metabolites in Overweight Subjects (270.7). *FASEB J.* 28, 270–277. doi:10.1096/fasebj.28.1_supplement.270.7
- Shang, L., Smith, J. A., and Zhou, X. (2020). Leveraging Gene Co-expression Patterns to Infer Trait-Relevant Tissues in Genome-wide Association Studies. *PLoS Genet.* 16 (4), 1008734. doi:10.1371/journal.pgen.1008734
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303
- Sharma, A., Lee, J. S., Dang, C. G., Sudrajat, P., Kim, H. C., Yeon, S. H., et al. (2015). Stories and Challenges of Genome Wide Association Studies in Livestock—A Review. *Asian-Australasian J. Animal Sci.* 28 (10), 1371. doi:10.5713/ajas.14.0715
- Shi, T. P., and Zhang, L. (2019). Application of Whole Transcriptomics in Animal Husbandry. *Yi Chuan= Hered.* 41 (3), 193–205. doi:10.16288/j.yczz.18-218
- Shumbusho, F., Raoul, J., Astruc, J. M., Palhière, I., and Elsen, J. M. (2013). Potential Benefits of Genomic Selection on Genetic Gain of Small Ruminant Breeding Programs. *J. Animal Sci.* 91 (8), 3644–3657. doi:10.2527/jas.2012-6205
- Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: Anintegrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays. *Bioinformatics* 35, 3055–3062. doi:10.1093/bioinformatics/bty1054
- Sitzenstock, F., Ytournal, F., Sharifi, A. R., Caverio Täubert, D. H., Preisinger, R., and Simianer, H. (2013). Efficiency of Genomic Selection in an Established Commercial Layer Breeding Program. *Genet. Sel. Evol.* 45, 1–11. doi:10.1186/1297-9686-45-29
- Snelling, W. M., Cushman, R. A., Keele, J. W., Maltecca, C., Thomas, M. G., Fortes, M. R. S., et al. (2013). Breeding and Genetics Symposium: Networks and Pathways to Guide Genomic Selection. *J. Animal Sci.* 91 (2), 537–552. doi:10.2527/jas.2012-5784
- Speicher, D. W. (2004). “Overview of Proteome Analysis,” in *Proteome Analysis* (Philippines: Elsevier), 1–18. doi:10.1016/b978-044451024-2/50018-7
- Stock, J., Bennewitz, J., Hinrichs, D., and Wellmann, R. (2020). A Review of Genomic Models for the Analysis of Livestock Crossbred Data. *Front. Genet.* 11, 568. doi:10.3389/fgene.2020.00568
- Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and its Application. *Bioinforma. Biol. insights* 14, 1177932219899051. doi:10.1177/1177932219899051
- Sumreddee, P., Hay, E. H., Toghiani, S., Roberts, A., Aggrey, S. E., and Rekaya, R. (2021). Grid Search Approach to Discriminate between Old and Recent Inbreeding Using Phenotypic, Pedigree and Genomic Information. *BMC genomics* 22 (1), 538. doi:10.1186/s12864-021-07872-z
- Sun, H. Z., Plastow, G., and Guan, L. L. (2019). Invited Review: Advances and Challenges in Application of Feedomics to Improve Dairy Cow Production and Health. *J. Dairy Sci.* 102 (7), 5853–5870. doi:10.3168/jds.2018-16126
- Sun, H. Z., Wang, D. M., Wang, B., Wang, J. K., Liu, H. Y., Guan, L. L., et al. (2015). Metabolomics of Four Biofluids from Dairy Cows: Potential Biomarkers for Milk Production and Quality. *J. Proteome Res.* 14 (2), 1287–1298. doi:10.1021/pr501305g
- Sun, X., Cui, Y., Su, Y., Gao, Z., Diaio, X., Li, J., et al. (2021). Dietary Fiber Ameliorates Lipopolysaccharide-Induced Intestinal Barrier Function Damage in Piglets by Modulation of Intestinal Microbiome. *mSystems* 6 (2), doi:10.1128/msystems.01374-20
- Suravajhala, P., Kogelman, L. J., and Kadarmideen, H. N. (2016). Multi-omic Data Integration and Analysis Using Systems Genomics Approaches: Methods and Applications in Animal Production, Health and Welfare. *Genet. Sel. Evol.* 48 (1), 1–14. doi:10.1186/s12711-016-0217-x
- Tan, C., Bian, C., Yang, D., Li, N., Wu, Z. F., and Hu, X. X. (2017). Application of Genomic Selection in Farm Animal Breeding. *Yi Chuan* 39 (11), 1033–1045. doi:10.16288/j.yczz.17-286
- Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K. A., Grill, J., and Frouin, V. (2014). Variable Selection for Generalized Canonical Correlation Analysis. *Biostatistics* 15 (3), 569–583. doi:10.1093/biostatistics/kxu001
- The Bovine HapMap Consortium (2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* 324, 528–532. doi:10.1126/science.1169736
- The Bovine Genome Sequencing and Analysis ConsortiumElsik, C. G., Tellam, R. L., and Worley, K. C. (2009). The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* 324, 522–528. doi:10.1126/science.1169588
- Tiezzi, F., Fix, J., Schwab, C., Shull, C., and Maltecca, C. (2021). Gut Microbiome Mediates Host Genomic Effects on Phenotypes: a Case Study with Fat Deposition in Pigs. *Comput. Struct. Biotechnol. J.* 19, 530–544. doi:10.1016/j.csbj.2020.12.038
- Ueda, S., Iwamoto, E., Kato, Y., Shinohara, M., Shirai, Y., and Yamanoue, M. (2018). Comparative Metabolomics of Japanese Black Cattle Beef and Other Meats Using Gas Chromatography-Mass Spectrometry. *Biosci. Biotechnol. Biochem.* 83 (1), 137–147. doi:10.1080/09168451.2018.1528139
- Valerio, A., Casadei, L., Giuliani, A., and Valerio, M. (2020). Fecal Metabolomics as a Novel Noninvasive Method for Short-Term Stress Monitoring in Beef Cattle. *J. Proteome Res.* 19 (2), 845–853. doi:10.1021/acs.jproteome.9b00655
- Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., et al. (2022). Identifying Temporal and Spatial Patterns of Variation from Multimodal Data Using MEFISTO. *Nat. methods* 19, 179. doi:10.1038/s41592-021-01343-9
- Wadapurkar, R. M., Sivaram, A., and Vyas, R. (2021). Computational Studies Reveal Co-occurrence of Two Mutations in IL7R Gene of High-Grade Serous Carcinoma Patients. *J. Biomol. Struct. Dyn.* 2021, 1–15. doi:10.1080/07391102.2021.1987326
- Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., and Vitezica, Z. (2014). Genome-wide Association Mapping Including Phenotypes from Relatives without Genotypes in a Single-step (ssGWAS) for 6-week Bodyweight in Broiler Chickens. *Front. Genet.* 5, 134. doi:10.3389/fgene.2014.00134
- Wang, X., and Kadarmideen, H. N. (2019). An Epigenome-wide DNA Methylation Map of Testis in Pigs for Study of Complex Traits. *Front. Genet.* 10, 405. doi:10.3389/fgene.2019.00405
- Wang, X., and Kadarmideen, H. N. (2020). Metabolite Genome-wide Association Study (mGWAS) and Gene-Metabolite Interaction Network Analysis Reveal Potential Biomarkers for Feed Efficiency in Pigs. *Metabolites* 10 (5), doi:10.3390/metabo10050201
- Wang, Y., Li, J., Lu, D., Meng, Q., Song, N., Zhou, H., et al. (2022). Integrated Proteome and Phosphoproteome Analysis of Interscapular Brown Adipose and Subcutaneous White Adipose Tissues upon High Fat Diet Feeding in Mouse. *J. proteomics* 255, 104500. doi:10.1016/j.jpro.2022.104500
- Wickramasinghe, S., Cánovas, A., Rincon, G., and Medrano, J. F. (2014). Review: RNA-Seq Applications in Livestock. *Livest. Rev.* 166, 206–216. doi:10.1016/j.livsci.2014.06.015
- Willforss, J., Morrell, J. M., Resjö, S., Hallap, T., Padrik, P., Siino, V., et al. (2021). Stable Bull Fertility Protein Markers in Seminal Plasma. *J. proteomics* 236, 104135. doi:10.1016/j.jpro.2021.104135
- Williams, J. L. (2005). The Use of Marker-Assisted Selection in Animal Breeding and Biotechnology. *Revue Sci. Tech. Int. Office Epizootics* 24 (1), 379–391. doi:10.20506/rst.24.1.1571
- Wu, J., Ye, Y., Quan, J., Ding, R., Wang, X., Zhuang, Z., et al. (2021). Using Nontargeted LC-MS Metabolomics to Identify the Association of Biomarkers in Pig Feces with Feed Efficiency. *Porc. health Manag.* 7 (1), 39. doi:10.1186/s40813-021-00219-w
- Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., et al. (2018). Integrative Analysis of Omics Summary Data Reveals Putative Mechanisms Underlying Complex Traits. *Nat. Commun.* 9 (1), 918. doi:10.1038/s41467-018-03371-0
- Yadav, S. P. (2007). The Wholeness in Suffix-Omics, Omes, and the Word Om. *J. Biomol. Tech.* 18 (5), 277.

- Yamada, R., Okada, D., Wang, J., Basak, T., and Koyama, S. (2021). Interpretation of Omics Data Analyses. *J. Hum. Genet.* 66 (1), 93–102. doi:10.1038/s10038-020-0763-5
- Yang, A. Q., Chen, B., Ran, M. L., Yang, G. M., and Zeng, C. (2020). The Application of Genomic Selection in Pig Cross Breeding. *Yi chuan = Hered.* 42 (2), 145–152. doi:10.16288/j.yczz.19-253
- Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., et al. (2021). A Review of Computational Tools for Generating Metagenome-Assembled Genomes from Metagenomic Sequencing Data. *Comput. Struct. Biotechnol. J.* 19, 6301–6314. doi:10.1016/j.csbj.2021.11.028
- Yarmush, M. L., and Jayaraman, A. (2002). Advances in Proteomic Technologies. *Annu. Rev. Biomed. Eng.* 4 (1), 349–373. doi:10.1146/annurev.bioeng.4.020702.153443
- Ye, J., Yan, X., Qin, P., Gong, X., Li, H., Liu, Y., et al. (2022). Proteomic Analysis of Hypothalamus in Prepubertal and Pubertal Female Goat. *J. proteomics* 251, 104411. doi:10.1016/j.jprot.2021.104411
- Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4 (1), 4. doi:10.2202/1544-6115.1128
- Zhang, C., Gao, Y., Liu, J., Xue, Z., Lu, Y., Deng, L., et al. (2018). PGG.Population: a Database for Understanding the Genomic Diversity and Genetic Ancestry of Human Populations. *Nucleic acids Res.* 46 (D1), D984–d993. doi:10.1093/nar/gkx1032
- Zhang, S., Gao, X., Jiang, Y., Shen, Y., Xie, H., Pan, P., et al. (2021). Population Validation of Reproductive Gene Mutation Loci and Association with the Litter Size in Nubian Goat. *Arch. Anim. Breed.* 64 (2), 375–386. doi:10.5194/aab-64-375-2021
- Zhao, Y., Wang, Y., Guo, F., Lu, B., Sun, J., Wang, J., et al. (2021). iTRAQ-based Proteomic Analysis of Sperm Reveals Candidate Proteins that Affect the Quality of Spermatozoa from Boars on Plateaus. *Proteome Sci.* 19 (1), 9. doi:10.1186/s12953-021-00177-9
- Zierer, J., Jackson, M. A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., et al. (2018). The Fecal Metabolome as a Functional Readout of the Gut Microbiome. *Nat. Genet.* 50 (6), 790–795. doi:10.1038/s41588-018-0135-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chakraborty, Sharma, Kour, Sodhi, Gupta, Lee and Son. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Sunday O Peters,
Berry College, United States

REVIEWED BY

Tara G McDanel,
United States Department of
Agriculture, United States
Ikhide G. Imumorin,
Georgia Institute of Technology,
United States

*CORRESPONDENCE

Geoffrey E. Pollott,
gpollott@rvc.ac.uk

SPECIALTY SECTION

This article was submitted to Livestock
Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 09 August 2021

ACCEPTED 21 July 2022

PUBLISHED 29 August 2022

CITATION

Pollott GE, Piercy RJ, Massey C,
Salavati M, Cheng Z and Wathes DC
(2022), Locating a novel autosomal
recessive genetic variant in the cattle
glucokinase gene using only WGS data
from three cases and six carriers.
Front. Genet. 13:755693.
doi: 10.3389/fgene.2022.755693

COPYRIGHT

© 2022 Pollott, Piercy, Massey, Salavati,
Cheng and Wathes. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Locating a novel autosomal recessive genetic variant in the cattle glucokinase gene using only WGS data from three cases and six carriers

Geoffrey E. Pollott^{1*}, Richard J. Piercy², Claire Massey²,
Mazdak Salavati^{3,4}, Zhangrui Cheng³ and D. Claire Wathes³

¹Department of Pathobiology and Population Sciences, The Royal Veterinary College, London, United Kingdom, ²Comparative Neuromuscular Diseases Laboratory, Department of Clinical Sciences and Services, Royal Veterinary College, London, United Kingdom, ³Department of Pathobiology and Population Sciences, The Royal Veterinary College, Hatfield, United Kingdom, ⁴The Roslin Institute, Midlothian, United Kingdom

New Mendelian genetic conditions, which adversely affect livestock, arise all the time. To manage them effectively, some methods need to be devised that are quick and accurate. Until recently, finding the causal genomic site of a new autosomal recessive genetic disease has required a two-stage approach using single-nucleotide polymorphism (SNP) chip genotyping to locate the region containing the new variant. This region is then explored using fine-mapping methods to locate the actual site of the new variant. This study explores bioinformatic methods that can be used to identify the causative variants of recessive genetic disorders with full penetrance with just nine whole genome-sequenced animals to simplify and expedite the process to a one-step procedure. Using whole genome sequencing of only three cases and six carriers, the site of a novel variant causing perinatal mortality in Irish moiled calves was located. Four methods were used to interrogate the variant call format (VCF) data file of these nine animals, they are genotype criteria (GCR), autozygosity-by-difference (ABD), variant prediction scoring, and registered SNP information. From more than nine million variants in the VCF file, only one site was identified by all four methods (Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)). This site was a splice acceptor variant located in the glucokinase gene (*GCK*). It was verified on an independent sample of animals from the breed using genotyping by polymerase chain reaction at the candidate site and autozygosity-by-difference using SNP-chips. Both methods confirmed the candidate site. Investigation of the GCR method found that sites meeting the GCR were not evenly spread across the genome but concentrated in regions of long runs of homozygosity. Locating GCR sites was best performed using two carriers to every case, and the carriers should be distantly related to the cases, within the breed concerned. Fewer than 20 animals need to be sequenced when using the GCR and ABD methods together. The genomic site of novel autosomal recessive Mendelian genetic diseases can be located using fewer than 20 animals combined with two bioinformatic methods, autozygosity-by-difference, and genotype criteria. In

many instances it may also be confirmed with variant prediction scoring. This should speed-up and simplify the management of new genetic diseases to a single-step process.

KEYWORDS

cattle, glucokinase gene, recessive genetics, runs of homozygosity, WGS, Irish Moiled, perinatal mortality

Introduction

A review of 34 articles detailing work to map 38 novel autosomal recessive genetic conditions to their position on the genome (Pollott, 2018) suggested that finding the site of such a condition required the use of at least a two-stage methodology. First, the use of a suitable single-nucleotide polymorphism (SNP) chip in a case/control study to locate the *region* containing the new variant combined with a method searching for long runs of homozygosity (ROH), the more traditional chi-squared method being shown to be inadequate. The second stage used a range of “fine-mapping” methods to search within the highlighted region for the *site* of the new variant, many of which resulted in whole genome sequencing (WGS) of a few cases and controls. More recent methods have been developed, which use WGS as an initial step but such methods typically require additional resources or sequencing a large number of animals.

The objective of the current study was to see if it is possible to locate the position of a novel autosomal recessive genetic condition directly using the ideas contained in Pollott (2018) on WGS methodology and using a small number of cases and controls (in this case carriers) without recourse to more extensive resources, which may not always be available. To achieve this, we investigated combining this approach with a range of other bioinformatic tools and genetic ideas which may indicate the site of such a variant by reading the various signals in the WGS data. Using the variant call format (VCF; VCF (2019)) files from a suitable combination of cases and controls, it was suggested that typically about 16 animals would need to be sequenced in order to locate a new autosomal recessive variant using a “genotype criteria” approach (GCR; Pollott (2018)).

Whole genome sequencing is becoming more widely used to locate single novel variants with major effects, and a number of approaches have been used. In a large scale analysis of Holstein cattle WGSs, seven dominant conditions were located using the genome criteria approach (Bourneuf et al., 2017) involving one case for each of the seven conditions and a control population of 1,230 animals. The trio approach has been used by a number of authors (see for example Sayyab et al., 2016). This method takes WGSs of one affected offspring and its two parents and uses the genotype criteria method to find possible sites for the causative variant. The large number of sites identified is further reduced by a range of methods. Using one dog example (Sayyab et al., 2016), a filtering pipeline was established with seven steps, including genotype criteria and SIFT analysis, Sanger sequencing

verification and sequencing of an additional 24 cases/controls. Runs of homozygosity methods have been widely used with SNP-chip data (Pollott, 2018) and Letko et al. (2020) report an example of using this method in Zwartbles sheep to locate a novel autosomal recessive condition associated with type 1 primary hyperoxaluria. Their study relied upon additional data from both the Sheep genomes project and 79 publicly available genomes of various breeds to provide “control” data for the GCR method. The methods reviewed here all required further data and analyses in order to locate the novel causative variant. In this article, we have tried to minimize this by looking for complementary methods, which can be used on the dataset alone.

Here, we test these ideas on a novel genetic disease found in Irish Moiled cattle. The Irish Moiled is an ancient hornless cattle breed native to the island of Ireland. It was popular in the 1800s, but by the late 1970s the pedigree herd numbered only 30 breeding females and two bulls. In 1979 the Rare Breeds Survival Trust recognized the Irish Moiled cattle as endangered and placed the breed on its “critical” list (Irish Moiled Cattle Society, 2020). The population size now numbers about 875 females and 90 bulls. Fortunately, novel fatal genetic diseases are relatively rare but when they do occur it is important to find the cause and implement plans to manage the condition *via* selective breeding, as soon as possible. A number of Irish Moiled cattle breeders were concerned about the seemingly high occurrence of early calf deaths in their herds. Affected calves had the following characteristics: days 1–2, the calf appeared slightly hyperactive with more playing/skipping than normal, and may have been seen drinking water from troughs or puddles, and also urinating more frequently than normal. Days 2–4, the calf started to deteriorate and became dull and lethargic. Days 4–6, the calf became dehydrated, very weak, and unable to stand. Death followed soon after. A 2–4 day-old calf was clinically very similar to a calf with septicemia; however, in contrast to septicemic calves, these calves did not respond to antibiotics and fluids given *via* an intravenous drip. Also when the blood glucose level of such calves was tested levels exceeded 30 mmol/L (Normal = 8 mmol/L). Breeders referred to this condition as “diabetes.”

An initial analysis was undertaken which suggested that there was likely a genetic basis to the disease (see Supplementary File Page S18), and since it was fatal, it could only be inherited as a recessive condition.

Materials and methods

Throughout this article the ARS-UCD1.2 (GCF_002263795.1) build of the cattle genome was used and all chromosome positions quoted relate to it (Ensembl, 2020).

Sample collection

Farmers were contacted *via* the Irish Moiled breed society and asked to submit hair samples for analysis. Clean tufts of tail hairs were plucked from either live cows/bulls or recently dead calves (within 8 h postmortem) by the animals' owners. These were sealed in a paper envelope for posting to the laboratory. Information recorded included the sample type (bull, dam or dead calf); ear tag numbers of dam and sire; dam herd ID; calving date; calf sex; and time of calf death (stillborn/days after calving). DNA was then extracted from the hair follicles for processing as described in the [Supplementary File](#) (Page S2).

Whole genome sequencing

Nine DNA samples were used in these analyses comprising three dead calves (cases) and six carriers (controls), which were either parents of the cases or parents of other dead calves (and grandparents of the cases) as illustrated in the pedigree ([Supplementary Figure S3](#)). These nine samples were sequenced on an Illumina NGS platform after sample preparation as described in [Supplementary File](#) (Page S2). Briefly, 1 µg of DNA per sample was processed using a TruSeq Nano DNA LT Library Prep Kit (Illumina, United States), according to the supplied protocol. This produced randomly sheared 350 bp inserts. After end repair and adapter ligation, DNA was amplified *via* polymerase chain reaction (PCR), and the product was purified using AMPure XP (Beckman Coulter, United Kingdom). Size-selected DNA from each animal was sequenced on the HiSeq machine to achieve 150 bp paired-end reads to cover the bovine genome with an average 30X coverage (>90 Gbp raw data with >85% Q30 (Phred-scaled)). Alignment, mapping, variant calling, and preparation of the final VCF file were carried out on the subsequent reads as described in [Supplementary File](#) (Page S2).

Genotype criteria

Information derived from WGS data on a small sample of cases and parental carriers may contain a number of signals indicating the site of a novel autosomal recessive condition. “Across”-animal data should show a typical pattern of homozygous cases and heterozygous parental carriers at the

candidate site; the “genotype criteria” approach (Pollott, 2018). Considering a single base position with a reference allele A for the given species and a new variant C which causes a novel autosomal recessive genetic disease, then the expected outcomes from matings between carriers in the population will be offspring with the genotypes AA, AC, and CC in the classical Mendelian ratio of 1:2:1. Lethals (CC) would be observed in the dataset if the effect of the new homozygous variant occurred after the time of recording the animal's health status. The term “genotype criteria” (GCR) was used to mean the particular combination of case and control genotypes, which was required to indicate that a base position could harbor the novel lethal variant (Pollott, 2018). For example, in a dataset comprising five cases and 10 parental carriers, we would expect to find the novel lethal variant at a position showing CC genotypes in all five cases and a genotype containing the C allele in the 10 parental carriers, or all AC in the case of a biallelic position. The probability of occurrence of GCR under this condition would be $1/3^n$ in cases and $1/3^m$ in parental carriers, where n is the number of cases and m is the number of carriers that have been whole genome-sequenced (Pollott, 2018). If the VCF data file comprised 14 million positions (~0.005 of the cattle genome), then a minimum of 15 animals would probably need to be genotyped in order to find one position with the required genotype criteria, that is, $1/3^{15} \times 14 \text{ million} = 0.98$ (i.e. ~1), the expected number of sites with the “correct” genotype criteria from the genome of 15 animals.

A script was written in Perl 5.28 to scan the final VCF file (containing all nine WGS animals) for the expected GCR pattern across cases and controls (i.e. all cases homozygous for the same allele and all controls heterozygous and containing this allele). In order to qualify for selection, a site had to have all genotypes with a Phred-scaled quality score greater than 12 and a depth of coverage more than 11 reads (Broad Institute, 2020). The identity of these sites was stored along with their relevant VCF record for later scanning and use.

Autozygosity-by-difference and runs of homozygosity

“Within”-animal data should show long ROH around the new variant in cases, which are not present in controls. Variants causing a novel autosomal recessive genetic disease are expected to carry with them a very long haplotype originating from the animal in which the variant first arose. When a new case animal is formed then it contains two copies of this long haplotype, only broken up by any recombination events that have occurred since the formation of the original variant, and the new variant will be situated in a long ROH. This idea has been the basis for locating novel variants using SNP-chips for a number of years (see Pollott (2018) for a review) and can be used in a number of ways with VCF data. Long ROH throughout the genome could be found

and one would expect to find the novel variant in the longest ROH in cases, possibly with adjustment for the situation in controls.

The autozygosity-by-difference (ABD) method measures runs of homozygosity on each chromosome of each individual in the dataset, cases, and controls, using genomewide single-nucleotide variant (SNV) (or SNP) genotypes (Pollott, 2018). Mean ROH length, in Kb, at each SNV position is calculated for cases and controls separately and then their difference calculated as the ABD score. Missing genotypes were assumed to be homozygous reference allele. The likely site of the new variant is in the region with the greatest mean ROH in cases, after taking into account any breed-specific ROH found in controls, that is, the ABD score. The ABD method was programmed in Perl 5.28. The final VCF file was used as the basis to generate a file of sites as input to the ABD method. This method is sensitive to incorrectly called genotypes and so the VCF file was subjected to hard filtering as recommended by the Broad Institute (2020) in the absence of suitable databases to use for the recommended variant quality score recalibration (VQSR). A file of SNV were generated from the final VCF file, which passed the quality control tests shown in [Supplementary Table S1](#), following a summary of the quality statistics of the VCF file (see [Supplementary Figures S1, S2](#)).

The ABD scores were used to look for the potential site of the novel variant causing calf mortality in the Irish Moiled dataset. The probability of each ABD score was tested using 100 permutations of the data based on the random allocation of animals to phenotypes and recalculation of the ABD scores (Pollott, 2018). Significance at the $p < 0.01$ level was considered as an indicator of a possible site of the new variant.

Sorting intolerant from tolerant (SIFT) score

In a fatal genetic disease one would expect to find that the products from any change in the sequence would have a drastic effect on the phenotype of the animal so one could not only search for SNV with a potentially drastic effect but could also eliminate those with “silent” changes. The Variant Effect Predictor (VEP; McLaren et al., 2016) is a bioinformatic tool, which can take a change in a base at a given position on the genome and predict the outcome of that change on the corresponding coding or non-coding genomic feature. Using this method it is possible to model each base position in the VCF file to see the effect of the new variant on the phenotype in the form of a SIFT score (Ng and Henikoff, 2002).

The final VCF file from all the animals was annotated for variant effect prediction using Ensembl VEP command line v90.5 (McLaren et al., 2016), given the following flags: `--tab --fork 8 --offline --species bos_taurus_merged`. The VEP was used on all variant sites in the merged VCF file, and the results filtered for

HIGH SIFT scores using VCF (Danecek et al., 2011). Various outcomes are given in the VEP but here the HIGH outcome was used for the SIFT score since this was a lethal variant. SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. The variant impact categories are subjective agreements between the VEP and SNPEff databases. However, high-impact variants are considered to have protein level disruption or change, while modifier or moderate variants impact non-coding regions of the genome. The SIFT score closer to zero is mostly represented by HIGH or modifier impact categories, while tolerated levels (SIFT score of 0.05–1) would show “minimal” to “no consequence” for the function of the genes under said variants. All sites with high-impact scores were captured in a separate file for further processing.

Novel variants

The dbSNP database of NCBI (NCBI, 2019) contains data on variants already reported by researchers. Novel variants are unlikely to be contained in these datasets and so will not already have an RS number. Their absence may be another way to reduce the search area along the genome, as novel variants are likely to be found at sites that are not already logged in the relevant SNP database. Sites in the final VCF file that did not have an RS number were possible positions for the new variant. The VCF file was scanned for positions that did not have a previously allocated RS number, using a script written in Perl 5.28. Such sites were output for further analysis. The sites identified in this way were summarized by the number of genotypes containing the variant allele found at each site. Candidate sites contained a “potential” variant allele in all nine samples.

Comparing datasets

Putting all these ideas together should make locating the site of the new variant on the genome possible using WGS data from a small number of animals without the need for any other data sources. The aforementioned analyses resulted in four independently derived sets of data, each of which could contain an indication of where the new variant might be found on the genome. These were 1) the sites with the appropriate GCR, 2) sites in long ROH, with a high ABD score significant at $p < 0.01$, 3) sites with a high-impact SIFT score, and 4) sites with no RS number. Each dataset was derived from the final VCF file by a method independent of the other three. If a site appears in all four datasets this is likely to be the site of the new variant. The four datasets were compared for overlapping positions by reading them into an Access database and linking on the site position.

Predicting the effect of the novel variant

The SIFT scoring method described earlier is one method for modeling the effects of a new variant on the phenotype of the animal. PHYRE2 (Kelley et al., 2015) is an alternative approach, which searches for homologous sequences in a database of known proteins. The reference sequence and its equivalent using the new variant were entered into the PHYRE2 database to see the effect of the site of the proposed new variant on amino acid sequence and protein structure.

Methods used to confirm the likely base position of the variant

Two independent methods were used in an attempt to confirm the site derived from the methods used on the WGS data described earlier. A sample of Irish Moiled animals comprising three bulls, 42 cows, and 18 dead calves (male and female) were genotyped using SNP-chips. The DNA from these animals were then analyzed using 1) genotyping by PCR at the suggested site and 2) the ABD method on the SNP-chip data.

SNP processing

DNA samples extracted as described in the [Supplementary File](#) (Page S2) were genotyped in the Department of Pathology, University of Cambridge (United Kingdom) and Gen-Probe (Heron House, Oaks Business Park, Crewe Road, Wythenshawe, Manchester, M23 9HZ, United Kingdom) using either 1) the Illumina BovineSNP50 BeadChip (Version 1, Illumina Inc., San Diego, CA, United States) (50k SNP, $n = 17$); 2) the BovineHD Genotyping BeadChip (777k SNP, $n = 68$), or 3) both chips ($n = 7$).

The SNP genotypes were prepared for all subsequent analyses using PLINK 1.9 (Purcell et al., 2007; Chang et al., 2015; PLINK, 2017). Quality control parameters were used to edit the data. This involved setting a lower limit on both sample and SNP quality at a call rate greater than 90%, and SNPs were retained in the dataset if they were in Hardy–Weinberg equilibrium. This was determined using Fisher's Exact Test, with a probability threshold of 0.05 and using the mid-p adjustment described in Graffelman and Moreno, 2013. The latest SNP positions were updated to ARS-UCD1.2 (GCF_002263795.1) build of the bovine genome using SNPchiMp (Nicolazzi et al., 2014; Nicolazzi et al., 2015). In addition, a merged set of data was produced using SNPchiMp, combining all genotyped animals from both the LD and HD datasets with common SNP. This merged dataset was then used in KING (Manichaikul et al., 2010) to generate relatedness coefficients between all genotyped animals, based on whole genome SNP genotypes and to aid pedigree checking. Any animal whose pedigree did not match the relatedness information from the SNP data was discarded. In all

13 animals were discarded for both pedigree and quality control reasons. Because nine animals were also used for the WGS analysis, they too were excluded from the SNP ABD analysis, in order to produce a dataset of independent animals.

Autozygosity-by-difference

The ABD method (Pollott, 2018), described earlier, was used on the merged SNP-chip dataset. The probability of each SNP ABD score (difference between mean ROH length (Kb) from cases and controls at each SNP position) was tested using 1,000 permutations of the dataset based on random allocation of animals to phenotypes and recalculation of the ABD scores. Significance at the $p < 0.001$ level was considered as an indicator of a possible site of the new variant.

Genotyping by PCR analysis

Primers (5'-CATGAACCCAGTGTACAGC-3' and 5'-CTCTCCGTGGAAGAGCAGAT-3') were designed using Primer3 (version 4.1.0; <http://primer3.ut.ee>) to amplify a 218 bp product spanning the identified variant locus. The primer design was based on the published sequence for the *Bos taurus* (UMD3.1; GCF_000003055.6) glucokinase gene ENSBTAG00000032288. Exon/intron boundaries were derived from this in combination with mRNA RefSeq NM_001102302. PCR was performed using AmpliTaq Gold polymerase (Applied Biosystems), according to the manufacturer's protocol. Products were purified using the QIAquick PCR purification kit (Qiagen) and sequenced by Sanger sequencing using the forward and reverse primers. Sequence analysis was carried out in CLC Genomics Workbench (Qiagen, 2013). The candidate site was updated to the ARS-UCD1.2 (GCF_002263795.1) genome build using the UCSC Genome LiftOver facility (UCSC, 2020).

Results

The WGS data from all nine animals, three cases, and six carriers, resulted in a final VCF file comprising 8,234,367 biallelic autosomal single-nucleotide variants, which were to be used for all subsequent WGS analyses. These are summarized by chromosome in [Table 1](#) along with the length of each chromosome aligned in ARS-UCD1.2 (GCF_002263795.1) build of the cattle genome.

Genotype criteria

Searching the final VCF file for sites with the appropriate genotype criteria (all homozygous cases for the same genotype and all carriers heterozygous containing one allele forming the homozygote in cases) resulted in the identification of 730 sites. These are shown broken down by chromosome in [Table 1](#). Applying the formula, $1/3^n$ cases and $1/3^m$ parental carriers to

TABLE 1 Summary of results by chromosome.

BTA	Length (bp)	Number of SNV in VCF file	Number of indels in VCF file	Estimated number of GCR	Actual number of GCR	Number of SIFT sites	Number of NoRS9 sites
1	158,534,110	535,135	101,947	32	582	251	694
2	136,231,102	452,534	82,255	27	4	279	165
3	121,005,158	394,128	70,162	24	0	383	271
4	120,000,601	373,759	72,506	23	22	485	347
5	120,089,316	401,056	71,836	24	2	458	173
6	117,806,340	362,930	71,311	22	4	174	221
7	110,682,743	358,947	67,655	22	2	491	250
8	113,319,770	319,261	62,634	19	3	204	486
9	105,454,467	328,559	62,502	20	2	179	185
10	103,308,737	353,791	63,775	21	1	293	162
11	106,982,474	324,386	58,178	19	9	337	134
12	87,216,183	335,302	63,750	20	21	139	148
13	83,472,345	232,616	43,728	14	2	249	188
14	82,403,003	262,990	49,112	16	1	108	135
15	85,007,780	285,748	54,791	17	2	374	160
16	81,013,979	268,622	49,356	16	0	225	388
17	73,167,244	278,406	50,191	17	5	198	108
18	65,820,629	194,623	37,810	12	0	522	384
19	63,449,741	227,124	39,664	14	0	449	110
20	71,974,595	230,092	43,456	14	1	110	84
21	69,862,954	202,249	38,102	12	2	205	261
22	60,773,035	180,922	33,961	11	1	147	164
23	52,498,615	255,467	42,220	15	58	566	337
24	62,317,253	228,628	39,037	14	3	83	109
25	42,350,435	142,293	25,212	9	0	225	43
26	51,992,305	177,147	33,190	11	1	133	120
27	45,612,108	175,471	32,421	11	1	99	88
28	45,940,150	179,521	31,214	11	1	92	99
29	51,098,607	172,660	31,952	10	0	306	358
Total	2,489,385,779	8,234,367	1,523,928	496	730	7,764	6,372

BTA, chromosome number; GCR, genotype criteria sites; NoRS9, number of sites with no RS number and with at least one alternate allele in all nine genotypes. "Estimated number of GCR sites" assumes an even spread across the genome. "Number of SIFT sites" was the number of sites with a "HIGH" SIFT score.

over 9 million SNV and indels, we would expect to find ~496 sites fitting the genotype criteria. There were clearly more GCR sites than expected in this set of animals. Chromosomes 1 and 23 appeared to have more sites than expected (Table 1).

Autozygosity-by-difference method

In order to run the ABD method on the final VCF file, the hard-filtering criteria shown in Supplementary Table S1 were used on the extracted biallelic SNVs for the dataset. This resulted in a file of 629,716 SNV for the ABD analysis. ABD software was used to generate the Manhattan plots shown in Figure 1 and Supplementary Figure S4. The two plots in S4 show the mean

ROH length at each of the base positions in the VCF file for cases and carriers, respectively, while Figure 1, the ABD score, shows the difference between them. Long ROH were found on BTA4 and BTA18. Probabilities for the ABD scores were generated from 100 permutations of the dataset, and the regions of the genome with $p < 0.01$ are summarized in Supplementary Table S2. The 0.01 probability level was computed to be at an ABD score of 9,034 Kb. Supplementary Table S2 shows that the length of BTA4 above the 0.01 probability threshold was 7.804 Mb. The highest mean ROH length in cases was 14.197 Mb so the long ROH found on BTA4 continued on either side of the significant region. Similarly on BTA18, the highest mean ROH score in cases was 22.4 Mb long.

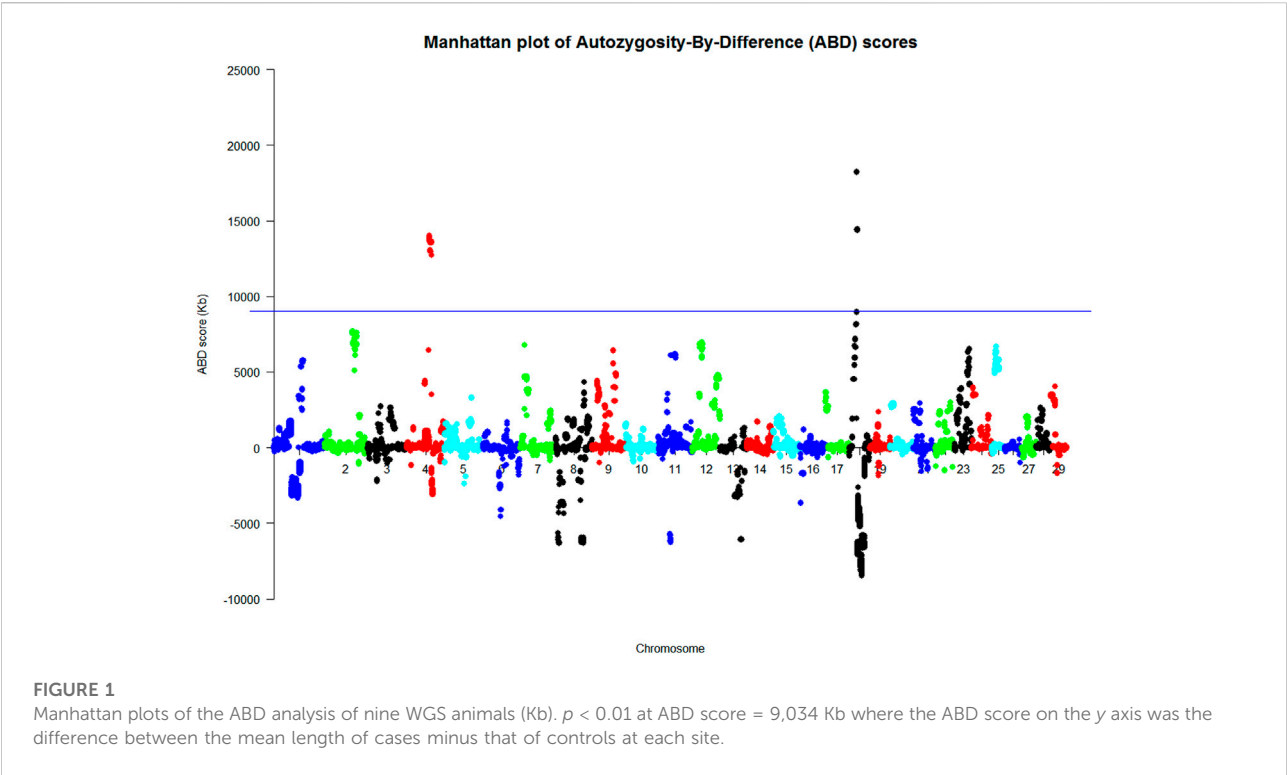


TABLE 2 Overlap between the four methods for locating a likely novel variant site.

Method	Genotype criteria (GCR)	Autozygosity-by-difference (ABD)	High-impact SIFT score (SIFT)	No registered SNP number (NoRS9)
GCR	730 (22)			
ABD	22 (10)	896 (635)		
SIFT	1 (1)	12 (12)	7,764 (12)	
NoRS9	78 (8)	41 (40)	8 (1)	6,372 (40)
GCR+ABD			1 (1)	8 (8)
GCR+SIFT				1 (1)
ABD+SIFT				1 (1)
GCR+ABD+SIFT				1 (1)

The table shows the number of sites in the final VCF file identified by each method (numbers in the BTA4 high-ABD region shown in parentheses).

SIFT score

Using the VEP to estimate the effect of each of the SNVs and indels in the final VCF file resulted in 65,961 records of HIGH impact SNVs located at 7,764 different autosomal positions. The distribution of these sites is summarized by chromosome in Table 1.

Sites with no RS number

The VCF file contained 340,893 sites with no RS number. Only 11% (6,372) of the sites had genotypes, other than the

homozygous reference genome, in all nine cases and carriers (NoRS9). The breakdown of these by chromosome is summarized in Table 1. These are likely to contain the novel variant.

Overlap of GCR, ABD, NoRS9, and SIFT results

So far, four possible datasets were generated that might contain the site of the novel variant causing this new autosomal recessive condition. The overlap between the

TABLE 3 Animal status by genotype for the 41 Sanger-sequenced animals at Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)).

Animal status	AA	AT	TT	Total
Calves	4	2	7	13
Known adult carriers (live)	0	6	0	6
Status unknown adults (live)	13	9	0	22
Total	17	17	7	41

four datasets is summarized in [Table 2](#). The genotype criteria method resulted in the fewest sites identified (730), with the other methods increasing in the order autozygosity-by-difference, no RS number with nine genotypes, and high SIFT score. Combining the GCR method with each of the others in turn allowed the identification of 22 (ABD), 1 (HIGH SIFT), and 78 (NoRS9) sites in common. One site appeared in all four datasets located at position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)). We may tentatively conclude that this is the site of the novel variant, causing early calf death in the Irish moiled breed.

PHYRE2 prediction

The PHYRE2 prediction ([Kelley et al., 2015](#)) of secondary structures between the reference genome and new variant *GCK* model at the beginning of exon 8, which contains the possible site of the new variant, Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)), predicted the amino acid sequence NPGQQLWY from the reference genome being changed to NPGQQLLY with the new variant.

Independent confirmation of the results

Sanger sequencing

A sample of 41 animals was Sanger sequenced at position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)), following PCR of the region surrounding this site. [Table 3](#) shows the Fisher's exact test ([Fisher, 1922](#)) results ([Freeman and Halton, \(1951\)](#)) for animals falling into three categories; calves, carriers, and live animals of the unknown status by three genotypes.

[Table 3](#) shows that all TT animals were calves, all of which died of the symptoms described earlier. All live animals were either AA or AT. The overall results were significant with a probability = 1.868×10^{-5} (0.00001868) for this 3×3 table arising by chance, thus indicating a likely association between genotype and health status at this site.

ABD method based on genotypes derived from the PCR analysis

The merged dataset (HD and LD chip data merged using SNPchiMp) comprised 63 animals (19 cases and 44 controls) and 42,453 SNPs after quality control conditions were met. The 42 animals used in the PCR analysis were selected for the ABD analysis, which excluded the WGS animals so that this analysis was independent of the WGS ABD analysis. The animals were allocated to their case/control status based on their PCR genotype at the highlighted location. The results of the ABD analyses are summarized in [Figure 2](#) and [Supplementary Figure S5](#) and showed a 20.8 Mb length of BTA4 with a permuted probability <0.001, from 1,000 permutations, equivalent to an ABD score greater than 7,023 Kb. This region was from position Chr4: g.62872037 to g.83635054 (ARS-UCD1.2 (GCF_002263795.1)), which includes the site highlighted as the putative causal variant from the WGS analyses.

The results in [Supplementary Figure S5](#) show a long ROH on BTA21 but this was present in both the case and control animals, which negated each other in the ABD score analysis. This is a good example of the benefit of the ABD method. Also, the long ROH found in WGS cases on BTA18 ([Figure 1](#)) was not a feature of this larger set of results. There was reduced variability of these results with a higher number of animals compared to those from the WGS dataset analysis with only nine animals.

Discussion

This work had two objectives. One general and the other more specific. The generally applied objective was to test the idea that it is possible to find the site of a novel autosomal recessive variant using just a small number of whole genome-sequenced animals and appropriate bioinformatic methods, thus circumventing the need for the commonly-used two-stage approach highlighted by the review of [Pollott \(2018\)](#) or the collection and/or use of further data. The specific objective was to find the site of a new autosomal recessive condition thought to exist in Irish Moiled cattle.

Bioinformatic methods used with WGS data to find the site of a new autosomal recessive variant using a small number of cases and controls

In the current study, four bioinformatic methods were tested to try to find the location of the new variant causing early calf death in the nine Irish Moiled animals and relied on the "correct" site appearing in all four methods. No additional data from the Irish Moiled or any other breeds were used. Two of the methods

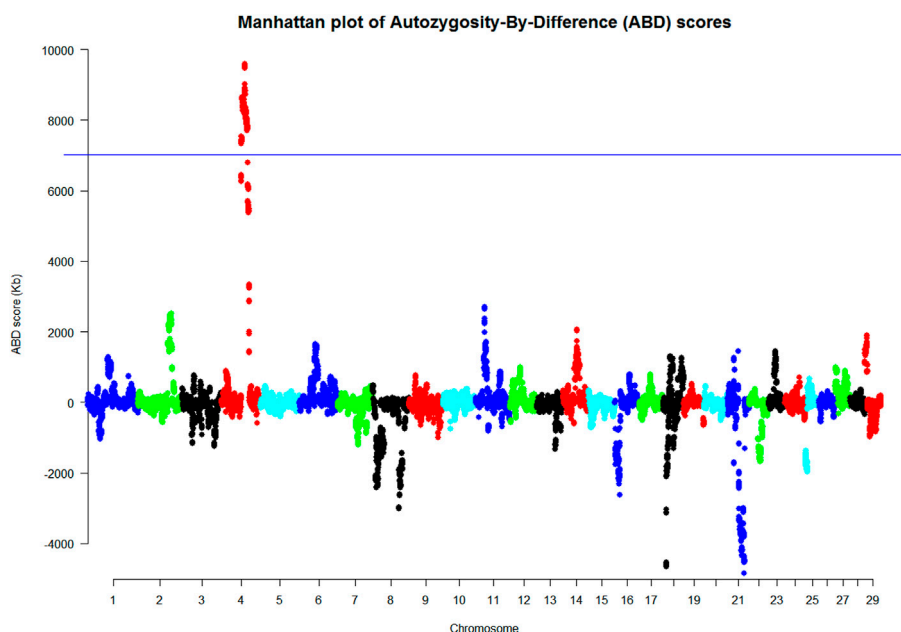


FIGURE 2

Manhattan plot of the ABD analysis of the SNP-chip analysis based on the genotypes found in the PCR analysis (Kb). These results were based on animals with phenotyping informed by the PCR results ($p < 0.001$ at ABD score = 7,023 Kb where the ABD score on the y axis was the difference between the mean length of cases minus that of controls at each site).

(GCR and ABD) do not require any prior information about genes, SNP, or other genomic features but rely on across- and within-animal patterns of information contained in the genotypes at each SNV/indel found in the VCF file. Ideally one would like to use these two methods alone since they are not only independent of any prior knowledge about genome features (except the reference genome for the alignments and generation of the VCF file) but they will also be able to find a new variant causing an autosomal recessive condition anywhere on the genome, even when located outside a protein-coding region: a useful feature of the two methods. As has been seen, using three cases and six controls with the ABD method and GCR combined revealed 10 possible sites in a 7.795 Gb stretch of Chr4 between g.70889821 and g.78684588 (ARS-UCD1.2 (GCF_002263795.1)), involving 18,705 SNVs. The underlying implication of this approach is that with more animals, either cases or controls, we would find fewer sites and so make the search considerably more straightforward and find just a single causative variant site.

It is worth noting that when using WGS methods, the depth of coverage can have an important effect on the results and is related to the costs of genotyping. In this article, a 30X coverage of the genome was used. In a study looking at the relationship between depth of coverage and the ability of an experiment to locate novel variants, Jiang et al. (2019) found 10X to be an ideal balance between cost and accuracy. Not surprisingly, the higher

the depth of coverage, the greater is the ability to discover novel variants.

The genotype criteria approach

The method used to find the sites meeting the genotype criteria was based on a number of implied assumptions not stated by Pollott (2018). First, GCR sites would be evenly distributed across the genome. Second, the higher the number of animals used the greater the chances of finding the GCR site of the new variant. Third, a GCR site was not dependent on the balance of cases and controls in the samples. Fourth, the location of a single GCR site was independent of the genetic relationship between cases and controls. Each assumption was tested using the data analyzed in this study, either the final VCF file for BTA4 or the SNP-chip data with phenotypes allocated by the PCR results as appropriate. The detail of these investigations is given in the [Supplementary File](#) (Pages S10–S17).

Evenly spaced GCR sites across the genome

The results in Table 1 show that some chromosomes contain no GCR candidate sites at all (BTAs 3, 16, 18, 19, 25, and 29). Many chromosomes contained far fewer GCR sites than expected whereas others contained a much greater number than expected (BTAs 1 and 23). A GCR site (in this case) comprises two components; the 0/1 in all controls and the 1/1 in all cases (using 0 to mean the reference allele and 1 the new or alternative

variant). We might expect the chromosomes containing long ROH in cases potentially to have many more GCR sites than others. Inspection of [Supplementary Figure S4](#) shows BTA4 and BTA18 as having the longest mean ROH but only BTA1 had a large excess of GCR sites. Using the information shown in [Supplementary File](#), the original implied assumption from [Pollott \(2018\)](#) of an even distribution of GCR sites across the genome has not been verified here, and this has implications for the number of animals required to find a new variant site using WGS data alone. Clearly, the long ROH on BTA4 were linked to a large number of GCR sites but that was not so on BTA18.

Finding a causative variant

The second assumption about the use of GCR sites to locate a novel autosomal recessive variant was that the greater the number of animals that are genotyped, the better the chance of locating the new variant. It appears, from the work reported in [Supplementary File](#), that the number of genotyped animals required to find the single candidate sites is three to four times greater than the predicted original formula of [Pollott \(2018\)](#); one case and 29 controls appears to require the fewest total animals genotyped to find the single candidate site in the SNP-chip dataset used. However, this “long tail” is due to several GCR sites being close together around the candidate site and always being “found” in the generated datasets. Differentiating between them may require another method or using a different set of controls, perhaps from more distantly related individuals.

The balance of cases and controls

The basic calculation of the number of animals required to find a GCR site is independent of the balance between the number of cases and controls used. [Supplementary Figure S10](#) demonstrates that the lower the number of cases used, the fewer the total number of animals required to be genotyped. At first sight, these are rather startling results. However, outside the candidate site, it is much more unlikely to find all controls with a heterozygote genotype, whereas there will be many sites with all homozygous genotypes in cases; after all long ROH imply many 1/1 genotypes and so more sites potentially could meet the genotype criteria. The information in the [Supplementary File](#) (pages S10–S16) clearly demonstrates that the number of animals required is much closer to the theoretical numbers when large ROH regions are excluded from the analyses. The minimum number of animals required to find the candidate site is still slightly greater than the theoretical figure but this may be due to some other small areas of ROH not removed from the dataset. There are an enduring number of GCR sites in the high-ROH regions, which inflate the results in contrast to the theoretical number of sites expected. This illustrates why theory and “practice” may differ.

The genetic relationship between cases and controls

The nine animals used in the WGS analysis comprised three cases and six parental controls. In order to investigate whether

the closely related controls might inflate the number of GCR sites found, an alternative SNP-chip dataset was derived using controls that were most distantly related to the cases ([Supplementary File](#) and [Supplementary Figure S12](#)). In this case, the number of animals required to find the new variant site was much closer to the theoretical expectation than with the parental controls.

The autozygosity-by-difference method

The ABD method was developed originally to locate regions of the genome likely to contain a new autosomal recessive variant using SNP data (*see for example* [Posbergh et al., 2018](#)). In the current analysis, it has been applied to WGS data from animals for the first time, as well as being used to confirm the results in an independent sample of SNP-genotyped animals. Because the method is sensitive to incorrectly called genotypes, a feature of WGS data, it was necessary to employ hard filtering criteria (*see* [Supplementary Table S1](#)) of both sites and genotypes in order to get a useable set of data. In this case, the VCF file was reduced from ~12 million to ~630,000 sites but this would differ under alternative hard-filtering criteria. Since VQSR methods were not available in the current situation (due to the species and number of animals genotyped) an alternative approach was taken; selecting sites and genotypes to fall within ± 2 s.d. of the mean (or peaks in the case of bimodal variables). This allowed the location of several long ROH, one of which was found, by additional methods, to contain the new variant.

As well as using alternative hard-filtering criteria, it may be possible to use other approaches, including site sampling or sliding windows, to locate the region containing the new autosomal recessive variant. Using a site-sampling approach with a VCF file one could randomly select, say, 5–10% of sites evenly spread across the genome with the ABD method. Repeated samples of these SNVs, (say 100), could be randomly drawn and the 100 sets of ABD results averaged at each site. Alternatively, one could use a sliding window of, say, 10,000 base positions and count the number of homozygous variant case and control genotypes in each window. The window would then be moved along the genome at a given interval, say every 1,000 base positions, and the results plotted. One would expect to find the new variant causing the autosomal recessive condition in the region with the highest ABD-type score. Both these methods would overcome the problem of a single incorrectly called genotype disrupting the long ROH in the ABD results and the need for hard filtering.

Using ABD on the hard-filtered WGS data resulted in the identification of two regions of the genome having an ABD score above the 0.01 probability threshold, and therefore likely to contain the new variant ([Figure 1](#)). Two aspects of [Figure 2](#) and [Supplementary Figure S4](#) are of note. First, there were a number of long ROH found in the controls throughout the genome with BTA18 having the largest mean ROH length. This was also found in the cases, but the effect of combining

the two sets of data in the ABD score was to remove many of these “breed-specific” ROH and leave those which probably harbored the new variant. This is one of the advantages of the ABD method, particularly in rare breeds, but it has also been shown to be effective in removing long breed-specific ROH in studies of new variants in the myostatin gene in both Texel sheep (Pollott, 2013) and Piedmontese cattle (Biscarini et al., 2013).

The ABD method was also used for another purpose in this work; to confirm the WGS results on an independent set of animals with SNP-chip-derived data (Figure 2). As with the WGS ABD results, there were many long ROH in the SNP-chip dataset in both cases and controls. In this instance, there were two very long ROH on BTA4 and BTA21, but interestingly not BTA18 as was found in the WGS data. The smaller number of animals used in the WGS analysis probably resulted in BTA18 having a long ROH due to sampling of closely related individuals. Supplementary Figure S5 also shows BTA18 to have a long ROH but it was less pronounced in this larger dataset. In Figure 2, the result of subtracting the control ROHs from that of cases at each site reduced the noise considerably and left BTA4 as the only significant peak by a considerable margin. Once again it has been demonstrated that the power of the ABD method to remove noise works effectively to highlight the region containing the potential causal variant.

GCR, ABD, SIFT score, and RS number

The approach in this work has been to use several bioinformatic methods on WGS data to see if they can pinpoint the site of a causal variant of a new autosomal recessive condition. Table 2 has highlighted a significant region on BTA4 using the ABD method that contained 12 sites meeting the genotype criteria. The earlier discussion has suggested that using more animals may have reduced the number of candidate sites by a small amount, but a greater use of unrelated controls may have reduced the number of GCR sites in the target area more effectively.

In this set of results, the SIFT scores were the crucial factor in determining the suggested site of the new variant. Position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1) was the only one of the 22 GCR sites in the target area to have a high-impact SIFT score (0–0.05). However, only sites in or near a coding region are scored using the SIFT method so it is not always going to find the causative site if it is located outside these regions of the genome.

The use of the absence of an RS number could be useful but, in this case, did not prove to be the final factor locating the novel variant site.

Other types of inheritance and effects

This article reports the search for a new autosomal recessive variant causing a fatal condition in calves using a range of

bioinformatic methods. It raises questions relating to whether the methods would work with autosomal recessive conditions with differing phenotypes or with other modes of inheritance.

In this example, the phenotype was mortality in early postnatal life. This had the advantage of being an obvious phenotype. As the deaths occurred on widely scattered farms it was, however, not possible in this instance to collect suitable pre- or postmortem samples for follow-up analyses. A new non-fatal condition may be less easy to identify initially but there are likely to be more opportunities to collect appropriate samples to confirm phenotype diagnosis.

Having suggested this WGS approach for finding a new autosomal recessive variant, the question arises about its general usefulness with variants involving other modes of inheritance. A dominance mode of inheritance can be thought of as the reverse of the recessive mode. One would expect to find cases to be 0/1 or 1/1 and controls to be 0/0 so the genotype criteria would be different compared to the recessive mode of inheritance. However, the number of animals required to use the GCR method would be very similar with cases being $2/(3^n)$ and controls $1/(3^m)$; the numerator having little effect with such a large denominator. The ABD method could only be used if all cases were 1/1 but that is unlikely with a dominant condition due to the large number of heterozygotes likely to be in the population. Alternatively, if there was some way to phenotypically distinguish 1/1 from 1/0 cases this would be useful. The 0/0 controls are unlikely to be situated in long ROH since they are likely to have been subjected to many generations of recombination, so alternative methods may be required. The new dominant variant would be situated in a long haplotype so it may be possible to adapt haplotype discovery methods to this situation. Both the SIFT score and RS number methods would be applicable but they are less powerful than the other two because they rely on previous knowledge and, in the case of SIFT, it only works for a limited distance around a protein-coding region.

These methods could be used for a recessive sex-linked new variant, that is, one found on the X chromosome. Males would provide no useful data in this case so only females would be required. Both the ABD and the GCR methods would work the same way but with a lot fewer sites to search (only the X chromosome data would be needed).

Finding the causative variant for a perinatal mortality syndrome in Irish Moiled cattle

The likely site for the causative variant of this fatal perinatal condition in Irish Moiled animals has been successfully located using just six parental control animals and three cases. Perinatal mortality (within 24 h of birth) typically occurs in about 6–10% of calves born (Brickell et al., 2009), with a further 3–4% dying in their first month, mainly from infectious disease (Johnson et al.,

2017). The site highlighted at Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1) was located in the glucokinase gene (*GCK*) and is a splice acceptor variant. Analysis of the OMIA website (Nicholas and Hobbs, 2013; OMIA, 2020) showed splice acceptor variants to be responsible for ~8% of known variants in non-laboratory animals. There was a clear difference in the PHYRE2 prediction of the secondary structures between the reference genome and new variant *GCK* model. As observed from the SIFT score results, this is expected to have a disruptive effect on the operation of the *GCK* gene.

Glucokinase is a key enzyme found in the liver, pancreas, brain, and endocrine cells of the gut. It catalyzes the starting point of glycolysis by phosphorylating glucose to form glucose-6-phosphate (Matschinsky et al., 1993). The crystal structure has revealed that glucose binds in a deep cleft between a large and small domain of *GCK*, resulting in a conformational change and enzyme activation (Kamata et al., 2004). Glucokinase stimulates glucose uptake, glycolysis and glycogen synthesis by hepatocytes, whereas in pancreatic β -cells it plays a crucial role in glucose-stimulated insulin secretion. Glucose homeostasis is essential in mammals and is under tight endocrine control, with insulin acting as the key regulator.

There are currently 922 SNPs listed within the bovine *GCK* gene (NCBI, 2019) but the closest to the new variant site flanking either side were at Chr4: g.77173441 (an intron variant) and Chr4: g.77174392 (ARS-UCD1.2 (GCF_002263795.1)), some 46 and 905 bp away, respectively. A segment of the ARS-UCD1.2 (GCF_002263795.1) genome 30 bp either side of the candidate variant was selected and BLASTn (Altschul et al., 1990) was used with the 61 bp sequence to find any homologous region on the human genome. A 40 bp length of sequence was found with 35 identical bases and a score of 50.9 bits (55) and no gaps. This was located on the reverse strand of human Chr7: g.44146590 to g.44146629 (GRCh38.p13 (GCF_000001405.39)), in the *GCK* gene. The location on the human genome equivalent to the candidate variant found in Irish Moiled calves was at Chr7: g.44146620 (GRCh38.p13 (GCF_000001405.39)). This was a highly conserved site with 96 out of the 100 vertebrate genomes shown on the UCSC (UCSC, 2020) genome browser, all having a T on the forward strand, the remaining four being not reported. No SNP was found at this site in the human database but there was an SNP reported at the adjacent position (Chr7: g.44146619; GRCh38.p13 (GCF_000001405.39)), which was cataloged as rs1167675604, a C>T change on the forward strand. This site was also highly conserved in 96 out of the 100 vertebrate genomes on the UCSC genome browser and was also a splice site acceptor variant. The ClinVar (ClinVar, 2020) record for this variant states that “the variant disrupts a canonical splice site and is therefore predicted to result in the loss of a functional protein, found in at least one symptomatic patient, and not found in general population data.” Its incidence was estimated to be well below 0.001% of the population. In addition, the Varsome (Varsome, 2020) record

for this SNP states that the effect of the variant was “Very Strong,” which means “Null variant (intronic within ± 2 of splice site) affecting gene *GCK*, which is a known mechanism of disease (gene has 378 known pathogenic variants, which is greater than minimum of 3), associated with diabetes mellitus, permanent neonatal 1, maturity onset diabetes of the young, type 2, and hyperinsulinemic hypoglycemia, familial 3.”

The mouse genome was also investigated in the same way but no SNPs were found in the candidate region.

Over 600 variants have been reported in the human *GCK* gene, which have varying effects depending on their location (Osak et al., 2009; OMIM 138079). Heterozygous inactivating variants cause a condition known as maturity onset diabetes of the young, characterized by mild fasting hyperglycaemia. Homozygotes are much rarer in the human population, and neonates present earlier with permanent neonatal diabetes mellitus. In mice, however, pups born with global *GCK* knockout (–/–) are slightly smaller than wild-type animals (+/+), have glucose levels about eight-fold higher and die within 3–5 days (Grupe et al., 1995). Tissue specific β -cell knockouts die within 4 days of birth, whereas hepatic knockout impairs glucose utilization and glycogen synthesis but with only mild hyperglycaemia (Postic et al., 1999).

Pregnancy outcome in women depends on a combination of the genotype of both mother and fetus (Spyer et al., 2001). When the fetus carries a single *GCK* variant, this affects glucose homeostasis with reduced insulin secretion, so both placental and birth weight are reduced (Hattersley et al., 1998; Spyer et al., 2008). During pregnancy, the fetal glucose supply is derived almost entirely from the dam across the placenta using facilitated diffusion by glucose transporters. In ruminants, this uptake is regulated sequentially by GLUT1 and GLUT3 (SLC2A1 and SLC2A3) (Wooding et al., 2005).

The fetus has a low capacity for endogenous glucose production but this increases in late gestation, in response to the pre-term increase in glucocorticoid production, together with catecholamine and thyroid hormone stimulation. These promote hepatic glycogen synthesis and gluconeogenesis, which are essential in providing the neonatal calf with an adequate glucose supply as milk lactose on its own is insufficient (Hammon et al., 2013). The postnatal maturation in the regulation of energy supply may thus explain why lack of *GCK* activity is fatal at this stage of life.

Concluding remarks

The original intention for this work was to locate the site of a potential novel variant, causing perinatal mortality in Irish Moiled calves. This has been achieved, and shown to be located in the *GCK* gene, but in the process it became apparent that there were no straightforward ways to achieve this objective. At best, a two-stage approach was required, involving genotyping a group of cases and

controls, identifying the genomic region likely to contain the novel variant followed by further work to sequence the identified region and look for appropriate signals in the data. Consequently, a further objective was set in order to simplify the process and investigate whether it would be possible to use a single whole genome sequencing stage with appropriate bioinformatic methodology to find the candidate site. This too has been achieved by sequencing nine animals, three cases, six parental controls, and applying four methods to the data. In the process, it has been possible to investigate some of these methods in more detail and arrive at some general conclusions to aid future such studies.

The VCF file format has proven to be a very practical source of data for this study particularly because it reduced the search “area” from over 2.5 billion base positions down to one involving only 9 million sites. In addition, the VCF file format facilitated finding the novel site when combined with methods to interrogate it for genotype criteria, long runs of homozygosity, and the predicted effects of variants on the phenotype of the animal. Using these three methods allowed the identification of a single variant site, which was found to have both the genomic and biological properties associated with this novel condition.

In the process of carrying out this work it has been possible to refine the genotype criteria method to demonstrate that in reality only a small number of cases and controls *are* required, and controls should outnumber cases by 2:1 and controls should be more distantly related to cases. In addition, it has been possible to show that using a runs-of-homozygosity method, previously used only on SNP-chip genotype data with whole genome sequence data, it was possible to locate the region of the genome containing the novel variant.

In future it should be possible to use the combination of genotype criteria and runs of homozygosity methods with the appropriate number of cases and controls, suitably distantly related, to locate the site of any new autosomal recessive genetic condition in a relatively short time. This should then facilitate a more speedy elimination of the harmful variant from the population by using an appropriate genetic test on available animals.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://rvc-repository.worktribe.com/output/1551650>.

Ethics statement

Ethical review and approval was not required for the animal study because it was based on collection of farm data. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

GP and DCW designed the study. GP analyzed the data, wrote ABD software, carried out the bioinformatic analyses, and wrote the first draft of the article. ZC prepared the samples for analysis. MS produced the alignments, VCF files, SIFT, and PHYRE2 scores. RP and CM carried out the Sanger sequencing. GP, DCW, RP, CM, and MS contributed written material to the final article.

Funding

We would like to acknowledge the Biotechnology and Biological Sciences Research Council for Stephanie Hill's funding under the Vacation Bursaries scheme in the early stages of this research. The Farm Animal Genetics and Genomics (Genesis) Faraday Partnership - LS1618 is also acknowledged for funding some of the SNP genotyping work used in the confirmatory stage of this research.

Acknowledgments

We would like to acknowledge the help given to us by Irish Moiled cattle breeders, who gave us access to their records, and to the breed society for making all the pedigree data available. Lauren Hamstead is thanked for her technical assistance with the laboratory analyses and Stephanie Hill for her farm survey of Irish Moiled data. This article is submitted as RVC manuscript number PPS_02318.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.755693/full#supplementary-material>

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2
- Biscarini, F., Corvo, M., delStella, A., Albera, A., Ferenčakovic, M., and Pollott, G. E. (2013). Looking for the mutations for arthrogryposis and macroglossia in piedmontese cattle: Preliminary results. *J. sobre Prod. Anim. Zaragoza* 14, 538–540. y 15 de mayo de 2013.
- Bourneuf, E., Otz, P., Pausch, H., Jagannathan, V., Michot, P., Grohs, C., et al. (2017). Rapid discovery of de novo deleterious mutations in cattle enhances the value of livestock as model species. *Sci. Rep.* 7, 11466. doi:10.1038/s41598-017-11523-3
- Brickell, J. S., McGowan, M. M., Pfeiffer, D. U., and Wathes, D. C. (2009). Mortality in Holstein-Friesian calves and replacement heifers, in relation to body weight and IGF-I concentration, on 19 farms in England. *Animal* 3, 1175–1182. doi:10.1017/S175173110900456X
- Broad Institute (2020). Available at <https://gatk.broadinstitute.org/hc/en-us> (Accessed July 26, 2021).
- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7–015. doi:10.1186/s13742-015-0047-8
- ClinVar (2020). Available at <https://clinvarminer.genetics.utah.edu/> (Accessed October 20, 2020).
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* 27, 2156–2158. doi:10.1093/bioinformatics/btr330
- Ensembl (2020). Available at http://oct2018.archive.ensembl.org/Bos_taurus/Info/Annotation (Accessed October 20, 2020).
- Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* 85, 87–94. doi:10.2307/2340521
- Freeman, G. H., and Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38, 141–149. doi:10.2307/2332323
- Graffelman, J., and Moreno, V. (2013). The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat. Appl. Genet. Mol. Biol.* 12, 433–448. doi:10.1515/sagmb-2012-0039
- Grupe, A., Hultgren, B., Ryan, A., Ma, Y. H., Bauer, M., and Stewart, T. A. (1995). Transgenic knockouts reveal a critical requirement for pancreatic beta cell glucokinase in maintaining glucose homeostasis. *Cell* 83, 69–78. doi:10.1016/0092-8674(95)90235-x
- Hammon, H. M., Steinhoff-Wagner, J., Flor, J., Schönhusen, U., and Metges, C. C. (2013). Lactation biology symposium: Role of colostrum and colostrum components on glucose metabolism in neonatal calves. *J. Anim. Sci.* 91, 685–695. doi:10.2527/jas.2012-5758
- Hattersley, A. T., Beards, F., Ballantyne, E., Appleton, M., Harvey, R., and Ellard, S. (1998). Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nat. Genet.* 19, 268–270. doi:10.1038/953
- Irish Moiled Cattle Society (2020). Available at <https://www.irishmoiledcattlesociety.com/breed-history/> (Accessed October 20, 2020).
- Jiang, Y., Jiang, Y., Wang, S., Zhang, W., and Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinforma.* 20, 556. doi:10.1186/s12859-019-3164-z
- Johnson, K. F., Chancellor, N., Burn, C. C., and Wathes, D. C. (2017). Prospective cohort study to assess rates of contagious disease in pre-weaned UK dairy heifers: Management practices, passive transfer of immunity and associated calf health. *Vet. Rec. Open* 4, e000226. doi:10.1136/vetreco-2017-000226
- Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J., and Nagata, Y. (2004). Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* 12, 429–438. doi:10.1016/j.str.2004.02.005
- Kelley, L., Mezulis, S., Yates, C., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi:10.1038/nprot.2015.053
- Letko, A., Dijkman, R., Strugnell, B., Häfliger, I. M., Paris, J. M., Henderson, K., et al. (2020). Deleterious AGXT missense variant associated with Type 1 primary hyperoxaluria (PH1) in Zwartbles sheep. *Genes* 11, 1147. doi:10.3390/genes1101147
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi:10.1093/bioinformatics/btq559
- Matschinsky, F., Liang, Y., Kesavan, P., Wang, L., Froguel, P., Velho, G., et al. (1993). Glucokinase as pancreatic beta cell glucose sensor and diabetes gene. *J. Clin. Invest.* 92, 2092–2098. doi:10.1172/JCI116809
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The Ensembl variant effect predictor. *Genome Biol.* 17, 122. doi:10.1186/s13059-016-0974-4
- NCBI (2019). Available at <https://www.ncbi.nlm.nih.gov/snp/> (Accessed October 20, 2020).
- Ng, P. C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12, 436–446. doi:10.1101/gr.212802
- Nicholas, F. W., and Hobbs, M. (2013). Mutation discovery for mendelian traits in non-laboratory animals: A review of achievements up to 2012. *Anim. Genet.* 45, 157–170. doi:10.1111/age.12103
- Nicolazzi, E. L., Caprera, A., Nazzicari, N., Cozzi, P., Strozzi, F., Lawley, C., et al. (2015). SNPchiMp v.3: Integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genomics* 16, 283. doi:10.1186/s12864-015-1497-1
- Nicolazzi, E. L., Picciolini, M., Strozzi, F., Schnabel, R. D., Lawley, C., Pirani, A., et al. (2014). SNPchiMp: A database to disentangle the SNPchip jungle in bovine livestock. *BMC Genomics* 15, 123. doi:10.1186/1471-2164-15-123
- OMIA (2020). *Online mendelian inheritance in animals, OMIA*. The University of Sydney: Sydney School of Veterinary Science. [Accessed October 20, 2020].
- Osbak, K. K., Colclough, K., Saint-Martin, C., Beer, N. L., Bellanné-Chantelot, C., Ellard, S., et al. (2009). Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Hum. Mutat.* 30, 1512–1526. doi:10.1002/humu.21110
- PLINK (2017). Available at <https://zzz.bwh.harvard.edu/plink/contact.shtml> (Accessed October 20, 2020).
- Pollott, G. E. (2013). “Do selective sweeps in sheep breeds indicate the genomic sites of breed characteristics?,” in *Book of abstracts of the 64th European association for animal production annual meeting, nantes, France* (Netherlands: Wageningen Academic Publishers), 627.
- Pollott, G. E. (2018). Invited review: Bioinformatic methods to discover the likely causal variant of a new autosomal recessive genetic condition using genome-wide data. *Animal* 12, 2221–2234. doi:10.1017/S1751731118001970
- Posbergh, C. J., Pollott, G. E., Southard, T. L., Divers, T. J., and Brooks, S. A. (2018). A non-synonymous change in adhesion G protein-coupled receptor L3 associated with risk for Equine Degenerative Myeloencephalopathy in the Caspian Horse. *J. Equine Vet. Sci.* 70, 96–100. doi:10.1016/j.jevs.2018.08.010
- Postic, C., Shiota, M., Niswender, K. D., Jetton, T. L., Chen, Y., Moates, J. M., et al. (1999). Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *J. Biol. Chem.* 274, 305–315. doi:10.1074/jbc.274.1.305
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795
- Qiagen (2013). QIAGEN CLC genomics Workbench 20.0. Available at <https://digitalinsights.qiagen.com/>.
- Sayyab, S., Viluma, A., Bergvall, K., Brunberg, E., Jagannathan, V., Leeb, T., et al. (2016). Whole-Genome sequencing of a canine family trio reveals a FAM83G variant associated with hereditary footpad hyperkeratosis. *G3 (Bethesda)* 6, 521–527. doi:10.1534/g3.115.025643
- Spyer, G., Hattersley, A. T., Sykes, J. E., Sturley, R. H., and MacLeod, K. M. (2001). Influence of maternal and fetal glucokinase mutations in gestational diabetes. *Am. J. Obstet. Gynecol.* 185, 240–241. doi:10.1067/mob.2001.113127
- Spyer, G., Slingerland, A. S., Knight, B. A., Ellard, S., Clark, P. M., Hauguel-de Mouzon, S., et al. (2008). Mutations in the glucokinase gene of the fetus result in reduced placental weight. *Diabetes Care* 31, 753–757. doi:10.2337/dc07-1750
- UCSC (2020). Available at <https://genome.ucsc.edu/> (Accessed October 20, 2020).
- Varsome (2020). Available at <https://varsome.com/variant/hg19/rs1167675604> (Accessed October 20, 2020).
- Wooding, F. B., Fowden, A. L., Bell, A. W., Ehrhardt, R. A., Limesand, S. W., and Hay, W. W. (2005). Localisation of glucose transport in the ruminant placenta: Implications for sequential use of transporter isoforms. *Placenta* 26, 626–640. doi:10.1016/j.placenta.2004.09.013

Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

