frontiers | Research Topics

# Omics applied to livestock genetics, volume II

**Edited by**
Lucas Lima Verardo, Nuno Carolino and Ana Fabrícia Braga Magalhães

**Coordinated by**
Gabriel Dallago

**Published in**
Frontiers in Genetics
Frontiers in Veterinary Science

ISSN 1664-8714
ISBN 978-2-8325-5401-2
DOI 10.3389/978-2-8325-5401-2

## 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 applied to livestock genetics, volume II

**Topic editors**

Lucas Lima Verardo — Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

Nuno Carolino — National Institute for Agricultural and Veterinary Research (INIAV), Portugal

Ana Fabrícia Braga Magalhães — Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

**Topic coordinator**

Gabriel Dallago — University of Manitoba, Canada

**Citation**

Verardo, L. L., Carolino, N., Braga Magalhães, A. F., Dallago, G., eds. (2024). *Omics applied to livestock genetics, volume II*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5401-2

# Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 23 August 2024
DOI 10.3389/fgene.2024.1477826



OPEN ACCESS

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

*CORRESPONDENCE
Lucas Lima Verardo,
lucas.verardo@ufvjm.edu.br

RECEIVED 08 August 2024
ACCEPTED 19 August 2024
PUBLISHED 23 August 2024

CITATION
Verardo LL, Carolino N, Duarte MR, Rodrigues Almeida EA, Dallago G and Braga Magalhães AF (2024) Editorial: Omics applied to livestock genetics: volume II. *Front. Genet.* 15:1477826. doi: 10.3389/fgene.2024.1477826



# Editorial: Omics applied to livestock genetics: volume II

Lucas Lima Verardo[1]*, Nuno Carolino[2], Marcela Ramos Duarte[1], Emily Alves Rodrigues Almeida[1], Gabriel Dallago[3] and Ana Fabrícia Braga Magalhães[1]

[1]Laboratory of Animal Breeding, Department of Animal Science, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, Brazil, [2]Instituto Nacional Investigação Agraria e Veterinaria (INIAV), Oeiras, Portugal, [3]Department of Animal Science, University of Manitoba, Winnipeg, MB, Canada

KEYWORDS

**data integration, epigenomics, farm animals, genomics, multiomics, proteomics, transcriptomics**

Editorial on the Research Topic
Omics applied to livestock genetics: volume II

After the first sequencing of a mammalian genome, several studies have been published with a variety of omics datasets aiming to unravel the biological aspects that influence the phenotypic expression of complex traits (e.g., Gallagher and Chen-Plotkin, 2018; Costa et al., 2013; Legrain et al., 2011; Portela and Esteller, 2010; Marcobal et al., 2013; Fonseca et al., 2018). These studies revolutionized genome translation into phenome in the last 2 decades, including the development of important tools for the livestock sector. Projects, initiatives, and databases provide knowledge of genetic variations for the main traits of livestock species that are economically, environmentally, and socially important. For instance, the AnimalQTLdb project (Hu et al., 2007) has curated genomic information of many quantitative trait loci (QTL) identified in cattle, pigs, chicken, sheep, and other species.

The large-scale datasets generated by livestock "omics" projects have been made publicly available to researchers aiming to generate knowledge and translation tools for improving animal production and sustainability. For instance, the Functional Annotation of Animal Genomes (FAANG) project has generated datasets to decipher the function of genome segments, and it has analyzed samples from approximately 15 species, including pigs, cattle, sheep, and salmon (Giuffra et al., 2019). Moreover, the "omics" approaches can be holistically evaluated (Vieira et al., 2024) and applied to improve animal breeding strategies based on biology-driven genomic predictions, in addition of contributing to a better understanding of the genomic background of phenotypic variability in livestock systems (Chakraborty et al., 2022).

The Research Topic titled "Omics Applied to Livestock Genetics II" presents a collection of the latest findings in livestock genetics based on omics approaches. Studies focusing on livestock animals, such as pigs, cattle, ducks, camels, rabbits, donkeys, and sheep, involving omics data revealed genetic information related to various relevant traits. In volume I of Omics Applied to Livestock Genetics, the two most used approaches were genomics and transcriptomics applied, with cattle and pigs being the main studied species. Now, at volume II, fish was also highlighted (Figure 1). The results presented in this Research Topic provide significant advancements toward understanding farm animal genetics

10.3389/fgene.2024.1477826

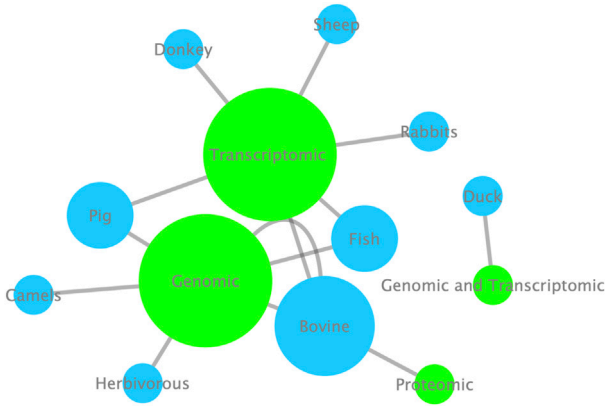



FIGURE 1
Enrichment analysis highlighting the most omic approaches (green nodes) used and the most livestock species (blue nodes) studied in the Research Topic. The size of the nodes corresponds to the Cytoscape network enrichment analysis; the bigger the nodes, the more used and/or studied the corresponding nodes.

The genomic knowledge of many commercially important fish species is still incipient, as is the case with some catfish species. The study by Ye et al. provided the first high-quality chromosomal-level reference genome map for the *Hemibagrus macropterus*. Using homology-based, *de novo*, and RNA-seq methods, various genes were identified and functionally annotated through a functional database. The assembled genome will be beneficial for exploring genome evolution and sexual determination mechanisms as well as facilitate future comparative genomics and conservation studies in Siluriformes. In another study also using catfish as a model, Wang et al. analyzed different organs in two species of catfish. The study performed RNA sequencing (RNA-seq) experiments to establish transcriptomic resources for both species' tissues (heart, liver, intestine, mucus, and muscle). A variable number of genes were identified as being expressed in each tissue, with a notable high expression of genes in the mucus. Gene ontology (GO) analysis revealed the functional specificity of differentially expressed genes in each tissue, with significant enrichment in metabolic pathways, immune activity, and stress responses. Tissue-specific genes such as *lrrc10*, *fabp2*, *myog*, *pth1a*, *hspa9*, *cyp21a2*, *agt*, and *ngtb* were identified. This study demonstrates that transcriptomics may be used to support further investigations into the molecular basis underlying environment-dependent heterosis and advance genetic breeding efforts of hybrid catfish

Studying camels, Abri et al. characterized the genetic diversity and selection signatures in camels from the Oman region. Using SNP (Single Nucleotide Polymorphism) genotyping data, genetic differentiation was observed due to evolutionary processes for adaptation between Muscat dromedaries and Al-Batinah and Al-Sharqiya populations. Candidate genes such as *SLC2A9*, *LEP*, and *PTPN22* were identified, which are involved in biological processes influencing the survival and reproduction of dromedaries in arid environments. For instance, *SLC2A9* is cited to be involved in glucose transport and may play a role in regulating energy metabolism. Genes associated with functional categories related to immune response, lipid metabolism, energy expenditure, optical and auditory functions, and long-term memory were also identified.

In this Research Topic, bovine was the most studied species. Rosyada et al. assessed the semen of native Indonesian bulls. Through proteomic investigations, they found that bull fertility is associated with many proteins involved in spermatogenesis. They identified 15 proteins linked to metabolic pathways and the tricarboxylic acid cycle, contributing to sperm energy production. Proteins related to thermal stress and predictors of thermotolerance, such as *HSPA9* and *HSPA2*, were identified as protective agents for sperm. Jang et al. studied Hanwoo cattle, a native breed from Korea known for high fertility but slow growth rates. This study collected samples from 22 different tissues of castrated males and utilized RNA-seq technology for gene expression profiling, integrating eQTL analysis to elucidate the genetic mechanisms influencing weight. By integrating the results from eQTL analyses and differentially expressed genes (DEGs), genomic regions that may regulate the expression of candidate genes, such as *TRIM31*, were identified. Reduced expression of *TRIM31* was associated with weight gain, which can be explained by cis-eQTL candidate genotypes and their effect on differential gene expression between lower and higher weight groups.

Also studying bovine specie, Wang et al. utilized Specific Locus Amplified Fragment (SLAF) sequencing to examine the genetic structure and diversity of Xinjiang Brown (XBG) cattle. Selection signature analysis revealed differentiated patterns between XBG and the ancestral breeds Swiss Brown (BS) and Kazakh (KZ). Besides, candidate genes enriched with GO terms related to disease resistance and the endocrine system were identified. This research enhances the understanding of genetic diversity in Xinjiang Brown cattle and provides valuable insights for future selection and genetic breeding practices. The fourth study using bovine as a model, Nawaz et al., sought to identify genomic regions under selection in Hanwoo and Angus cattle using advanced genetic analysis methods based on imputed whole-genome sequencing variants. The study employed allele frequency-based and haplotype-based methods, including runs of homozygosity and extended haplotype homozygosity, to detect selection signals within each breed and between the two breeds. In Angus cattle, 27 genomic regions housing 360 genes were identified. The identified genes are associated with growth, immunity, reproductive development, feed efficiency, and environmental adaptation, suggesting that selection processes in this breed focused on productivity and environmental adaptability. In contrast, Hanwoo cattle displayed 17 genomic regions containing 59 genes. Candidate genes indicated that selection in Hanwoo prioritized traits related to meat quality and sensory perception. This study enhances the understanding of the genetic architecture of selection in Angus and Hanwoo cattle, highlighting breed-specific adaptations and priorities.

Considering the studies focusing on pigs, Xiao et al. studied the Acetyl-Coenzyme A Acyltransferase-1 (*ACAA1*) gene, which is involved in fatty acid metabolism. The study assessed the mRNA expression levels of *ACAA1* in various tissues (heart, liver, spleen, lung, kidney) from 6-month-old Xiangsu pigs. The mRNA expression was also evaluated in the *Longissimus dorsi* muscle at different growth stages (newborn, 6 months, and 12 months) using RT-qPCR. Additionally, the relationship between single-nucleotide polymorphisms (SNPs) of *ACAA1* gene and growth traits in 6-month-old and 12-month-old Xiangsu pigs was investigated on 184 healthy Xiangsu pigs. The expression of *ACAA1* was detected in

multiple tissues of 6-month-old Xiangsu pigs, with the highest expression observed in the liver. In the *L. dorsi*, expression decreased as the animals grew. Significant associations between SNPs and these growth traits suggest that the *ACAA1* gene is a potential marker for genetic selection in Xiangsu pigs. In another study, Wang et al. investigated the correlation between intramuscular fat content (IFC) and meat color (CIE) in pigs using an integrative approach that combines differential gene expression analysis, gene co-expression network analysis (WGCNA), functional enrichment analysis, and the identification of candidate and hub genes. The results revealed 485 and 394 DEGs and identified 47 and 53 candidate genes affecting IFC and CIE, respectively. Protein-protein interaction (PPI) network analysis of the candidate genes identified 5 hub genes affecting IFC and 13 hub genes affecting CIE value. Four crucial hub genes (*MYC*, *SOX9*, *CEBPB*, and *PPARGC1A*) were shared between these two traits. The authors propose that the *SOX9/CEBPB/PPARGC1A* axis may co-regulate lipid metabolism and the redox response of myoglobin, contributing to a better understanding of the molecular mechanism underlying the co-regulation of IFC and CIE value.

Using an integrative approach, the study by Li et al. on Jianshui yellow-brown ducks used resequencing and transcriptomic data and revealed several SNPs and InDels, with variants identified in genes associated with muscle development and fat metabolism. This study used phylogenetic trees, PCA, and admixture analysis to investigate the population genetic structure of Jianshui yellow-brown ducks by comparing their selection signals with those of ancestral mallard ducks and meat-type Pekin ducks. Selection signal analysis indicated significant selection pressure on genes related to meat traits (*ELOVL2*, *ELOVL3*, *GDF10*, *VSTM2A*, *PHOSPHO1*, and *IGF2BP*1) in Jianshui yellow-brown ducks and mallard ducks. While transcriptomic analysis suggested that *ELOVL3*, *PHOSPHO1* and *GDF10* are candidate genes influencing meat production and quality in Jianshui yellow-brown ducks.

Another species of significant economic interest presented in this Research Topic was the rabbit (*Oryctolagus cuniculus*), in which Jia et al. obtained samples from the *Longissimus dorsi* muscle of male and female rabbits and utilized Oxford Nanopore Technologies long-read sequencing technology to investigate the association between gene expression levels and growth traits through large-scale transcriptome-wide association studies (TWAS). This contributed to the improvement of rabbit genome annotation. However, the transcriptome-wide association studies did not identify statistically significant genes or transcripts associated with the growth traits examined, highlighting the need for other omics studies in this species aiming to a better understanding of growth traits' genetic architecture.

Aiming to investigate the transcriptomic screening of lncRNAs and mRNA associated with skin development and collagen organization, Wang et al. obtained skin tissue samples from Dezhou donkeys at different stages, including the 8-month fetal stage, and at ages of 2 and 8 years. Through enrichment analyses and functional analyses, it was possible to identify specific lncRNAs and interactions between mRNA and lncRNA. Specific lncRNAs, including *ENSEAST00005041187*, *ENSEAST00005038497* and *MSTRG.17248.1*, which potentially regulate the *COL1A1* gene that is responsible for the type I collagen chain I, were identified through interaction networks. Collagen organization and skin development pathways were also observed, including protein digestion and absorption, metabolic pathways, phosphatidylinositol 3-kinase-protein kinase B signaling pathway (PI3K-Akt signaling pathway), extracellular matrix-receptor interaction (ECM-receptor interaction) and relaxin signaling and biological function processes. The *COL1A1*, *COL3A1* and *LOXL2* genes were involved in the regulation of these pathways.

Finally, Liu et al. reviewed Copy Number Variants (CNVs) in the genomes of herbivorous livestock species, including cattle, sheep, horses, and donkeys. They presented a brief elucidation of the fundamental concepts underlying CNVs, their mutational mechanisms, and the diverse array of detection methods that can be employed to identify these structural variations within genomes. The review highlighted the role of CNVs in shaping various phenotypic traits, including growth and reproductive traits, pigmentation, disease resistance, etc. In conclusion, the authors stated that CNVs represent a valuable and dynamic field of study poised to impact the genetic improvement of herbivorous livestock species, ultimately benefiting both human society and the global livestock industry.

As observed in the first volume, the main livestock species have been studied through omics approaches. However, multiomic analyses are still scarce, and the generation and sharing of multiomic datasets are crucial for further advancing research in this field. Functional genomic analyses and high-throughput phenotyping are relevant for providing a clearer picture of the genome-to-phenome paradigm in livestock systems. Moreover, the integration of omics technologies with phenomics into the breeding programs, which is still lacking in this Research Topic, may be helpful for increasing the rate of genetic progress in sustainable breeding programs.

## Author contributions

LV: Writing–original draft, Writing–review and editing. NC: Writing–original draft, Writing–review and editing. Marcela Ramos: Writing–original draft, Writing–review and editing. ER: Writing–review and editing, Writing–original draft. GD: Writing–original draft, Writing–review and editing. AB: Writing–original draft, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This article was partly supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico—(CNPq) (process 314532/2021-8) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais—FAPEMIG (process APQ-02638-24).

## Acknowledgments

The editors of this Research Topic thank the valuable contributions of all the authors to this Research Topic, the constructive comments and suggestions from the reviewers, and

the editorial support from Frontiers team throughout the publication process.

## 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



## References

Chakraborty, D., Sharma, N., Kour, S., Sodhi, S. S., Gupta, M. K., Lee, S. J., et al. (2022). Applications of omics technology for livestock selection and improvement. *Front. Genet.*, 13. doi:10.3389/fgene.2022.774113

Costa, V., Aprile, M., Esposito, R., and Ciccodicola, A. (2013). RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.*, 21(2), 134–142. doi:10.1038/ejhg.2012.129

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(10), p.e0205295. doi:10.1371/journal.pone.0205295

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, 102(5), 717–730. doi:10.1016/j.ajhg.2018.04.002

Giuffra, E., Tuggle, C. K., and Faang Consortium. (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Annu. Rev. Anim. Biosci.*, 7, 65–88. doi:10.1146/annurev-animal-020518-114913

Hu, Z. L., Fritz, E. R., and Reecy, J. M. (2007). AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Res.*, 35, D604–D609. doi:10.1093/nar/gkl946

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., et al. (2011). The human proteome project: current state and future direction. *Mol Cell Proteomics*, 10(7). doi:10.1074/mcp.M111.009993

Marcobal, A., Kashyap, P. C., Nelson, T. A., Aronov, P. A., Donia, M. S., Spormann, A., et al. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME J.*, 7(10), 1933–1943. doi:10.1038/ismej.2013.89

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. *Nat. Biotechnol.*, 28(10), 1057–1068. doi:10.1038/nbt.1685

Vieira, J. I. G., Braga, L. G., Chud, T. C., Ferreira, P. H., Guimarães, S. E. F., Martins, M. F., et al. (2024). Resequencing of Brazilian locally adapted cattle breeds revealed variants in candidate genes and transcription factors for meat fatty acid profile. *J. Anim. Breed. Genet.*, 00, 1–15. doi:10.1111/jbg.12869

frontiers | Frontiers in Genetics

TYPE Data Report
PUBLISHED 01 November 2023
DOI 10.3389/fgene.2023.1297119



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Mingzheng Duan,
Zhaotong University, China
Zhongwei Wang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Yong Xie,
799091142@qq.com
Chuangju Li,
lcj@yfi.ac.cn

RECEIVED 19 September 2023
ACCEPTED 23 October 2023
PUBLISHED 01 November 2023

CITATION
Ye H, Fan J, Hou Y, Yue H, Ruan R, Li S, Hu C, Xie Y and Li C (2023), Chromosome-level genome assembly of the largefin longbarbel catfish (*Hemibagrus macropterus*). *Front. Genet.* 14:1297119. doi: 10.3389/fgene.2023.1297119



# Chromosome-level genome assembly of the largefin longbarbel catfish (*Hemibagrus macropterus*)

Huan Ye[1], Jiahui Fan[1], Yanling Hou[1], Huamei Yue[1], Rui Ruan[1], Shuang Li[2], Chongjiang Hu[2], Yong Xie[2]* and Chuangju Li[1]*

[1]Key Laboratory of Freshwater Biodiversity Conservation, Ministry of Agriculture and Rural Affairs, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, China, [2]Chongqing Fishery Sciences Research Institute, Chongqing, China

The largefin longbarbel catfish, *Hemibagrus macropterus*, is an economically important fish species in southwestern China, with males growing faster than females. This study presents a high-quality chromosome-level genome assembly of the largefin longbarbel catfish, generated by integrating Illumina short reads, PacBio HiFi long reads, and Hi-C data. The assembled genome size was 858.5 Mb, with a contig and scaffold N50 of 5.8 Mb and 28.4 Mb, respectively. A total of 656 contigs were successfully anchored to 30 pseudochromosomes with a BUSCO score of 97.7%, consistent with the number of chromosomes analyzed by karyotype. The genome contained 29.5% repeat sequences, and a predicted total of 26,613 protein-coding genes, of which 25,769 (96.8%) were functionally annotated in different databases. Evolutionary analysis showed that *H. macropterus* was most closely related to *H. wyckioides*, with a divergence time of approximately 16.3 million years. Chromosomal syntenic relationships among *H. macropterus*, *H. wyckioides*, and *Pelteobagrus fulvidraco* revealed a one-to-one relationship for most chromosomes, except for break, fission, and inversion of some chromosomes. The first high-quality reference genome will not only provide a valuable genetic resource for the study of sex determination mechanisms and genetic breeding of largefin longbarbel catfish, but also contribute to comparative analyses of genome and chromosome evolution within Siluriformes.

KEYWORDS

***Hemibagrus macropterus*, genome assembly, genome annotation, comparative genomics, Hi-C, PacBio**

## Introduction

Catfish (order: Siluriformes) are a highly diverse and globally distributed group of actinopterygian fish, generally characterized by the whisker-like barbels, lack scales, and intramuscular spines (Gisbert et al., 2022). They comprise more than 4,500 species and account for nearly 12% of teleost fish (Fricke et al., 2022). Catfish are one of the most important aquaculture species worldwide (Gisbert et al., 2022). The number of chromosomes in catfish ranges from 2n = 24 to 100, with mainly continuous variation from 2n = 48 to 60 (Zhu and Pan, 2021). Consequently, catfish are considered suitable for studying genomic and chromosomal evolution in fish. With the rapid development of sequencing technologies, the chromosome-level genomes of more than 10 catfish species have been assembled, including

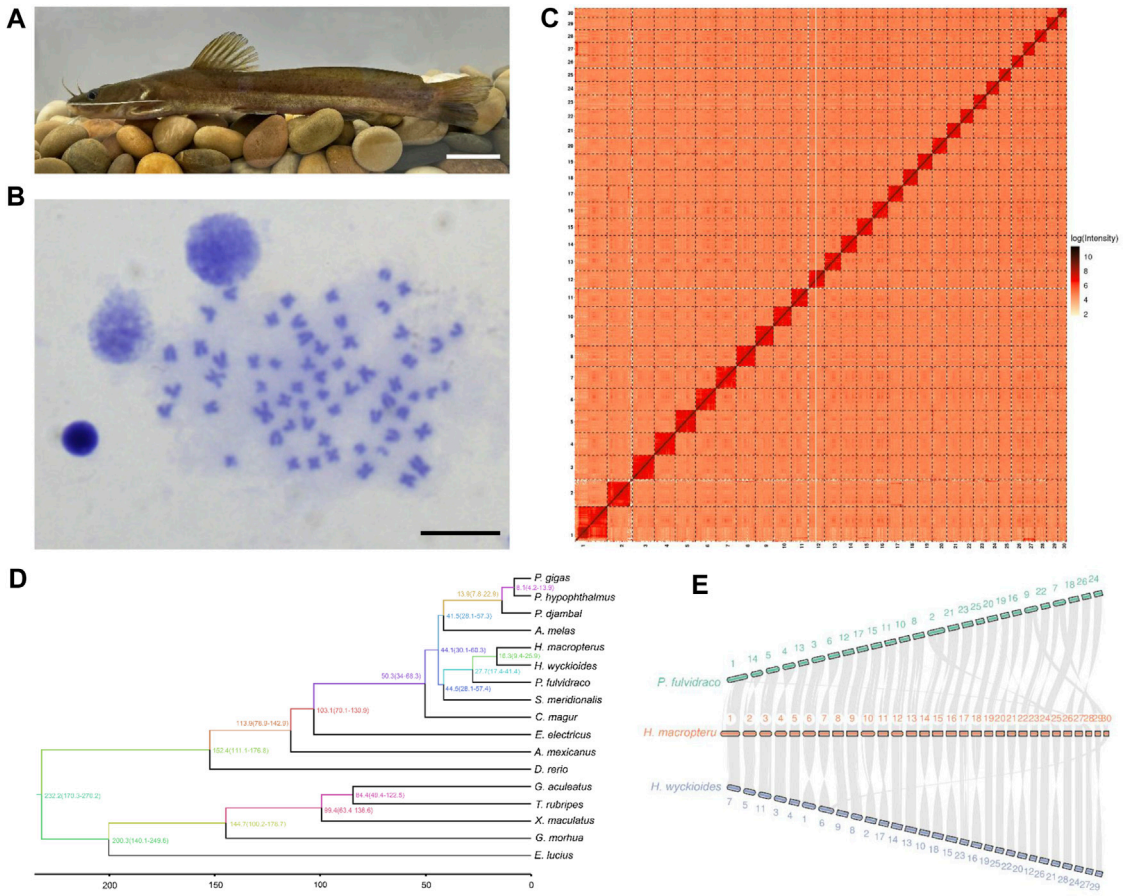



FIGURE 1
Chromosome-level genome assembly of *Hemibagrus macropterus* and comparative genomics analysis. **(A)** Representative image of *H. macropterus*. Scale bar represents 5 cm. **(B)** Karyotype of male *H. macropterus*. Scale bar represents 10 μm. **(C)** Heatmap of Hi-C interactions among 30 pseudochromosomes. Colour depth represents the density of the Hi-C interactions. **(D)** Phylogenetic relationships between *H. macropterus* and 16 other fish species based on 3,105 single-copy orthologous genes. **(E)** Chromosomal syntenic relationships among *H. macropterus*, *H. wyckioides*, and *Pelteobagrus fulvidraco*.

*Ictalurus punctatus* (Liu et al., 2016), *Pelteobagrus fulvidraco* (Gong et al., 2018), *Bagarius yarrelli* (Jiang et al., 2019), *Silurus meridionalis* (Zheng et al., 2021), *Leiocassis longirostris* (He et al., 2021), *Hemibagrus wyckioides* (Shao et al., 2021), *Pangasianodon hypophthalmus* (Gao et al., 2021), *Pseudobagrus ussuriensis* (Zhu et al., 2022), *Cranoglanis bouderius* (Xu et al., 2022), *Ictalurus furcatus* (Wang et al., 2022), and *Ancistrus triradiatus* (Lemopoulos and Montoya-Burgos, 2022). These genomic resources facilitate studies of sex determination mechanisms (Bao et al., 2019; Gong et al., 2022), chromosomal and genome evolution (Zhu et al., 2022), ecological adaptation, and gene evolution and function (Liu et al., 2016; Zhou et al., 2023).

The largefin longbarbel catfish (*Hemibagrus macropterus*) (Figure 1A), belonging to the Bagridae family (Siluriformes), is an important commercial fish in southwestern China because of its high nutritional value (Zhang et al., 2009). It is a benthic dweller naturally distributed in the mainstream and tributaries of the Yangtze and Pearl River Basins (Zhu et al., 2007). Largefin longbarbel catfish exhibit sexual size dimorphism, with males growing faster than females. Two karyotypes (2n = 56 and 60) have been reported for this species among different populations (Hong and Zhou, 1984; Ma et al., 2013). However, the genetic resources of largefin longbarbel catfish are limited, which is unfavorable for understanding its genetic characteristics and developing breeding programs.

In this study, we present the first high-quality chromosome-level reference genome of *H. macropterus*. The assembled genome will be beneficial for exploring the genome evolution, sex determination mechanisms, and genetic breeding of largefin longbarbel catfish. Furthermore, this contribution to the genomic resources of Siluriformes will facilitate future comparative genomic studies among catfish.

# Data

## Genome assembly

A total of 31.4 Gb Illumina clean data were used to assess genome size and heterozygosity in *H. macropterus*. The predicted genome size was approximately 873.7 Mb and the estimated heterozygosity rate was 0.37%. For *de novo* genome assembly, 41.8 Gb PacBio HiFi reads were preliminarily assembled into 691 contigs with an N50 length of 5.8 Mb, covering 98.3% of the

**TABLE 1 Statistics of *Hemibagrus macropterus* genome assembly and annotation.**

| Item | Category | Number |
|---|---|---|
| Sequencing data | PacBio HiFi (Gb) | 41.8 |
| | Illumina short WGS (Gb) | 31.4 |
| | Hi-C(Gb) | 98.4 |
| Assembly | Estimated genome size (Mb) | 873.7 |
| | Assembled genome size (Gb) | 858.5 |
| | Contig number | 35 |
| | Contig N50 (Mb) | 5.8 |
| | Scaffold number | 30 |
| | Scaffold N50 (Mb) | 28.4 |
| | Longest scaffold (Mb) | 57.5 |
| Annotation | GC content (%) | 40.0 |
| | Repeat sequences (%) | 29.5 |
| | Number of protein-coding genes | 26,613 |
| | Number of functional annotated genes | 25,769 |
| | Average gene length (bp) | 25,071.3 |
| | Average exon length (bp) | 171.4 |
| | Average intron length (bp) | 2,880.7 |
| | Average exon per gene | 9.2 |

estimated genome. Using the Hi-C technique, a total of 656 contigs were successfully anchored to 30 pseudochromosomes (Figure 1C), consistent with the number of chromosomes analyzed by karyotype (Figure 1B) and reported in a previous study (Hong and Zhou, 1984). The assembled chromosome-level genome consisted of 35 contigs and 30 scaffolds, with a contig and scaffold N50 length of 5.8 and 28.4 Mb, respectively (Table 1), which represented 98.3% of the estimated genome. The guanine-cytosine (GC) content was 40%, similar with that of other Bagridae (Zhu et al., 2022). Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis revealed 97.7% of BUSCO genes identified in the genome (Supplementary Table S1), indicating high completeness for the genome assembly.

## Genome annotation

The genome of *H. macropterus* contained 29.5% repetitive sequences (Supplementary Table S2), with transposable element (TE) accounting for 18.06% of the assembled genome. The largest proportion of TE was terminal inverted repeats (9.29%), followed by long terminal repeat retrotransposons (7.43%). Together with homology, *de novo*, and RNA-seq prediction methods, 26,613 protein-coding genes were annotated (Supplementary Table S3). The average length of gene, exon, and intron was 25,071, 171, and 2,881 bp, respectively (Table 1). BUSCO assessments showed that 98.2% complete BUSCO genes were predicted, including 96.1% single copy and 2.1% duplicated genes (Supplementary Table S4). These results indicated high-quality genome assembly and annotation of *H. macropterus*.

## Genome evolution analysis

The evolutionary relationships between *H. macropterus* and other teleosts were determined based on the analysis of 3,105 single-copy orthologous genes from of 17 fish genomes (Supplementary Figure S1). *H. wyckioides* was most closely related to *H. macropterus*, consistent with their taxonomic relationship (Fricke et al., 2022), and clustered with *P. fulvidraco* (Figure 1D). The nine species of Siluriformes formed a monophyletic clade, and then together with *Electrophorus electricus* (Gymnotiformes), *Astyanax mexicanus* (Characiformes), and *Danio rerio* (Cypriniformes), formed the clade of Ostariophysan. According to the fossil calibration times, the estimated divergence time between *H. macropterus* and *H. wyckioides* was approximately 16.3 Mya, and the divergence time between *M. macropterus* and *P. fulvidraco* was around 27.7 Mya.

Through comparative genomic analysis, we identified 60 and 28 gene families, respectively, that underwent significant expansion and contraction in *H. macropterus* (Supplementary Figure S2). Enrichment analysis revealed that the expanded and contracted genes were enriched in 18 and 7 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, respectively (Supplementary Tables S5, S6), with most involved in the immunity, metabolism, and hormone biosynthesis. These results provide valuable preliminary information on the biological properties of the species.

## Synteny anslysis

We compared the chromosome syntenies between *H. macropterus* and two other catfish species. The karyotype of these three species (*P. fulvidraco*, *H. macropterus*, and *H. wyckioides*) was 2n = 52, 2n = 60, and 2n = 59, respectively. Most of the chromosomes between *P. fulvidraco* and *H. macropterus* exhibited a one-to-one relationship (Figure 1E), whereas the chromosomes (Chr) 1, 2, 7, and 9 of *P. fulvidraco* broke into two chromosomes in *H. macropterus*. *H. macropterus* and *H. wyckioides* displayed a strong one-to-one correspondence among their chromosomes, except for the fission of *H. wyckioides* Chr6 into Chr7 and 30, and inversion of some chromosomes (Chr6, 7, 10, 11, 14, 15, 16, 18, 21, 22, 25, and 29) in *H. macropterus*. It was recently reported that the sex-determining region of *P. fulvidraco* and *H. wyckioides* was located on the Chr 2 and 26, respectively (Gong et al., 2022), whereas the correspondent chromosomes in *H. macropterus* were Chr14 and 15 and 24, respectively, indicating the complexity of the sex-determining region or chromosome in catfish.

# Materials and methods

## Sample collection and sequencing

A male *H. macropterus* was collected for genome sequencing from the Wuhan section of the Yangtze River. After anesthesia with

3-aminobenzoic acid ethyl ester methanesulfonate-222 (MS-222) (Sigma-Aldrich, St. Louis, MO, United States), muscle tissue was collected, frozen in liquid nitrogen, and stored at −80°C, while other tissues including brain, gills, heart, intestine, kidney, liver, spleen, and testis were stored in RNAlater solution (Sigma-Aldrich). All experiments involving in the handling and treatment of fish were conducted in accordance with the Guidelines for the Care and Use of Laboratory Animals of the Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences.

Total genomic DNA was isolated from frozen muscle tissue using a Blood & Cell Culture DNA Kit (Qiagen, Hilden, Germany). DNA quality and purity were determined using a Nanodrop 2000 (Thermo Fisher Scientific, Waltham, MA, United States) and agarose gel electrophoresis. A 350 bp paired-end library was constructed using an Illumina TruSeq DNA Nano Preparation Kit (Illumina, San Diego, CA, United States) and sequenced on an Illumina HiSeq 2500 platform (Illumina). Approximately 5 μg of genomic DNA was used to construct a PacBio SMRTbell library. The library was sequenced using a PacBio Circular Consensus Sequencing (CCS) Platform (PacBio, Menlo Park, CA, United States). A Hi-C library was prepared using a GrandOmics Hi-C kit according to the manufacturer's instructions and sequenced on an Illumina NovaSeq platform (Illumina). Total RNA was extracted from different tissues of *H. macropterus* using an RNeasy Plus Mini Kit (Qiagen). A complementary DNA library was constructed using a TruSeq Stranded mRNA-Seq kit on the Illumina HiSeq 2500 platform (Illumina).

## Preparation of chromosome metaphases

Three male *H. macropterus* juveniles were injected twice with phytohemagglutinin (PHA) at 10 μg/g body weight with a 12 h interval, injected with colchicine at 10 μg/g body weight for 3 h, and anaesthetised using MS-222. Kidney cells were collected by hypotonic and fixation treatments as previously described (Zhu and Gui, 2007) and the number of mitotic metaphase chromosomes was counted in 100 cells.

## Genome assembly and evaluation

Illumina clean short reads were used to estimate genome size and heterozygosity based on k-mer frequency distribution analysis using Jellyfish (Marçais and Kingsford, 2011). The PacBio long reads were assembled *de novo* into contigs using Hifiasm (Cheng et al., 2021), and then polished using Illumina short reads and NextPolish (Hu et al., 2019). For chromosome-level assembly of the *H. macropterus* genome, clean Hi-C reads were mapped to the primary genome using Bowtie2 (Langmead and Salzberg, 2012). HiC-Pro was used to validate interacting paired reads (Servant et al., 2015). Primary assembly scaffolds were oriented, ordered, and clustered on pseudochromosomes using LACHESIS (Korbel and Lee, 2013). JuiceBox was used to adjust the placement and orientation errors (Durand et al., 2016), and a Hi-C heat map was constructed.

Two strategies were used to assess genome completeness. The BUSCO completeness score of the assembled genome was evaluated using the Actinopterygii database (Simão et al., 2015), and RNA-seq data were mapped back to the genome using HISAT2 with default settings (Kim et al., 2015).

## Genome annotation

Repetitive elements of the *H. macropterus* genome were annotated using both homology and *de novo* strategies. According to the structural features, tandem and simple sequence repeats were predicted using TRF (Benson, 1999) and MISA (Beier et al., 2017), respectively, with default parameters. The transposable elements were identified using LTR_Finder (Ou and Jiang, 2019), LTRharverst (Ellinghaus et al., 2008), and LTR_retriver (Ou and Jiang, 2017). *De novo* annotation of other repeat sequences was performed using RepeatModeler (Price et al., 2005), followed by genome-scale detection using RepeatMasker (Chen, 2004). The combined results of these two predictions provided the final annotation of the non-redundant repeat elements in the genome.

We combined *de novo*, homology, and transcriptome-based methods to predict protein-coding genes. For *de novo* prediction, we used Augustus (Keller et al., 2011), GlimmerHMM (Majoros et al., 2004), and Geneid (Blanco et al., 2007) with their default parameters. Protein sequences of *D. rerio*, *E. electricus*, *Esox lucius*, *Gadus morhua*, *H. wyckioides*, *Silurus meridionalis*, *P. fulvidraco*, and *Takifugu rubripes* were aligned to the genome of *H. macropterus* using TBLASTN (Camacho et al., 2009). GeneWise (Birney et al., 2004) was used to predict the gene structure according to homology alignments. For transcriptome-based prediction, protein-coding regions were identified by aligning the transcripts with the assembled genome using PASA (Haas et al., 2003). Transposons were removed using TransposonPSI, and the final non-redundant reference gene set was obtained using EVidenceModeler (Haas et al., 2008).

For the functional annotation, the gene set was aligned to proteins deposited in the SwissProt and NCBI non-redundant protein databases using BLASTP. KEGG pathways were annotated by the KEGG Automatic Annotation Server (Moriya et al., 2007). Gene Ontology and protein domains were identified using InterProScan (Jones et al., 2014) with default parameters.

## Phylogenetic and comparative genomic analyses

Orthologous gene families were identified by comparing the predicted protein sequences of H. macropterus with those of 16 other fish using OrthoMCL (Li et al., 2003). Single-copy gene orthogroups among these species were selected and aligned using MAFFT V7 (Yamada et al., 2016). A maximum likelihood phylogenetic tree was constructed using RAXML7 (Stamatakis, 2015) with 1,000 bootstrap replicates. Divergence time was estimated using MCMCTREE in PAML4 (Yang, 2007), and calibrated using fossil divergence times from the TimeTree database (http://www.timetree.org/). Based on the results of OrthoMCL, expanded and contracted gene families were analyzed via CAFE (De Bie et al., 2006), and functional

enrichment analysis was performed by alignment homologues against the KEGG pathway database.

### Chromosomal syntenic analysis

To investigate the chromosomal syntenic relationships among *H. macropterus*, *H. wyckioides*, and *P. fulvidraco*, MCscan (Tang et al., 2008) was applied to determine the syntenic blocks. Proteomes were compared between the pairs of species using BLASTP with an e-value of 1e-5, and a minimum of four genes in each block were used for synteny calling.

## 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 studies were approved by the Animal Experimental Ethical Inspection of Laboratory Animal Centre, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was obtained from the owners for the participation of their animals in this study.

## Author contributions

HYe: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. JF: Data curation, Investigation, Methodology, Validation, Writing–original draft. YH: Data curation, Investigation, Validation, Visualization, Writing–original draft. HYu: Investigation, Methodology, Writing–original draft. RR: Writing–original draft. SL: Data curation, Resources, Writing–original draft. CH: Resources, Writing–original draft. YX: Conceptualization, Funding acquisition, Resources, Supervision, Writing–original draft. CL: Conceptualization, Writing–original draft, Writing–review and editing.

## Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Key R&D Program of China (2022YFD2400101); Chongqing Science and Technology Bureau (cstc2021jxjl80023); and the Science and Technology Project of Guizhou Province, China [(2020)4Y027].

## 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.2023.1297119/full#supplementary-material

**SUPPLEMENTARY FIGURE S1**
Comparison of the number of homologous genes present in the genomes of *Hemibagrus macropterus* and 16 other fish species.

**SUPPLEMENTARY FIGURE S2**
Expansion and contraction of *Hemibagrus macropterus* gene families. The number and sector with green, red, and blue represent the specific values of the gene families of expansion, contraction, and stability for each species, respectively. MRCA, most recent common ancestor.

## References

Bao, L., Tian, C., Liu, S., Zhang, Y., Elaswad, A., Yuan, Z., et al. (2019). The Y chromosome sequence of the channel catfish suggests novel sex determination mechanisms in teleost fish. *BMC Biol.* 17 (1), 6. doi:10.1186/s12915-019-0627-7

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

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27 (2), 573–580. doi:10.1093/nar/27.2.573

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14 (5), 988–995. doi:10.1101/gr.1865504

Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. *Curr. Protoc. Bioinform.* 18 (1), Unit 4.3–4.3.28. doi:10.1002/0471250953.bi0403s18

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinforma.* 10 (1), 421. doi:10.1186/1471-2105-10-421

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* 5 (1), Unit 4.10–14.10.14. doi:10.1002/0471250953.bi0410s05

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi:10.1038/s41592-020-01056-5

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22 (10), 1269–1271. doi:10.1093/bioinformatics/btl097

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Sys* 3 (1), 99–101. doi:10.1016/j.cels.2015.07.012

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinforma.* 9 (1), 18. doi:10.1186/1471-2105-9-18

Fricke, R., Eschmeyer, W. N., and Van der Laan, R. (2022). Eschmeyer's catalog of fishes: genera, species, references. Available at: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp.Electronic version (Accessed January 23, 2023).

Gao, Z., You, X., Zhang, X., Chen, J., Xu, T., Huang, Y., et al. (2021). A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). *Genomics* 113 (5), 3349–3356. doi:10.1016/j.ygeno.2021.07.026

Gisbert, E., Luz, R. K., Fernández, I., Pradhan, P. K., Salhi, M., Mozanzadeh, M. T., et al. (2022). Development, nutrition, and rearing practices of relevant catfish species (Siluriformes) at early stages. *Rev. Aquac.* 14 (1), 73–105. doi:10.1111/raq.12586

Gong, G., Dan, C., Xiao, S., Guo, W., Huang, P., Xiong, Y., et al. (2018). Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *GigaScience* 7 (11), giy120. doi:10.1093/gigascience/giy120

Gong, G., Xiong, Y., Xiao, S., Li, X.-Y., Huang, P., Liao, Q., et al. (2022). Origin and chromatin remodeling of young X/Y sex chromosomes in catfish with sexual plasticity. *Natl. Sci. Rev.* 10 (2), nwac239. doi:10.1093/nsr/nwac239

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr, Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31 (19), 5654–5666. doi:10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9 (1), R7. doi:10.1186/gb-2008-9-1-r7

He, W. P., Zhou, J., Li, Z., Jing, T. S., Li, C. H., Yang, Y. J., et al. (2021). Chromosome-level genome assembly of the Chinese longsnout catfish *Leiocassis longirostris*. *Zool. Res.* 42 (4), 417–422. doi:10.24272/j.issn.2095-8137.2020.327

Hong, Y. H., and Zhou, T. (1984). Karyotype of nine species of Chinese catfishes (Bagridae). *Zool. Res.* 5, 21–28. (in Chinese).

Hu, J., Fan, J., Sun, Z., and Liu, S. (2019). NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36 (7), 2253–2255. doi:10.1093/bioinformatics/btz891

Jiang, W., Lv, Y., Cheng, L., Yang, K., Bian, C., Wang, X., et al. (2019). Whole-genome sequencing of the giant devil catfish, *Bagarius yarrelli*. *Genome Biol. Evol.* 11 (8), 2071–2077. doi:10.1093/gbe/evz143

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30 (9), 1236–1240. doi:10.1093/bioinformatics/btu031

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27 (6), 757–763. doi:10.1093/bioinformatics/btr010

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

Korbel, J. O., and Lee, C. (2013). Genome assembly and haplotyping with Hi-C. *Nat. Biotechnol.* 31 (12), 1099–1101. doi:10.1038/nbt.2764

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi:10.1038/nmeth.1923

Lemopoulos, A., and Montoya-Burgos, J. I. (2022). Whole genome assembly of the armored loricariid catfish *Ancistrus triradiatus* highlights herbivory signatures. *Mol. Genet. Genomics* 297 (6), 1627–1642. doi:10.1007/s00438-022-01947-6

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13 (9), 2178–2189. doi:10.1101/gr.1224503

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., et al. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat. Commun.* 7 (1), 11757. doi:10.1038/ncomms11757

Ma, Y. G., Zhang, F., and Yuan, W. A. (2013). Analysis the karotype of *Mystus macropterus* bleeker in the jialin river. *Mod. Agr. Sci. Technol.* 2, 260–261. (in Chinese).

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 20 (16), 2878–2879. doi:10.1093/bioinformatics/bth315

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27 (6), 764–770. doi:10.1093/bioinformatics/btr011

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35 (2), W182–W185. doi:10.1093/nar/gkm321

Ou, S., and Jiang, N. (2017). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176 (2), 1410–1422. doi:10.1104/pp.17.01310

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* 10 (1), 48. doi:10.1186/s13100-019-0193-0

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21 (1), i351–i358. doi:10.1093/bioinformatics/bti1018

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., et al. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16 (1), 259. doi:10.1186/s13059-015-0831-x

Shao, F., Pan, H., Li, P., Ni, L., Xu, Y., and Peng, Z. (2021). Chromosome-level genome assembly of the Asian red-tail catfish (*Hemibagrus wyckioides*). *Front. Genet.* 12, 747684. doi:10.3389/fgene.2021.747684

Simão, 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 (19), 3210–3212. doi:10.1093/bioinformatics/btv351

Stamatakis, A. (2015). Using RAxML to infer phylogenies. *Curr. Protoc. Bioinform.* 51 (1), 6.14.11–6. doi:10.1002/0471250953.bi0614s51

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18 (12), 1944–1954. doi:10.1101/gr.080978.108

Wang, H., Su, B., Butts, I. A. E., Dunham, R. A., and Wang, X. (2022). Chromosome-level assembly and annotation of the blue catfish *Ictalurus furcatus*, an aquaculture species for hybrid catfish reproduction, epigenetics, and heterosis studies. *GigaScience* 11, giac070. doi:10.1093/gigascience/giac070

Xu, Y., Shao, F., Chen, W., Ni, L., and Peng, Z. (2022). A chromosome-level genome of the helmet catfish (*Cranoglanis bouderius*). *Front. Genet.* 13, 962406. doi:10.3389/fgene.2022.962406

Yamada, K. D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics* 32 (21), 3246–3251. doi:10.1093/bioinformatics/btw412

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24 (8), 1586–1591. doi:10.1093/molbev/msm088

Zhang, M., Liu, W., and Li, G. (2009). Isolation and characterisation of collagens from the skin of largefin longbarbel catfish (*Mystus macropterus*). *Food Chem.* 115 (3), 826–831. doi:10.1016/j.foodchem.2009.01.006

Zheng, S., Shao, F., Tao, W., Liu, Z., Long, J., Wang, X., et al. (2021). Chromosome-level assembly of southern catfish (*silurus meridionalis*) provides insights into visual adaptation to nocturnal and benthic lifestyles. *Mol. Ecol. Resour.* 21 (5), 1575–1592. doi:10.1111/1755-0998.13338

Zhou, Y. L., Wu, J. J., Gong, G. R., Liu, M., Li, Z., Guo, X. F., et al. (2023). Barbel regeneration and function divergence in red-tail catfish (*Hemibagrus wyckioides*) based on the chromosome-level genomes and comparative transcriptomes. *Int. J. Biol. Macromol.* 232, 123374. doi:10.1016/j.ijbiomac.2023.123374

Zhu, H. P., and Gui, J. F. (2007). Identification of genome organization in the unusual allotetraploid form of *Carassius auratus* gibelio. *Aquaculture* 265 (1), 109–117. doi:10.1016/j.aquaculture.2006.10.026

Zhu, C., Pan, Z., Yao, Y., Wu, D., Yuan, M., Gan, L., et al. (2021). Progress of karyotype studies on Siluriformes in China. *J. Fish. China* 46 (4), 657–663. (in Chinese). doi:10.12122/j.issn.1673-4254.2021.05.04

Zhu, C. K., Zheng, Y. H., Gao, Y. Y., and Li, Y. J. (2007). Biological characteristics and prospects for captive aquaculture of the *Mystus macropterus*. *Reserv. Fish.* 153, 86–89. (in Chinese).

Zhu, C., Liu, H., Pan, Z., Cheng, L., Sun, Y., Wang, H., et al. (2022). Insights into chromosomal evolution and sex determination of *Pseudobagrus ussuriensis* (Bagridae, Siluriformes) based on a chromosome-level genome. *DNA Res.* 29 (4), dsac028. doi:10.1093/dnares/dsac028

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 30 November 2023
DOI 10.3389/fvets.2023.1296610



OPEN ACCESS

EDITED BY
Ana Fabrícia Braga Magalhães,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Mohammad Hossein Banabazi,
Swedish University of Agricultural Sciences, Sweden
Szilvia Kusza,
University of Debrecen, Hungary
Pamela Burger,
University of Veterinary Medicine Vienna, Austria

*CORRESPONDENCE
Mohammad Al Abri
abri1st@squ.edu.om
Hussain Bahbahani
hussain.bahbahani@ku.edu.kw

RECEIVED 18 September 2023
ACCEPTED 17 November 2023
PUBLISHED 30 November 2023

CITATION
Al Abri M, Alfoudari A, Mohammad Z, Almathen F, Al-Marzooqi W, Al-Hajri S, Al-Amri M and Bahbahani H (2023) Assessing genetic diversity and defining signatures of positive selection on the genome of dromedary camels from the southeast of the Arabian Peninsula.
*Front. Vet. Sci.* 10:1296610.
doi: 10.3389/fvets.2023.1296610



# Assessing genetic diversity and defining signatures of positive selection on the genome of dromedary camels from the southeast of the Arabian Peninsula

Mohammad Al Abri[1]*, Ahmad Alfoudari[2], Zainab Mohammad[2], Faisal Almathen[3,4], Waleed Al-Marzooqi[1], Salim Al-Hajri[5], Mahmood Al-Amri[5] and Hussain Bahbahani[2]*

[1]Department of Animal and Veterinary Sciences, Sultan Qaboos University, Muscat, Oman, [2]Department of Biological Sciences, Faculty of Science, Kuwait University, Safat, Kuwait, [3]Department of Veterinary Public Health and Animal Husbandry, College of Veterinary Medicine, King Faisal University, Al-Ahsa, Saudi Arabia, [4]Camel Research Center, King Faisal University, Al-Ahsa, Saudi Arabia, [5]Laboratories and Research Administration, Directorate General of Veterinary Services, Royal Court Affairs, Muscat, Oman

Dromedary camels (*Camelus dromedarius*) are members of the Camelini tribe within the Camelidae family. They are distributed throughout North Africa, the Arabian Peninsula and Southeast Asia. This domestic species is characterized by its superior adaptability to the harsh desert environment. In this study, whole autosomal data of 29 dromedary samples from the Southeast Arabian Peninsula in Oman; 10 from Muscat, 14 from Al-Batinah, and 5 from Al-Sharqiya, were investigated to assess their genetic relationship and to define candidate signatures of positive selection. A minimal genetic distinction that separates Muscat dromedaries from the other two populations was observed, with a degree of genetic admixture between them. Using the de-correlated composite of multiple signals (DCMS) approach, a total of 47 candidate regions within the autosomes of these dromedary populations were defined with signatures of positive selection. These candidate regions harbor a total of 154 genes that are mainly associated with functional categories related to immune response, lipid metabolism and energy expenditure, optical and auditory functions, and long-term memory. Different functional genomic variants were called on the candidate regions and respective genes that warrant further investigation to find possible association with the different favorable phenotypes in dromedaries. The output of this study paves the way for further research efforts aimed at defining markers for use in genomic breeding programs, with the goal of conserving the genetic diversity of the species and enhancing its productivity.

KEYWORDS

**de-correlated composite of multiple signals, environmental adaptation, signatures of selection, genetic diversity, dromedary camels**

## Introduction

Dromedary camels are considered an integral part of both Bedouin and non-Bedouin societies in the Arabian Peninsula. They were indispensable means of transportation, with which owners covered vast areas of the desert for commuting and trading. These animals are patient,

sturdy, able to withstand harsh environmental conditions, and can navigate through treacherous sand storms. Settlers utilized them for land plowing and transporting heavy loads over long distances. In both societies, camels provided meat, milk, hide, and a living asset that can be cashed in times of need (1). Camels still provide meat and milk today due to increasing demand for both products. However, as lifestyles and priorities in the nations of the Arabian Peninsula slowly changed with modernity, other uses for camels emerged (2). Historically, camel racing competitions have been well-known in the Arabian Peninsula. However, in modern times, racing and beauty competitions have attracted significantly more participants (3). As a result, camel owners have increasingly focused on selecting animals that excel in both racing and beauty competitions over the last few decades.

Oman's weather is characterized by warm and dry conditions with average temperatures ranging between 16°C (in winter) and 31°C (in summer) but can soar up to 50°C during the summer (4). The exception to this weather pattern is in the southern region of Dhofar, which is affected by monsoon winds during the summer, locally known as Al-Khareef. These winds give rise to dazzle fog and occasional rain, resulting in lush vegetation and pastures (5). The Omani coastline spans over 2,000 km along the Arabian Sea and the Gulf of Oman, leading to an increase in relative humidity throughout the year in areas close to coastline (4). The average annual rainfall ranges from 50 mm in plain areas to over 300 mm in the southern region, with an overall average of about 100 mm per year (6).

Oman harbors approximately 280,000 dromedary camels, mainly used for milk production, racing or beauty competitions (7). As observed by Al Askar et al. (8), Omani camels are genetically distinct and have little to no admixture with other camel populations in the region. The majority of Omani camels, about 60%, are found in the southern region of Oman in the Governorate of Dhofar (9). It is believed that they have descended from the southern part of the Arabian Peninsula (10) and are highly regarded as some of the finest camels in the region (9), often prized for their racing abilities and stamina (8).

Assessing the genetic diversity and relationship of different dromedary camel populations was confined previously to using autosomal microsatellites (8, 11) and mitochondrial DNA markers (10, 12). Recent efforts have also utilized genotyping-by-sequencing (GBS) (13) and whole genome sequence data (14). A recent study by Almathen et al. (11) differentiated dromedary camel populations from the Arabian Peninsula into three groups utilizing 17 microsatellite loci. These groups correspond to three geographical locations: (1) North, Central, and West, (2) Southwest, and (3) Southeast. Similarly, Bahbahani and Almathen (14) observed the same geographic genetic distinctions based on whole genome sequence analyses.

Signatures of selection analyses have been performed on different livestock species, such as cattle (15–17), sheep (18, 19) and goat (20, 21), to determine candidate regions and genes associated with different favorable traits. Recent releases of the dromedary draft reference genome starting with the Arabian Peninsula (22), followed by the North African dromedary genome draft (23), and recently the chromosome-level draft by Elbers et al. (24) have all encouraged scientists to investigate the genome of dromedary camels in search for natural selection footprints (13, 25, 26). Releasing the first draft of the dromedary camel genome in 2014 revealed several gene ontologies to be under adaptive evolution, such as fat and water metabolism, response to heat stress, and salt metabolism (22). Bahbahani et al. (13) investigated the genome of dromedary racing and packing camels from Sudan for signatures of selection using genotyping-by-sequencing data where they found natural selection signals on genes associated with energy homeostasis, chondrogenesis, milk content, and immune response. Recently, a study by Khalkhali-Evrigh et al. (26) on Iranian dromedary camels also defined genes related to energy metabolism, reproduction, and long-term memory to be under natural positive selection.

Previously mentioned studies on dromedary camels relied on separate single statistical tests to detect signatures of selection. However, to improve the accuracy and resolution of detecting selection signatures, several composite analyses have been proposed that combine the signals of different statistics. These include Composite of Multiple Signals (CMS) (27), Meta-analysis of Selection Signals (28), and Composite Selection Signals (CSS) (29). While these approaches were successfully used to define candidate regions under selection in humans (27) and cattle (30), they either require accurate demographic models, as in the case of CMS, which was investigated by Fitak et al. (31), or they do not account for the covariance structure of the different single statistics employed. To address these limitations, Ma et al. (32) introduced a new approach called the de-correlated composite of multiple signals (DCMS). Compared to meta-SS and CSS, DCMS has generally shown higher power in detecting selection signatures. This approach has been previously employed to look for signatures of selection in Swedish cattle breeds (33), Russian cattle breeds (34), Russian sheep breeds (35), and Welsh sheep breeds (36).

In this study, the genetic diversity and relationship of Omani dromedary camels from the southeast of the Arabian Peninsula were assessed using whole genome sequence data. Signatures of natural positive selection were also investigated in the genome of Omani dromedary camels using the DCMS approach to defined candidate regions and genes under natural selection.

## Materials and methods

### Dromedary samples whole genome sequence data

Twenty hair bulbs were collected from each of 29 dromedary camels selected from the southeast of the Arabian Peninsula in the north of Oman: 10 from Muscat, 14 from Al-Batinah and 5 from Al-Sharqiya governorates. The samples were collected from different owners to avoid a close relationship. Hair bulbs genomic DNA was extracted using Gentra DNA purification kit as in Cook et al. (37), and sequenced using 150 bp paired-end libraries on an Illumina Hiseq 2000 platform at IGA Technology Services (Udine, Italy).

### Whole genome sequence data processing and variants calling

The adaptor-free sequence reads were mapped against the Arabian dromedary camel reference genome assembly (GCF_000803125.2)

TABLE 1 Quality control criteria and the number of excluded and remaining SNPs for the genetic diversity and signatures of selection analyses.

| Dataset | | Number of SNPs | |
|---|---|---|---|
| Raw autosomal SNPs | | 5,099,313 | |
| Quality control criteria | | Number of excluded SNPs | |
| Diversity analyses | Signatures of selection analyses | Diversity analyses | Signatures of selection analyses |
| Genotypic call rate < 100% | Genotypic call rate < 100% | 851,275 | 851,275 |
| MAF ≤ 5% | MAF ≤ 5% | 1,039,085 | 1,039,085 |
| HWE ($p$-value $<1\times10^{-6}$) | HWE ($p$-value $<1\times10^{-6}$) | 70,023 | 70,023 |
| Linkage disequilibrium ($R^2>0.5$) | | 2,930,406 | |
| Final number of SNPs | | 208,524 | 3,138,930 |

using the *bwa-mem* algorithm of Burrows-Wheeler Aligner (BWA) version 0.7.17 (38). Reads were sorted by coordinates using the *SortSam* algorithm and SORT_ORDER = coordinate option, and duplicates were marked and excluded using *MarkDuplicates* algorithm and REMOVE_DUPLICATES = true option in Picard tools version 3.0.0.[1] Summary statistics calculated for mapped reads included: the proportion of reference genome covered, mean depth of coverage, and percentage of mapped reads via the *coverage* and *flagstat* tools in SAMTools software version 1.13 (39).

Single Nucleotide Polymorphisms (SNPs) were called across all samples using the *HaplotypeCaller* tool in *GVCF* mode of GATK version 4.2.5.0 (40). The variants were subsequently combined and genotyped using the GATK *CombineGVCFs* and *GenotypeGVCFs* tools resulting in a total of 8,010,983 SNPs. After selecting autosomal SNPs, variants were hard-filtered using the Var*iantFiltration* tool of GATK to exclude: (1) variants with high probability strand bias between reference and alternate alleles (FS > 60); (2) variants with low quality by depth (QD < 2); (3) variants with a low root mean square mapping quality (MQ < 40); (4) variants with low phred-scaled variant probability (QUAL < 30); (5) variants with strand bias in mapping quality between reads supporting reference or alternate alleles (MQRankSum < −12.5); and (6) variants where the position of the alternate allele exhibits a bias toward the ends of the reads (ReadPosRankSum < −8). SNPs with a depth of coverage ranging between two reads and three standard deviations from the mean depth of coverage across samples were retained to end up with a total of 5,099,313 autosomal SNPs.

## SNPs quality control pruning

The retained autosomal SNPs underwent two separate quality control pruning criteria using PLINK v1.9 (41) for each of the genetic diversity and signature of selection analyses. For the diversity analysis, autosomal SNPs were pruned if: (1) their call rate was <100% of the genotyped samples; (2) they departed from the Hardy–Weinberg equilibrium ($p$-value $<1\times10^{-6}$); or (3) they had a minor allele frequency (MAF) ≤5%. Linkage disequilibrium pruning was also implemented using the PLINK option (*--indep-pairwise 50 10 0.5*), as in Ming et al. (42), to exclude SNPs with a correlation coefficient ($r^2$) > 0.5. Average $r^2$ was calculated between SNP pairs using the PLINK option (--r2 --ld-window 1000000 --ld-window-kb 2000 --ld-window-r2 0.09) (Supplementary Figure S1). The same quality control criteria were applied to SNPs for the signatures of selection analyses except for linkage disequilibrium, which is considered as a signal of selection. The final number of SNPs for the genetic diversity and signatures of selection analyses remaining were 208,524 and 3,138,930, respectively (Table 1). Samples were filtered out if they had a genotyping call rate <100% or a maximum pairwise identity-by-state (IBS) ≥95%, in which case the sample with the lower call rate was excluded. No samples were excluded due to these criteria.

## Genetic diversity analyses

Observed homozygosity and inbreeding coefficient (*Fis*) were computed for the different dromedary camel populations using the *hom* function of the GenABLE package (43) in R software version 4.1.0 (44). Two-sample Mann–Whitney U test was used to test for statistically significant differences in the homozygosity and *Fis* values between different dromedary populations. One-sample Mann–Whitney U test was used to check if the *Fis* values of each of the dromedary populations were significantly different from zero. Principal Component Analysis (PCA) was conducted on the filtered SNP data to determine the genetic relationship between the dromedary populations. *Prcomp* function implemented in R software was used to define the different principal components and the amount of variation explained by each component. The first two components were plotted using the ggplot2 package (45) of R software. Local ancestry proportions of the different dromedary samples were estimated using the ADMIXTURE 1.23 software (46). Ancestral cluster ($K$) values ranging from 1 to 3 were assumed to reflect the total number of populations in the dataset. A total of 200 bootstrap iterations were performed for each $K$ analysis. The $K$ cluster with the lowest cross-validation error (cv) was considered as the optimal number of clusters fitting the dataset.

## Signatures of selection analysis

The de-correlated composite of multiple signals (DCMS) approach (32) was used to detect signatures of selection on the autosomes of the dromedary camels. This approach employed four statistical tests: two allele-frequency spectrum-based statistics

1 http://broadinstitute.github.io/picard/index.html

[Tajima's D index (47) and nucleotide diversity (pi) (48)]; and two intra-population haplotype-based statistics [integrated haplotype score (iHS)] (49) and number of segregating sites by length (nSL) (50).

Tajima's D statistic was estimated in sliding 100 kb windows with a 25 kb step using the *Tajima* flag implemented in the VCF-kit tool version 0.2.6 (51). Nucleotide diversity (pi) was calculated for each chromosome separately on a per-site basis using the *--site-pi* function implemented in VCFtools version 0.1.16 (52). For calculating iHS and nSL scores, Beagle (53) was used in its default parameters for haplotype phasing to determine the individual haplotypes. After haplotype phasing, the iHS and nSL scores were obtained for every SNP using the selscan tool and default parameters (54). Before combining the four statistics, the nucleotide diversity, iHS, and nSL values were all standardized individually in sliding 100 kb windows with a 25 kb step and yielded mean values using an in-house R script.

In the final step, the 100 kb windows of the four statistics were all processed using the MINOTAUR package (55) in R software. After excluding windows with less than 10 SNPs, the results of the four statistics were converted to *p*-values based on fractional ranks using the *stat_to_pvalue* function. The *p*-values were then transformed to rank-based *p*-values based on either the one-tailed test (Tajima's D and pi - left-tailed) or two-tailed tests (iHS and nSL). Then, the correlation between the four statistics was calculated by constructing a covariance matrix using the *Cov-NAMcd* function with $\alpha = 0.75$ and $nsamp = 50{,}000$. This matrix was then used to adjust for correlation among the statistics and obtain the DCMS values concurrently with the *DCMS* function. *p*-values for the DCMS values were calculated using the *pnorm* function in R. Windows with $-\log_{10}$ (*p*-values) $\geq 4$, which is equivalent to *p*-values $\leq 0.0001$, were defined as candidate windows with signatures of selection. Finally, overlapping windows were merged into a single candidate region under selection.

## Functional annotation of the candidate signatures of selection regions

The coordinates of the candidate regions were cross-referenced against the dromedary camel reference genome assembly (GCF_000803125.2) genes list using the *GenomicRanges* package (56) in R. Functional profiling of the overlapping genes was conducted using the (g: GOSt) function of the *gProfiler* web server (57), which determined the functionally enriched terms for the Gene Ontology (GO), biological processes and molecular functions. The gprofiler (g: SCS) algorithm was used to compute multiple testing corrections for *p*-values from GO and pathway enrichment analyses. All the identified genes were also processed using the functional annotation enrichment tool implemented in *DAVID* Bioinformatics resources 6.7 (58) to determine enriched functional terms. An enrichment score of 1.3, which is equivalent to the Fisher exact test *p*-value of 0.05, was used as a threshold to define the significantly enriched functional terms in comparison to the dromedary reference genome background. The genes were then cross-referenced with literature to evaluate their biological functions. Genomic variants in the candidate regions were annotated with SnpEff software version 4.3 (59).

# Results

## Summary statistics of mapped sequence reads

The depth of coverage of the mapped sequence reads among the Omani dromedary samples ranged from 7.93X to 22.5X with a mean of 11.4X (Supplementary Table S1). On average, 99.8% of the sequence reads were mapped to the dromedary reference genome, out of which 96.2% of them were properly paired. These mapped reads covered an average of 94.8% of the reference genome (Supplementary Table S1).

## Genetic diversity analyses

The mean observed homozygosity values were estimated as 0.576 for Muscat, 0.575 for Al-Batinah, and 0.563 for Al-Sharqiya dromedaries, which were not significantly different from each other (*p*-value > 0.05). Negative *Fis* values were calculated for the different dromedary camels: −0.062 for Muscat, −0.062 for Al-Batinah, and −0.093 for Al-Sharqiya camels. The different *Fis* values were not significantly deviated from zero and did not show significant differences among each other (*p*-value > 0.05).

The principal component analysis showed a degree of genetic separation between the Omani camels from Muscat and the other Omani camels through the first principal component, which explained 4.7% of the total variation. Along this component, the Omani camels from Al-Sharqiya were slightly separated from Al-Batinah camels (Figure 1).

The optimal number of clusters determined by the admixture analysis was $K = 1$ (Supplementary Table S2). At $K = 2$, a genetic ancestry background related to Omani camels from Muscat was observed. A substantial degree of genetic admixture also observed among the Omani dromedaries analyzed at all *K* values tested (Figure 2).

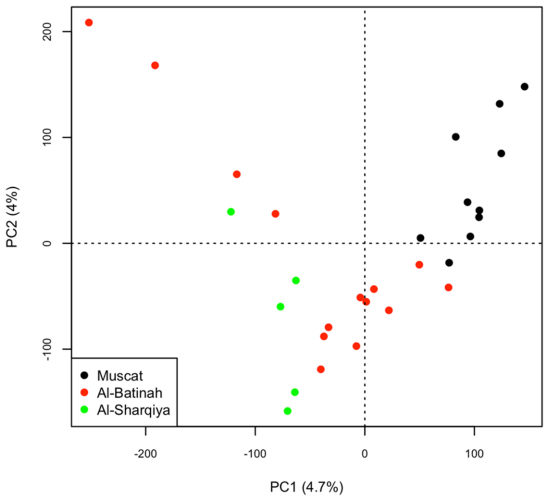



FIGURE 1
Principal Component Analysis (PCA) plot on the Muscat, Al-Batinah and Al-Sharqiya dromedaries.

10.3389/fvets.2023.1296610


FIGURE 2
Admixture analysis plots of $K = 2$ and 3 on the Muscat, Al-Batinah, and Al-Sharqiya dromedaries.

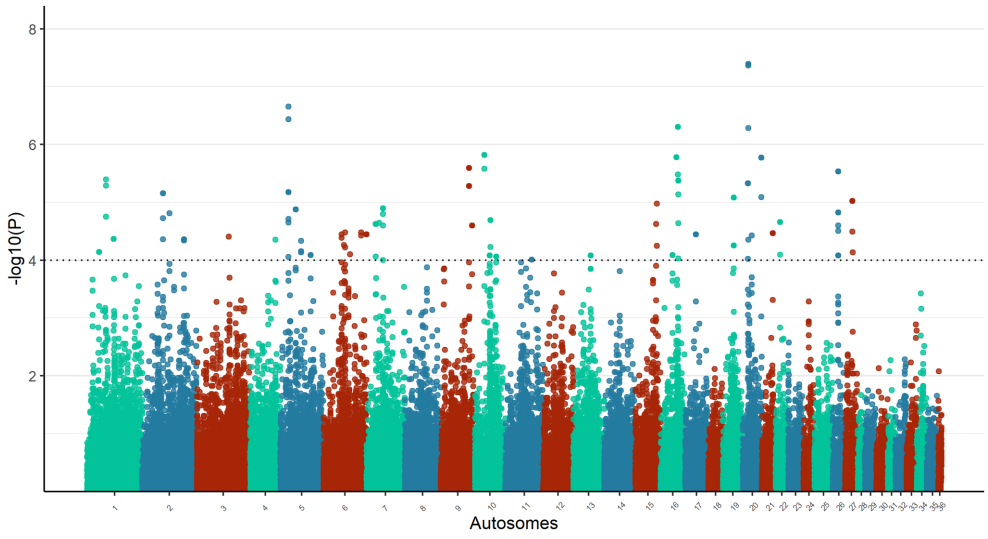



FIGURE 3
Manhattan plot of DCMS analysis on the autosomes of Omani dromedary camels. The horizontal line represents a significant $p$-value threshold of $1 \times 10^{-4}$.

## Signatures of selection analysis

The calculated DCMS values for the total 77,133 windows on the genome of the Omani dromedary camels ranged from −14.74 ($p$-value = 1) to 14.67 ($p$-value = $4.04 \times 10^{-8}$), in which a total of 87 candidate windows passed the threshold of $p$-value $< 1 \times 10^{-4}$ (Figure 3, Supplementary Table S3). The candidate windows were merged into 47 regions ranging in size from 100 kb to 225 kb with a mean size of 123.94 kb ± 32.54 kb. These regions were distributed among 20 autosomes with chromosome 6 having the highest number of regions ($n = 6$) (Supplementary Table S4). The largest candidate regions were in chromosomes 20 (117.5 Mb to 119.8 Mb) and 7 (209.0 Mb to 211.3 Mb), while the most significant candidate window was in chromosome 20 (118.3 Mb to 119.3 Mb) (Supplementary Table S4).

10.3389/fvets.2023.1296610

TABLE 2 Examples of candidate genes and their number of variants in different functional categories within the candidate regions.

| Functional category | Gene ID | Candidate region (Chr: start-stop) | Total number of variants |
|---|---|---|---|
| Immune response | *GMFG* | 9:64.00 Mbp-64.17 Mbp | 8 |
| | *IL17RD* | 17:23.72 Mbp-23.82 Mbp | 133 |
| | *SLC22A7* | 20:11.75 Mbp-11.97 Mbp | 40 |
| Lipid metabolism and energy expenditure | *SELENOV* | 9:64.00 Mbp-64.17 Mbp | 11 |
| | *STEAP2* | 7:20.90 Mbp-21.12 Mbp | 31 |
| | *NBEAL1* | 5:68.40 Mbp-68.50 Mbp | 109 |
| Optical function | *RLBP1* | 27:16.52 Mbp-16.67 Mbp | 47 |
| Long-term memory | *SYT3* | 9:70.90 Mbp-71.00 Mbp | 64 |
| | *TTBK1* | 20:11.75 Mbp-11.97 Mbp | 142 |
| Auditory function | *CRIP3* | 20:11.75 Mbp-11.97 Mbp | 32 |
| | *SLC22A7* | 20:11.75 Mbp-11.97 Mbp | 40 |
| Fertility | *CFAP69* | 7:20.90 Mbp-21.12 Mbp | 32 |
| | *SRF* | 20:11.75 Mbp-11.97 Mbp | 32 |

TABLE 3 Types, counts, and percentages of genomic variants called in the candidate regions.

| Type of variants | Count | Percentage |
|---|---|---|
| Intron variant | 17,514 | 59.50% |
| Intergenic region | 4,908 | 16.67% |
| Downstream gene variant | 2,967 | 10.08% |
| Upstream gene variant | 2,322 | 7.89% |
| Missense variant | 409 | 1.39% |
| 3-prime UTR variant | 361 | 1.23% |
| Synonymous variant | 358 | 1.22% |
| Noncoding exon variant | 239 | 0.81% |
| Intergenic variant | 133 | 0.45% |
| 5-prime UTR variant | 117 | 0.40% |
| 5-prime UTR premature start codon gain variant | 40 | 0.14% |
| Splice region variant. Intron variant | 35 | 0.12% |
| Nonsense variant | 16 | 0.05% |
| Missense variant. Splice region variant | 7 | 0.02% |
| Splice region variant | 6 | 0.02% |
| Splice acceptor variant. Intron variant | 1 | 0.01% |
| Splice region variant. Noncoding exon variant | 1 | 0.01% |
| Splice region variant, synonymous variant | 1 | 0.01% |

## Functional profiling of the candidate regions

A total of 154 genes were found within the 47 candidate regions relating to different functional categories, such as immune response, lipid metabolism and energy expenditure, optical function, long-term memory, and fertility (Table 2, Supplementary Table S5). The functional profiling of the defined genes indicated different functionally enriched molecular and biological processes relating to molecular adaptor activity, dopamine signaling, olfactory behavior, and Wnt signaling (Supplementary Table S6). Based on the DAVID analysis, eight functional clusters were identified, showing enrichment for functions relating to transcription (enrichment score = 1.04), diseases (enrichment score = 0.8), Wnt signaling pathway (enrichment score = 0.66), beta-transducing repeats (enrichment score = 0.58), acidic and basic amino acids (enrichment score = 0.44), extracellular regions (enrichment score = 0.36), immunoglobulin domains (enrichment score = 0.35), and transmembrane helices (enrichment score = 0.26) of which none were significantly enriched (Supplementary Table S7). A total of 12,089 variants were called on the defined candidate regions, which were classified into different types and the majority of them being intronic variants (59.5%), followed by intergenic region variants (16.67%), downstream gene variants (10.08%) and upstream gene variants (7.89%) (Table 3, Supplementary Table S8).

# Discussion

The analyzed Omani dromedary camel populations showed a minimal degree of genetic distinction mainly related to their distinct geographical origins: Muscat, Al-Batinah, and Al-Sharqiya. Such genetic distinction might also be related to the different types of dromedaries analyzed, as dromedary camels from Muscat are known as production camels while camels from Al-Batinah and Al-Sharqiya are non-production type, i.e., beauty and racing types. According to camel owners, Omani production camels are rarely interbred with beauty or racing camels. Conversely, limited interbreeding occurs between beauty and racing camels. This type-wise genetic differentiation needs to be further validated by sequencing more dromedary samples from these different classifications.

A substantial level of genetic introgression was also observed among the Omani dromedary populations, which may reflect the historical use of dromedaries in transportation and trading linking the different parts of the Arabian Peninsula through the "Incense Route" (60). Similar genetic phenomena have also been observed in dromedary camels from the Arabian Peninsula based on autosomal microsatellite data (11), whole genome sequence data (14), and in African dromedary camels (13). The practice of outbreeding among camel owners, characterized by random breeding between different camel populations, could also contribute to the observed genetic admixture. This is supported by the mean negative *Fis* values calculated for the different dromedary populations analyzed.

Several candidate genome regions were defined with signatures of positive selection based on the DCMS approach on Omani dromedary camels. This approach has the advantage of combining the signals of the different signature of selection statistics and outperforms the power of any single statistic (32). The use of such composite statistics, as seen with meta-SS, CMS, and CSS, will improve the resolution of localizing the selection hotspot (27, 28, 32).

The defined candidate regions harbor genes mainly related to different signaling pathways, such as Wnt signaling and dopamine signaling, immunity, hematopoiesis, fat metabolism and energy expenditure, and thermoregulation. These biological functions might be related to the adaptations of dromedary camels within their surrounding environment and habitat. The Wnt signaling pathway, a functionally enriched pathway identified among the genes in the

candidate regions, has a main role in myogenesis. The canonical and non-canonical Wnt signaling pathways are involved in regulating the differentiation of muscle stem cells and the growth of skeletal muscle fibers, respectively, which have a significant impact on the ability of dromedary camels to endure long-distance transportation along the sandy deserts (61).

Genes with biological roles associated with immunity, such as glia maturation factor gamma (*GMFG*) interleukin 17 receptor D (*IL17RD*), and solute carrier family 22 member 7 (*SLC22A7*) were found within three candidate regions on chromosomes 9, 17, and 20, respectively. These biological functions are considered natural selection hotspots in dromedaries, as they enable them to tolerate infections and pathogen loads in the desert environment. The *GMFG* gene plays an important role in maintaining effective cellular immunity. The encoded protein is a component of T-lymphocytes pseudopodia and is hence involved in regulating their migration (62). Interleukin 17 receptor D is involved in controlling inflammation upon mediating IL-17A-induced proinflammatory gene expression as observed in keratinocytes and psoriasis-like skin inflammation (63). *IL17A* has also been found to play a role in tissue repair by enhancing cellular adaptation to chronic hypoxia upon activating hypoxia-inducible factor 1a (HIF1a) pathway in epithelial cells (64). *SLC22A7* is an organic anion transporter involved in regulating blood uric acid level by the renal systema (65). Uric acid plays a role in triggering interleukin-mediated inflammation and induction to type 2 immune response (66).

Energy and fat metabolism, as well as visual system, were found to be under adaptive evolution in camelids by Wu et al. (22). Genes related to these functional categories have also been found in the defined candidate regions in this study. The Selenoprotein V (*SELENOV*) gene in the candidate region (9: 64–64.17 Mb) is an example of a gene involved in fat metabolism and energy expenditure. In a mouse knockout experiment by Chen et al. (67), high-fat mass accumulation and a decrease in energy release have been observed upon *SELENOV* depletion. *STEAP2* gene in the candidate region (7: 20.9–21.1 Mb) is another example of a gene related to fat metabolism. This gene has been found to be significantly up-regulated in myogenic precursors, which differentiates skeletal muscle fibers. The balance between myogenesis and adipogenesis during skeletal muscle development is related to intramuscular fat content (68). Neurobeachin like 1 (*NBEAL1*) gene in the candidate region (5: 68.4–68.5 Mb) is involved in cholesterol metabolism. The encoded *NBEAL1* protein regulates the expression of low-density lipoprotein (LDL) receptors, which are required to uptake extracellular cholesterol from LDL upon controlling the sterol regulatory element-binding protein 2 (*SREBP2*) processing (69). *NBEAL1* has also been involved in regulating body temperature in cattle, as highlighted in by Howard et al. (70). Interestingly, this gene has been considered as a candidate of selection pressure in locally adapted Mediterranean sheep (71), Ethiopian sheep (72), and Ugandan goats (73). Long-term ultraviolet radiation exposure is a selective force facing dromedary camels in the desert environment, which may target genes related to photoreception, such as the *RLBP1* gene found in the candidate region (27: 16.5–16.67 Mb). This gene codes for the cellular retinaldehyde-binding protein (*CRALBP*) in the retinal pigment epithelial and Muller cells of the retina. *CRALBP* is involved in the visual cycle (retinoid cycle) for continued light detection by rod and cone photoreceptor cells (74, 75).

Dromedary camels are characterized by a long-term memory that was commonly used to remember the routes of long journeys, hence genes related to neural development and differentiation may be considered as targets of selection. Synaptotagmin 3 (*SYT3*) gene was found in a candidate region on chromosome 9 (9: 70.9–71 Mb). The synaptotagmin 3 protein is an integral membrane protein localized in the postsynaptic endocytic zone of neurons. This protein is important to promote forgetting, which is crucial to maintain an acquired memory in changing environments. A study conducted by Awasthi et al. (76) demonstrated a lack of forgetting ability in *SYT3* knock-out mice. Tau-tubulin kinase 1 (*TTBK1*) found in the candidate region (20: 11.75–11.97 Mb) is another example of a gene related to neural development and memory. This gene has been found to encode neuron-specific protein that regulates the phosphorylation of the tau protein. Tau protein hyperphosphorylation and aggregate formation are correlated to various neurodegenerative disorders, such as Alzhiemer's disease, and dementia across several species (77).

The desert environment has a significant impact on the physical properties of sound, with factors such as humidity-related attenuation and sound propagation altering the way that sound travels. Unfortunately, these changes can also make dromedary camels vulnerable to abiotic noise, especially wind noise, which can disrupt their auditory awareness and balance (78). Genes related to hearing function have also been found in this study, particularly the genes *CRIP3* and *SLC22A7* both found in chromosome 20 (20: 11.75–11.97 Mb). Variants on the genes *CRIP3*, which plays a role in T-cell proliferation and metal-ion binding, and *SLC22A7*, have been found to be associated with hearing loss in humans (79).

A major challenge facing dromedary camels is their reproductive fitness in such a harsh desert environment. Therefore, genes involved in spermatogenesis, such as the *CFAP69* gene in the candidate region (7: 20.9–21.1 Mb), are a target of natural selection to maintain optimal sperm motility. This gene, which codes for cilia and flagella-associated proteins, has been found to be associated with multiple morphological abnormalities of sperm flagella (MMAF) syndrome and consequently male infertility. Knocking out this gene in mice revealed severe disruption to the sperm flagellum structure (80). The serum response factor (*SRF*) gene in the candidate region (20: 11.75–11.97 Mb) may also be another target of selection. This gene has been found to play a role in early embryonic development and its knockout leads to embryonic lethality by mid-gestation (81).

Several uncharacterized loci were found within the defined candidate regions, more specifically in chromosome 6 (6:38.57–38.7 Mb), which is one of the most significant regions defined in this study. These loci need to be further investigated to determine their biological roles that might be related to the dromedaries' adaptability. In addition to these loci, the most recent draft of the dromedary genome (CamDro3) contains 21,032 scaffolds unmapped to the autosomes or sex chromosomes and consequently were excluded from the analyses here. The genes on these scaffolds can be considered as potential targets of selection. Further sequencing efforts are required to map these DNA elements into their corresponding chromosome to be investigated for signatures of selection.

Different classes of genomic variants have been identified in the candidate regions, in which some of them are associated with functional impacts: such as amino acid changes, i.e., missense variants; truncated proteins, i.e., nonsense variants; or modifications in gene expression levels, i.e., upstream gene variants. The significance of these variants warrants further investigation that includes specific phenotypic data to find possible associations via genome-wide association analyses.

## Conclusion

Here, we have investigated, for the first time, the whole autosome of dromedary camels from southeast of the Arabian Peninsula to assess their genetic diversity and search for genomic signatures of selection. A degree of geographical-wise genetic distinction with a substantial level of introgression has been observed among the dromedary populations. Based on the de-correlated composite multiple signals (DCMS) approach, candidate regions and genes with signatures of positive selection were defined. Different environmental data; such as humidity, pathogens load, temperature, altitude and salinity, are needed to investigate possible correlations between these selection signatures and the dromedaries' adaptability to their environment. Moreover, such discoveries need to be further validated by including more diverse dromedary populations from different geographical regions and habitats in the Arabian Peninsula. An improvement of the reference genome draft would serve as an advantage to similar analyses whereby the functionality of the unplaced scaffolds can be further explored. This study represents the first milestone in developing genomic markers that can be used in designing genomic-informative breeding programs to conserve the genetic diversity of this well-adapted species and improve its productivity.

## 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 here: https://www.ebi.ac.uk/ena, PRJEB67314.

## Ethics statement

Ethical approval was not required for the study involving animals in accordance with the local legislation and institutional requirements because no approval was required as the research only required hair samples from the animals which were collected with minimum discomfort and with non invasively.

## Author contributions

MA: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. AA: Data curation, Formal Analysis, Investigation, Software, Writing – original draft. ZM: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing. FA: Conceptualization, Data curation, Investigation, Project administration, Writing – original draft, Writing – review & editing. WA-M: Funding acquisition, Writing – review & editing. SA-H: Validation, Writing – review & editing. MA-A: Funding acquisition, Writing – review & editing. HB: Conceptualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Ministry of Higher Education, Research and Innovation. Grant code: RC/RG-AGR/ANVS/19/01.

## Acknowledgments

We would like to thank the Royal Court Affairs and Royal Camel Corps for their support in sample collection and logistics of the project. We would also like to thank the camel owners for permission to sample their camels.

## 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



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2023.1296610/full#supplementary-material

## References

1. Al Abri MA, Faye B. Genetic improvement in dromedary damels: challenges and opportunities. *Front Genet*. (2019) 10:167. doi: 10.3389/fgene.2019.00167

2. Khalaf S. Poetics and politics of newly invented traditions in the gulf: camel racing in the United Arab Emirates. *Ethnology*. (2000) 39:243–61. doi: 10.2307/3774109

3. Al-Shorepy SS. Identification of environmental factors affecting the racing performance of race camels in the United Arab Emirates. *Emir J Food Agric*. (2011) 23:424–30.

4. Yousif JH, Al-Balushi HA, Kazem HA, Chaichan MT. Analysis and forecasting of weather conditions in Oman for renewable energy applications. *Case Stud Therm Eng*. (2019) 13:100355. doi: 10.1016/j.csite.2018.11.006

5. Tigani ElMahi A. Old ways in a changing space: the issue of camel pastoralism in Dhofar. *JAMS*. (2011) 16:51–64.

6. Al-Ajmi HA, Abdel Rahman HA. Water mangement intricacies in the Sultanate of Oman the augmenation—conservation conundrum. *Water Int*. (2001) 26:68–79. doi: 10.1080/02508060108686887

7. NCSI. *National Center for statistics & information* (2023).

8. AlAskar H, Alhajeri BH, Almathen F, Alhaddad H. Genetic diversity and population structure of dromedary camel-types. *J Hered*. (2020) 111:405–13. doi: 10.1093/jhered/esaa016

9. Shaat I, Al-Habsi R. Current status of animal genetic resources in Oman. *JASFT*. (2016) 2:139–46.

10. Almathen F, Charruau P, Mohandesan E, Mwacharo JM, Orozco-terWengel P, Pitt D, et al. Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *PNAS*. (2016) 113:6707–12. doi: 10.1073/pnas.1519508113

11. Almathen F, Bahbahani H, Elbir H, Alfattah M, Sheikh A, Hanotte O. Genetic structure of Arabian peninsula dromedary camels revealed three geographic groups. *Saudi J Biol Sci*. (2022) 29:1422–7. doi: 10.1016/j.sjbs.2021.11.032

12. Alaqeely R, Alhajeri BH, Almathen F, Alhaddad H. Mitochondrial sequence variation, haplotype diversity, and relationships among dromedary camel-types. *Front Genet*. (2021) 12:723964. doi: 10.3389/fgene.2021.723964

13. Bahbahani H, Musa HH, Wragg D, Shuiep ES, Almathen F, Hanotte O. Genome diversity and signatures of selection for production and performance traits in dromedary camels. *Front Genet*. (2019) 10:893. doi: 10.3389/fgene.2019.00893

14. Bahbahani H, Almathen F. Homogeneity of Arabian peninsula dromedary camel populations with signals of geographic distinction based on whole genome sequence data. *Sci Rep*. (2022) 12:130. doi: 10.1038/s41598-021-04087-w

15. Bahbahani H, Salim B, Almathen F, Al Enezi F, Mwacharo JM, Hanotte O. Signatures of positive selection in African Butana and Kenana dairy zebu cattle. *PLoS One*. (2018) 13:e0190446. doi: 10.1371/journal.pone.0190446

16. Tijjani A, Utsunomiya YT, Ezekwe AG, Nashiru O, Hanotte O. Genome sequence analysis reveals selection signatures in endangered Trypanotolerant west African Muturu cattle. *Front Genet*. (2019) 10:442. doi: 10.3389/fgene.2019.00442

17. Bahbahani H, Clifford H, Wragg D, Mbole-Kariuki MN, Van Tassell C, Sonstegard T, et al. Signatures of positive selection in east African shorthorn zebu: a genome-wide single nucleotide polymorphism analysis. *Sci Rep*. (2015) 5:11729. doi: 10.1038/srep11729

18. Fariello MI, Servin B, Tosser-Klopp G, Rupp R, Moreno C, San Cristobal M, et al. Selection signatures in worldwide sheep populations. *PLoS One*. (2014) 9:e103813. doi: 10.1371/journal.pone.0103813

19. Romanov MN, Abdelmanova AS, Fisinin VI, Gladyr EA, Volkova NA, Koshkina OA, et al. Selective footprints and genes relevant to cold adaptation and other phenotypic traits are unscrambled in the genomes of divergently selected chicken breeds. *J Anim Sci Biotechnol*. (2023) 14:35. doi: 10.1186/s40104-022-00813-0

20. Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, et al. Crepaldi, and c. the AdaptMap, signatures of selection and environmental adaptation across the goat genome post-domestication. *Genet Sel Evol*. (2018) 50:57. doi: 10.1186/s12711-018-0421-y

21. Wang J-J, Zhang T, Chen Q-M, Zhang R-Q, Li L, Cheng S-F, et al. Genomic signatures of selection associated with litter size trait in Jining gray goat. *Front Genet*. (2020) 11:286. doi: 10.3389/fgene.2020.00286

22. Wu H, Guang X, Al-Fageeh MB, Cao J, Pan S, Zhou H, et al. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun*. (2014) 5:5188. doi: 10.1038/ncomms6188

23. Fitak RR, Mohandesan E, Corander J, Burger PA. The *de novo* genome assembly and annotation of a female domestic dromedary of north African origin. *Mol Ecol Resour*. (2015) 16:314–24. doi: 10.1111/1755-0998.12443

24. Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, et al. Improving Illumina assemblies with hi-C and long reads: an example with the north African dromedary. *Mol Ecol Resour*. (2019) 19:1015–26. doi: 10.1111/1755-0998.13020

25. Bahbahani H, Al-Zoubi S, Ali F, Afana A, Dashti M, Al-Ateeqi A, et al. Signatures of purifying selection and site-specific positive selection on the mitochondrial DNA of dromedary camels (*Camelus dromedarius*). *Mitochondrion*. (2023) 69:36–42. doi: 10.1016/j.mito.2023.01.004

26. Khalkhali-Evrigh R, Hedayat N, Ming L, Jirimutu . Identification of selection signatures in Iranian dromedary and Bactrian camels using whole genome sequencing data. *Sci Rep*. (2022) 12:9653. doi: 10.1038/s41598-022-14376-7

27. Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*. (2010) 327:883–6. doi: 10.1126/science.1183863

28. Utsunomiya Y.T., Perez O'Brien A.M., Sonstegard T.S., Van Tassell C.P., A.S. do Carmo. Meszaros J. Solkner, Garcia J.F., Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One* 8 (2013):e64280, doi: 10.1371/journal.pone.0064280

29. Randhawa IA, Khatkar MS, Thomson PC, Raadsma HW. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet*. (2014) 15:34. doi: 10.1186/1471-2156-15-34

30. Bahbahani H, Tiijani A, Mukasa C, Wragg D, Almathen F, Nash O, et al. Signature of selection for environmental adaptation and zebu x taurine hybrid fitness in east African shorthorn zebu. *Front Genet*. (2017) 8:68. doi: 10.3389/fgene.2017.00068

31. Fitak RR, Mohandesan E, Corander J, Yadamsuren A, Chuluunbat B, Abdelhadi O, et al. Genomic signatures of domestication in Old World camels. *Commun Biol*. (2020) 3:316. doi: 10.1038/s42003-020-1039-5

32. Ma Y, Ding X, Qanbari S, Weigend S, Zhang Q, Simianer H. Properties of different selection signature statistics and a new strategy for combining them. *Heredity (Edinb)*. (2015) 115:426–36. doi: 10.1038/hdy.2015.42

33. Ghoreishifar SM, Eriksson S, Johansson AM, Khansefid M, Moghaddaszadeh-Ahrabi S, Parna N, et al. Signatures of selection reveal candidate genes involved in economic traits and cold acclimation in five Swedish cattle breeds. *Genet Sel Evol*. (2020) 52:52. doi: 10.1186/s12711-020-00571-5

34. Yurchenko AA, Daetwyler HD, Yudin N, Schnabel RD, Vander Jagt CJ, Soloshenko V, et al. Scans for signatures of selection in Russian cattle breed genomes reveal new candidate genes for environmental adaptation and acclimation. *Sci Rep*. (2018) 8:12984. doi: 10.1038/s41598-018-31304-w

35. Yurchenko AA, Deniskova TE, Yudin NS, Dotsev AV, Khamiruev TN, Selionova MI, et al. High-density genotyping reveals signatures of selection related to acclimation and economically important traits in 15 local sheep breeds from Russia. *BMC Genomics*. (2019) 20:294. doi: 10.1186/s12864-019-5537-0

36. Sweet-Jones J, Lenis VP, Yurchenko AA, Yudin NS, Swain M, Larkin DM. Genotyping and whole-genome resequencing of welsh sheep breeds reveal candidate genes and variants for adaptation to local environment and socioeconomic traits. *Front Genet*. (2021) 12:612492. doi: 10.3389/fgene.2021.612492

37. Cook D, Gallagher PC, Bailey E. Genetics of swayback in American Saddlebred horses. *Anim Genet*. (2010) 41:64–71. doi: 10.1111/j.1365-2052.2010.02108.x

38. Li H, Durbin R. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*. (2010) 26:589–95. doi: 10.1093/bioinformatics/btp698

39. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. (2011) 27:2987–93. doi: 10.1093/bioinformatics/btr509

40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. (2010) 20:1297–303. doi: 10.1101/gr.107524.110

41. Purcell NB, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. (2007) 81:559–75. doi: 10.1086/519795

42. Ming L, Yuan L, Yi L, Ding G, Hasi S, Chen G, et al. Whole-genome sequencing of 128 camels across Asia reveals origin and migration of domestic Bactrian camels. *Commun Biol*. (2020) 3:1. doi: 10.1038/s42003-019-0734-6

43. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. (2007) 23:1294–6. doi: 10.1093/bioinformatics/btm108

44. R-Core-Team. *R: A language and environment for statistical computing*. Vienna: R-Core-Team (2022).

45. Wickham H. *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag (2009).

46. 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

47. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. (1989) 123:585–95. doi: 10.1093/genetics/123.3.585

48. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. (1979) 76:5269–73. doi: 10.1073/pnas.76.10.5269

49. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. (2006) 4:e72. doi: 10.1371/journal.pbio.0040072

50. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol*. (2014) 31:1275–91. doi: 10.1093/molbev/msu077

51. Cook DE, Andersen EC. VCF-kit: assorted utilities for the variant call format. *Bioinformatics*. (2017) 33:1581–2. doi: 10.1093/bioinformatics/btx011

52. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. (2011) 27:2156–8. doi: 10.1093/bioinformatics/btr330

53. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. (2007) 81:1084–97. doi: 10.1086/521987

54. Szpiech ZA, Hernandez RD. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. (2014) 31:2824–7. doi: 10.1093/molbev/msu211

55. Verity R, Collins C, Card DC, Schaal SM, Wang L, Lotterhos KE. Minotaur: a platform for the analysis and visualization of multivariate results from genome scans with R shiny. *Mol Ecol Resour*. (2017) 17:33–43. doi: 10.1111/1755-0998.12579

56. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. (2013) 9:e1003118. doi: 10.1371/journal.pcbi.1003118

57. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Res*. (2019) 47:W191–8. doi: 10.1093/nar/gkz369

58. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists. *Nucleic Acids Res*. (2021) 50:W216–21. doi: 10.1093/nar/gkac194

59. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnpEff Fly*. (2012) 6:80–92. doi: 10.4161/fly.19695

60. Epstein H, Mason IL. *In evolution of domesticated animals*. New York: Longman Inc. (1984).

61. von Maltzahn J, Chang NC, Bentzinger CF, Rudnicki MA. Wnt signaling in myogenesis. *Trends Cell Biol*. (2012) 22:602–9. doi: 10.1016/j.tcb.2012.07.008

62. Lippert DND, Wilkins JA. Glia maturation factor gamma regulates the migration and adherence of human T lymphocytes. *BMC Immunol*. (2012) 13:21. doi: 10.1186/1471-2172-13-21

63. Girondel C, Meloche S. Interleukin-17 receptor D in physiology, inflammation and cancer. *Front Oncol*. (2021) 11:656004. doi: 10.3389/fonc.2021.656004

64. Wang J, Ding X. IL-17 signaling in skin repair: safeguarding metabolic adaptation of wound epithelial cells. *Signal Transduct Target Ther*. (2022) 7:359. doi: 10.1038/s41392-022-01202-9

65. Sato M, Mamada H, Anzai N, Shirasaka Y, Nakanishi T, Tamai I. Renal secretion of uric acid by organic anion transporter 2 (OAT2/SLC22A7) in human. *Biol Pharm Bull*. (2010) 33:498–503. doi: 10.1248/bpb.33.498

66. Ghaemi-Oskouie F, Shi Y. The role of uric acid as an endogenous danger signal in immunity and inflammation. *Curr Rheumatol Rep*. (2011) 13:160–6. doi: 10.1007/s11926-011-0162-1

67. Chen LL, Huang JQ, Wu YY, Chen LB, Li SP, Zhang X, et al. Loss of Selenov predisposes mice to extra fat accumulation and attenuated energy expenditure. *Redox Biol*. (2021) 45:102048. doi: 10.1016/j.redox.2021.102048

68. Qiu K, Zhang X, Wang L, Jiao N, Xu D, Yin J. Protein expression landscape defines the differentiation potential specificity of Adipogenic and myogenic precursors in the skeletal muscle. *J Proteome Res*. (2018) 17:3853–65. doi: 10.1021/acs.jproteome.8b00530

69. Bindesbøll C, Aas A, Ogmundsdottir MH, Pankiv S, Reine T, Zoncu R, et al. NBEAL1 controls SREBP2 processing and cholesterol metabolism and is a susceptibility locus for coronary artery disease. *Sci Rep*. (2020) 10:4528. doi: 10.1038/s41598-020-61352-0

70. Howard JT, Kachman SD, Snelling WM, Pollak EJ, Ciobanu DC, Kuehn LA, et al. Beef cattle body temperature during climatic stress: a genome-wide association study. *Int J Biometeorol*. (2014) 58:1665–72. doi: 10.1007/s00484-013-0773-5

71. Serranito B, Cavalazzi M, Vidal P, Taurisson-Mouret D, Ciani E, Bal M, et al. Local adaptations of Mediterranean sheep and goats through an integrative approach. *Sci Rep*. (2021) 11:21363. doi: 10.1038/s41598-021-00682-z

72. Edea Z, Dadi H, Dessie T, Kim K-S. Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes Genom*. (2019) 41:973–81. doi: 10.1007/s13258-019-00820-y

73. Onzima RB, Upadhyay MR, Doekes HP, Brito LF, Bosse M, Kanis E, et al. Genome-wide characterization of selection signatures and runs of homozygosity in Ugandan goat breeds. *Front Genet*. (2018) 9:318. doi: 10.3389/fgene.2018.00318

74. Saari JC, Nawrot M, Kennedy BN, Garwin GG, Hurley JB, Huang J, et al. Visual cycle impairment in cellular retinaldehyde binding protein (CRALBP) knockout mice results in delayed dark adaptation. *Neuron*. (2001) 29:739–48. doi: 10.1016/S0896-6273(01)00248-3

75. Xue Y, Shen SQ, Jui J, Rupp AC, Byrne LC, Hattar S, et al. CRALBP supports the mammalian retinal visual cycle and cone vision. *J Clin Invest*. (2015) 125:727–38. doi: 10.1172/JCI79651

76. Awasthi A, Ramachandran B, Ahmed S, Benito E, Shinoda Y, Nitzan N, et al. Synaptotagmin-3 drives AMPA receptor endocytosis, depression of synapse strength, and forgetting. *Science*. (2019) 363:eaav1483. doi: 10.1126/science.aav1483

77. Sato S, Cerny RL, Buescher JL, Ikezu T. Tau-tubulin kinase 1 (TTBK1), a neuron-specific tau kinase candidate, is involved in tau phosphorylation and aggregation. *J Neurochem*. (2006) 98:1573–84. doi: 10.1111/j.1471-4159.2006.04059.x

78. Ali S, Esmat A, Erasha A, Yasuda M, Alsafy M. Morphology and morphometry of the inner ear of the dromedary camel and their influence on the efficiency of hearing and equilibrium. *Zool Lett*. (2022) 8:12. doi: 10.1186/s40851-022-00196-0

79. Cornejo-Sanchez DM, Li G, Fabiha T, Wang R, Acharya A, Everard JL, et al. Rare-variant association analysis reveals known and new age-related hearing loss genes. *Eur J Hum Genet*. (2023) 31:638–47. doi: 10.1038/s41431-023-01302-2

80. Dong FN, Amiri-Yekta A, Martinez G, Saut A, Tek J, Stouvenel L, et al. Absence of CFAP69 causes male infertility due to multiple morphological abnormalities of the flagella in human and mouse. *Am J Hum Genet*. (2018) 102:636–48. doi: 10.1016/j.ajhg.2018.03.007

81. Holtz ML, Misra RP. Serum response factor is required for cell contact maintenance but dispensable for proliferation in visceral yolk sac endothelium. *BMC Dev Biol*. (2011) 11:18. doi: 10.1186/1471-213X-11-18

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 04 December 2023
DOI 10.3389/fvets.2023.1287676



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Mahak Singh,
The ICAR Research Complex for North Eastern Hill Region (ICAR RC NEH), India
Mustafa Hitit,
Kastamonu University, Türkiye

*CORRESPONDENCE
Bambang Purwantara
✉ purwantara@apps.ipb.ac.id
Zulfi Nur Amrina Rosyada
✉ nur.amrina@fkh.unair.ac.id

RECEIVED 02 September 2023
ACCEPTED 06 November 2023
PUBLISHED 04 December 2023

CITATION
Rosyada ZNA, Pardede BP, Kaiin EM, Gunawan M, Maulana T, Said S, Tumbelaka LITA, Solihin DD, Ulum MF and Purwantara B (2023) A proteomic approach to identifying spermatozoa proteins in Indonesian native Madura bulls.
*Front. Vet. Sci.* 10:1287676.
doi: 10.3389/fvets.2023.1287676



# A proteomic approach to identifying spermatozoa proteins in Indonesian native Madura bulls

Zulfi Nur Amrina Rosyada[1,2,3]*, Berlin Pandapotan Pardede[1,2], Ekayanti Mulyawati Kaiin[2], Muhammad Gunawan[2], Tulus Maulana[2], Syahruddin Said[2], Ligaya I. T. A Tumbelaka[1], Dedy Duryadi Solihin[4], Mokhamad Fakhrul Ulum[1] and Bambang Purwantara[1]*

[1]Division of Reproduction and Obstetrics, School of Veterinary Medicine and Biomedical Sciences, IPB University, Bogor, Indonesia, [2]Research Center for Applied Zoology, National Research and Innovation Agency (BRIN), Bogor, Indonesia, [3]Division of Veterinary Anatomy, Faculty of Veterinary Medicine, Universitas Airlangga, Surabaya, Indonesia, [4]Department of Biology, Faculty of Science, IPB University, Bogor, Indonesia

Proteins assist sperm mature, transit the female reproductive tract, and recognise sperm oocytes. Indigenous Indonesian bulls, Madura bulls, have not been studied for reproductive proteomics. As local Indonesian beef livestock, Madura cattle assist in achieving food security; hence, their number must be improved. Thus, the identification of molecular proteomics-based bull fertility biomarkers is needed. This study aimed to characterise the sperm fertility function of the superior Madura bull (*Bos indicus* × *Bos Javanicus*) spermatozoa proteome. Frozen semen from eight Madura superior bulls (*Bos indicus* × *Bos javanicus*) aged 4–8 years was obtained from the artificial insemination centre (AIC) in Singosari and Lembang. Madura superior bulls are those that have passed the bull breeding soundness evaluation. Frozen sperm were thawed and centrifuged at 3000 × g for 30 min. Proteins in sperm were characterised through proteomic analysis using liquid chromatography–tandem mass spectrometry (LC–MS/MS). The resulting gene symbols for each protein were then subjected to bioinformatics tools, including UniProt, DAVID, and STRING databases. Regarding sperm fertility, the analysis revealed that 15 proteins were identified in the sperm of Madura bulls. Amongst the identified proteins, the superior Madura bull sperm contained several motilities, energy-related proteins, and chaperone proteins. A substantial portion of characterised proteins are linked to metabolic pathways and the tricarboxylic acid (TCA) cycle, contributing to sperm energy production. In conclusion, the first in-depth proteome identification of sperm related to sperm quality and bull fertility of a unique indigenous Madura breed of Indonesia was performed using the LC–MS/MS proteomic method. These findings may serve as a reference point for further studies related to the functions of bovine sperm and biomarkers of fertility and sperm quality.

KEYWORDS

**fertility, food security, LC–MS/MS, Madura bulls, sperm proteins fertility, proteomic, sperm proteins**

# 1 Introduction

Recent developments in proteomics have significantly impacted our understanding of how sperm become fertile (1). One of the most differentiated cells is the sperm head, which has a highly compacted chromatin structure and an enormous midpiece that contains the machinery required to interact and transfer paternal genetic and epigenetic information to the oocyte (2). Due to their high level of differentiation, sperm is suitable for studying proteomic sections, such as the sperm membrane, which is the most crucial part because of its role in interacting with the environment and the oocyte (3). The fusion of a sperm and an oocyte requires complex membrane modifications of the sperm (4). Spermatozoa proteins that regulate normal/abnormal sperm function have been identified through proteomic investigation (5).

Thus, bull fertility is linked to many proteins involved in spermatogenesis. The significance of specific proteins in controlling sperm quality and fertilization is unknown, although their availability and quantity may alter sperm fertility (6). According to Peddinti et al. (1), many bovine fertility protein markers have been reported. Major semen proteins, such as binding sperm protein (BSP) in Frisian Holstein cattle (6), zona pellucida binding protein (ZPBP) in Bali Polled bulls (7), A-kinase anchoring protein 3 in buffalo (8), and osteopontin in Limousin cattle (9), have been widely reported. The sperm proteome profile of mammals has been studied, including pigs (10), equines (11), sheep (12), and bovines (1). Nevertheless, the sperm proteome remains unknown in Madura bulls.

Madura cattle (*Bos indicus* × *Bos javanicus*) are native to Madura, Indonesia. Small-scale producers raise Madura cattle as working cattle to maintain the regional culture, which still values Sonok and Karapan cattle (13). In addition, Madura cattle are currently being developed as beef cattle using artificial insemination and reproduction technology. It also ensures food security for the Indonesian population. In particular, these cattle are highly valued for their unique reproductive traits, with regular calving even under low-input regimes and in dry and arid regions (14). However, it remains unknown whether the proteome level of Madura bull sperm affects sperm function and fertility. Proteomics offers various methods. However, complex protein analysis using LC–MS has become the preferred analytical technique for quantitative proteomics (15, 16).

Ultimately, acquiring knowledge about the proteome of spermatozoa can provide a holistic understanding of reproductive processes, not only specific to the Madura bull breed but also applicable to bovine species. This study aims to provide a comprehensive profiling of the sperm proteins associated with fertility function in the superior Madura bull (*Bos indicus* × *Bos javanicus*) spermatozoa proteome. Additionally, we used the literature to explain how Madura bull sperm proteins affect bull fertility. Further, proteomic approaches have made it possible to identify proteins that have the potential to function as indicators of male fertility, as well as proteins that are involved in the functional characteristics of sperm.

# 2 Materials and methods

## 2.1 Frozen semen samples

A total of 8 superior Madura bulls (*Bos indicus* × *Bos javanicus*) aged 4 to 8 years with spermatozoa motility >70%, based on secondary data from each AIC in Lembang and Singosari, Indonesia, were included in this study. Frozen semen from each of the eight Madura superior bulls was obtained from the National AIC in Singosari and Lembang Bank of Semen. Madura superior bulls are those that have passed the bull breeding soundness evaluation. The Animal Care and Use Committee excluded this study from its ethical review because artificial vaginal semen collection did not alter the physiology of animals. The study was directed by veterinarians from both institutions and followed SNI ISO 9001: 2015 No. 824 100 16072 at Lembang AIC and SNI ISO 9001: 2015 No. G.01-ID0139-VIII-2019 at Singosari AIC. The ethics committees of Lembang AIC and Singosari AIC provided ethical guidance and sanctions for the responsible collection of bull sperm. Additionally, an experienced bull technician collected sperm using an artificial vagina.

Semen samples from Lembang AIC were processed according to the AIC's standard operating procedure (SOP), which used skim milk as the extender. The formulation of the skimmed milk diluent in a volume of 1000cc consisted of 100g of skimmed milk, 960cc of distilled water, and an antibiotic solution containing 3000000 IU of penicillin, 3g of streptomycin, and 30cc of distilled water. The ratio of skimmed milk diluent to antibiotic was 100:1. In contrast, the semen samples obtained from Singosari were cryopreserved utilising tris egg yolk as a cryoprotectant. The tris egg yolk extender was formulated using the following components: 20% egg yolk, 1.6% tris aminomethane, 1.4% lactose, 2.5% raffinose, 0.9% citric acid, and an antibiotic mixture consisting of 1000000 IU/L of penicillin, 1g/L of streptomycin, and 1g/L of distilled water. The AIC in Lembang and Singosari employ extenders consisting of skim milk and tris egg yolk supplemented with antibiotics. The extender utilised in every artificial insemination facility serves as a cryoprotectant, safeguarding sperm cells during the freezing process.

Additionally, it enhances semen volume, sustains sperm viability, and regulates sperm pH. Consequently, the inclusion of an extender is imperative in the manufacture of frozen semen to ensure the preservation of sperm quality. The cryopreserved sperm samples were transferred with a specialised transport container that maintained a temperature of −196°C through liquid nitrogen. The pieces were placed within the container for subsequent examination.

## 2.2 Sperm protein isolation

The collected semen samples of about ive straws of frozen semen from each of the eight bulls were washed in 2 mL of phosphate buffer saline and centrifuged twice (3000 × g for 10 min, 4°C) to clean and separate the spermatozoa, seminal plasma, and extender of frozen semen. The resulting sperm cell pellet was then resuspended in cell lysis buffer (2% SDS in 62.5 mM Tris–HCl, pH 6.8, 1.0 mM phenylmethanesulfonyl fluoride, and 23 mM benzidine as a protease inhibitor) and stored at −20°C for protein extraction. The sperm pellet was vortexed for 10 min before protein extraction. Next, sonication was used thrice for 20 s each to dissolve the protein. After centrifuging the cell lysates (10,000 × g, 10 min), protein lysates were isolated. The Bradford method (17) was used to estimate the total protein yield of the sperm protein lysate. The Bradford protocol was performed using the Coomassie protein kit instructions (Merck, Darmstadt, Germany). The data were processed using ThermoScan RE software version 3.2 Multiskan Go (Thermo Fisher Scientific, Waltham, Massachusetts, United States). Pools of sperm protein containing equal amounts from three samples were frozen at −20°C for later use. A summary of the experimental design is shown in Figure 1.

## 2.3 Fractionation by one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (1D-SDS–PAGE) and in-gel digestion

Following the separation of 35 µg of sperm protein (an equivalent quantity of protein amalgamated from three cases) by 12.5% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS–PAGE), the gel was stained with colloidal Coomassie brilliant blue R-250 (CAS number: 6104-59-2) from HiMedia. Subsequently, the gel bands were excised and digested using a previously described protocol (18). The proteins were reduced using 0.5 M dithiothreitol (DTT) at 55°C for 25 min, followed by alkylation with 14 mM iodoacetamide (IAA) for 40 min at room temperature without light. Gel fragments were treated with trypsin (Promega, Madison, Fitchburg, WI, USA) at a ratio of 1:50 (trypsin: protein) at 4°C for 30 min. The gel fragments were incubated for 18 h at 37°C in a ThermoMixer shaker incubator block (Eppendorf). The reaction in each vial was terminated by adding 1% trifluoroacetic acid (TFA). An extraction buffer (2% TFA in 20% acetonitrile) was used to extract peptide digests from the gel pieces, followed by 70% acetonitrile. The obtained peptide digests were subjected to a drying process and subsequently purified using a C18 spin column for desalination (Thermo Scientific, Pierce Biotechnology, N Meridian Rd., Rockford, IL, USA) and then preserved at −20°C until LC–MS/MS analysis.

## 2.4 Liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis

The peptides obtained from each sample were subjected to LC–MS/MS analysis using an Ultimate 3000 Nano LC system coupled with a Q-Exactive Plus Orbitrap HRMS system (Thermo Fisher Scientific, Bremen, Germany). The peptides were introduced onto a preanalytical column with dimensions of 75 µm ID, 15 cm length, and 100 pore size, packed with Acclaim PepMap C18 2 µm particles. The solvent used for loading was solvent A, consisting of 0.1% formic acid, and the flow rate was set at 300 nL/min. A Q-Exactive Plus Orbitrap mass analyser was used for subsequent analysis. The peptides were subsequently resolved on an analytical column (50 cm 75 m ID, Pep Map RSLC C18 2 m) at a flow rate of 300 nL/min for 120 min using an increasing gradient of 5–35% solvent B (98% acetonitrile and 0.1% formic acid). Peptide signals were collected using an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The MS spectra within the 200–2000 m/z range were obtained using an Orbitrap analyser with a resolution of 30,000 (at m/z 400). Subsequently, ten precursor MS scans were conducted using collision-induced dissociation fragmentation at 35% normalised collision energy (19).

## 2.5 Protein database searching and bioinformatics analysis

Protein analysis was performed on proteins that contained at least one unique peptide. The exclusion of specific peptides based on the "minimum two-peptide rule may result in the perpetual disregard of genuine peptides, and proteins must have a sequence score HT > 0. Thus, we manually confirmed single peptide spectral match (PSM)-identified protein MS/MS spectra. Protein data were acquired and analysed using the UniProt bovine protein database (http://www.uniprot.org) and Proteome Discover version 2.2 software (Thermo Fisher Scientific). Furthermore, the discovered proteins from spermatozoa were also put into functional groups based on Gene Ontology (GO) using the DAVID web resource (https://david.ncifcrf.gov/tools.jsp). The interactions between the proteins were obtained using STRING version 11.0 (https://string-db.org/).

# 3 Result

## 3.1 Protein characterisation of Madura bull spermatozoa

LC–MS/MS identified Madura bull sperm protein markers of fertility and environmental compatibility. All the databases contained

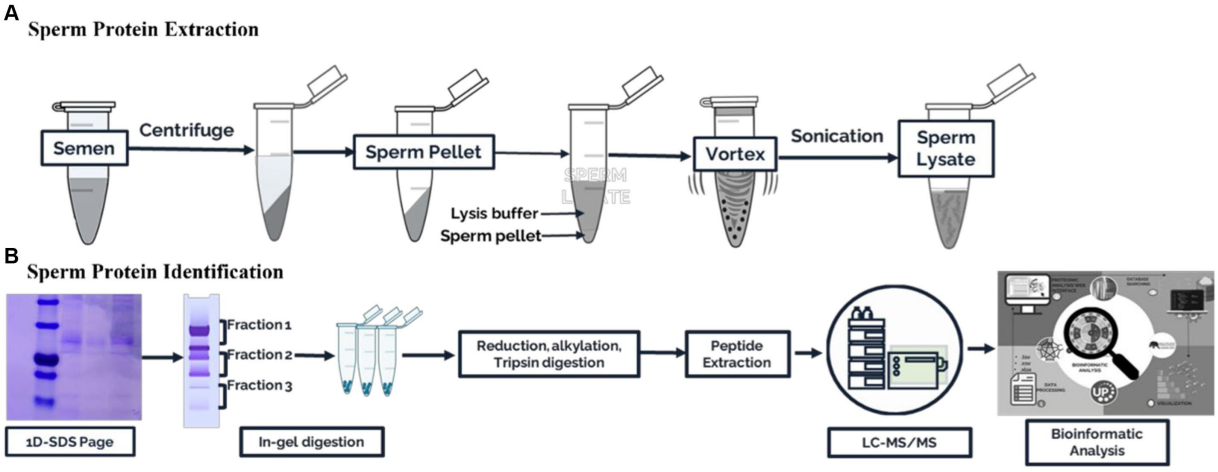



FIGURE 1
Schematic workflow for the proteomic analysis of Madura bull spermatozoa. **(A)** Sperm protein extraction. **(B)** Sperm protein identification.

in the database were only *Bos taurus* origin proteins since *Bos indicus* × *Bos javanicus* proteins are limited. This study analysis identified 15 proteins in Madura bull (*Bos indicus* × *Bos Javanicus*) sperm using UniProt and DAVID software (Table 1). Fifteen proteins were identified that were associated with sperm function categories (Table 1), including adenosine triphosphate (ATP) synthesis activity, cellular metabolic processes, cilia/flagella, sperm motility, capacitation and acrosome reaction, sperm-egg fusion, spermatogenesis and fertilization, and chaperone proteins. Functional annotation of proteins from Madura bull spermatozoa was further carried out using DAVID software and divided into several categories: "Biological Process" (BP), "Cellular Component" (CC), and "Molecular Function" (MF) (Figure 2). Although different categories were created for each division, the most numerous were those involved in the generation of precursor metabolites and energy (16%), urine ribonucleotide metabolic processes (16%), ATP metabolic processes (13%), and pyruvate metabolic processes (10%) in the case of BP (Figure 2A); cytoplasm (24%), supramolecular fibre (21%), and cell surface and microtubule (10%) in CC (Figure 2B); and ATP-binding proteins (16%), ATPase activity (12%) and unfolded protein binding (8%) in MF (Figure 2C).

## 3.2 Protein interaction networks and pathway enrichment analysis

All proteins identified in Madura bull spermatozoa were then searched using STRING software (version 11.0) for protein–protein interaction network analysis. Light blue lines show protein nodes, whereas pink lines reflect experimentally established interactions. The expected interactions are green for the gene neighbourhood, red for gene fusion, and dark blue for gene co-occurrence. Other protein relationships are shown by light green text mining, black co-expression, and blue protein homologies. A coloured node shows the query protein, and a white node leads the second shell of the interaction. Proteins with unknown 3D conformations had empty nodes, whereas those with known conformations had full nodes. The interactions between the 15 identified proteins are shown in Figure 3. Fifteen proteins were directly or indirectly connected through one or more interacting proteins, indicating functional links. We categorised proteins using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway terminology to study Madura bull sperm pathways. As expected, many of the discovered proteins were involved in metabolic pathways. The discovered proteins were associated with the TCA cycle (Figure 4). The sperm head, mitochondria (mid-piece), and flagellum/tail expressed the majority of these proteins (Figure 5).

# 4 Discussion

Fifteen proteins found in the sperm of Madura bulls were found to be related to fertility based on LC–MS/MS analysis (Table 1). The utilisation of known molecular weight standards for calibration is a fundamental aspect of size-exclusion chromatography that provides insights into the distribution of molecular weights. The current study has demonstrated that Madura bull sperm exhibits a molecular weight distribution of proteins within the 40 to 180 kDa range. This study utilised GO annotation to ascertain the cellular localisation, biological processes, and molecular functions of proteins specific to the sperm of Madura bulls.

## 4.1 Protein related to sperm fertility function

Effect of sperm motility on semen quality Male infertility is often attributed to reduced sperm motility. Therefore, the motility proteins of the Madura bull breed must be identified. We found many of these proteins in the spermatozoa. SPAM1 is a transmembrane protein present in ejaculated bull spermatozoa. It originates from the testis and occupies the anterior head region, where the orientation of the C-terminus with the zonal-binding domain facilitates interaction with zona pellucida after the acrosome reaction (20). This protein has also been detected in male sperm (21). The roteomee analysis carried out in this study from the sperm of Madura Bull also revealed the existence of the hyaluronidase protein, also known as SPAM1, isoform 62.4 kDa.

The flagellum of mammalian spermatozoa has intricate supplemental structures around the core axoneme, including the outer dense fibers (ODFs). ODFs protect the sperm tail against shear stress that may arise during epididymal transit and ejaculation. Four major proteins (ODF1, ODF2, ODF3, and ODF4) are amongst more than 14 polypeptides produced by mammalian ODFs. In mice, disrupting ODF2 expression lowers sperm motility, which is consistent with the characteristics of asthenozoospermia (22). Moreover, the cervical midpiece junction requires the presence of ODF2, which comprises part of the centrosome and is separated from the flagellum (23). Our analysis identified ODF2 in Madura bull spermatozoa.

In addition, we found that enolase protein, previously discovered in the plasma membrane of sperm from Bali bulls (24), was present in sperm samples from Madura bulls. During glycolysis and gluconeogenesis, enolase reversibly converts 2-phosphoglycerate into phosphoenolpyruvate. This enzyme produces sperm motility energy (25). Xi et al. (26) found a significant positive correlation between ENO3 and sperm motility parameters in sturgeons. Similarly, He et al. (25) discovered ENO3 in the midpiece and tail of ram sperm, which strongly correlated with sperm motility. The presence of ENO3 expression in Madura cattle sperm and its correlation with sperm function, particularly motility, can be readily explained.

The present study revealed the detectability of AKAP 4 in Madura bull sperm using LC–MS/MS. AKAP4, a marker involved in flagellar structure and motion, is associated with the equine sperm tail (27) and is substantially preserved in mice (28), bulls (29), and humans (30). AKAP4 attaches cAMP-dependent protein kinase A (PKA) to different subcellular locations. In the sperm fibrous sheath, AKAP4 binds PKA and interacts with other proteins to regulate motility. AKAP4, a critical fibrous sheath protein, bundles glycolytic enzymes and phosphorylation-signalling cascade components to supply a localised source of ATP and govern flagellar motion, sperm motility, and hypermotility. Many glycolytic enzymes closely linked to the fibrous sheath at the primary component of the flagellum are required for sperm motility (27).

Tubulins were also found to be another type of protein that contributes to sperm motility that was uncovered in this study. Tubulins are proteins found in the microtubules of sperm (30). Two tubulin proteins were identified as present in the sperm of the studied

10.3389/fvets.2023.1287676

TABLE 1 Proteins identified in Madura bull sperm associated with sperm function.

| Protein accession | Gene symbol | Protein name | Function | Molecular weight (kDa) |
|---|---|---|---|---|
| A0A4 W2HIH5 | AKAP4 | A-kinase anchoring protein 4 | flagellated sperm motility, motile cilium assembly | 99.6 |
| A0A4 W2BT96 | ATP5F1A | ATP synthase F1 subunit alpha | lipid metabolic process, ATP synthase activity | 63.1 |
| A0A452DII8 | ATP5F1B | ATP synthase F1 subunit beta | ATP synthesis activity, Hydrogen ion transport, Ion transport | 62.2 |
| A0A6P5BW43 | ACO2 | aconitase 2 | tricarboxylic acid cycle, citrate metabolic process | 85.3 |
| A0A4 W2CUL5 | COL4A2 | collagen type IV alpha 2 chain | cell membrane receptors, metabolic processes, spermatozoa differentiation, | 179.5 |
| A0A4 W2EVB5 | C1orf56 | Chromosome 1 open reading frame 56 | Capacitated sperm and acrosome reaction | 37.9 |
| A0A3Q1M0V5 | ENO3 | Enolase 3 | glycolytic process, pyruvate metabolism, ATP formation from ADP | 62.1 |
| A0A4 W2CEF3 | GAPDHS | glyceraldehyde-3-phosphate dehydrogenase, spermatogenic | glucose metabolism, glycolytic process, formation of ATP from ADP, spermatozoa energy precursor | 44.1 |
| A0A4 W2D4U4 | HSPA2 | heat shock protein family A (Hsp70) member 2 | male meiosis I, spermatogenesis, spermatid development, response to heat and cold stress, positive regulation of G2/M transition of the mitotic cell cycle, cell differentiation, positive control over ATPase activity, protein folding mediated by chaperones that require cofactors, the reaction of cells to unfolded proteins, and protein refolding | 69.81 |
| A0A4 W2DHF8 | HSPA9 | heat shock protein family A (Hsp70) member 9 | ATP binding, metabolic processes, mitochondrial protein folding, enzyme binding, stress response | 73.7 |
| A0A4 W2I3C7 | SPAM1 | sperm adhesion molecule 1 (Hyaluronidase) | Sperm egg recognition, fertilization | 62.4 |
| A0A6P5CLI1 | ODF2 | the outer dense fiber of sperm tails 2 | spermatogenesis, cell differentiation, cilium organisation, regulation of cilium assembly, | 105 |
| A0A6P5BWX7 | PDHA2 | pyruvate dehydrogenase E1 subunit alpha 2 | Carbohydrate metabolism, Glucose metabolism, | 43.3 |
| Q3MHM5 | TUBB4B | tubulin beta 4B class IVb | microtubule-based process, Cytoskeleton | 49.8 |
| F6RP72 | TUBB1A | Tubulin alpha chain | microtubule-based process, Cytoskeleton | 50.9 |

Madura bull samples. These proteins are known as TUBA1A and TUBB4B. These proteins are necessary for cilia and flagella formation; hence, their importance cannot be overstated.

## 4.2 Chaperone proteins: heat shock proteins

Heat stress affects cattle productivity and reproduction. Heat stress damages developing spermatozoa. However, heat tolerance can be inherited by offspring (3). According to the relevant literature, Madura cattle are a native breed of beef cow thought to have developed on the dry and barren island of Madura (3). Madura Island receives 1600 mm of rain annually and is hot and arid (31). HSPs are linked to heat tolerance and reproductive performance (3, 32). In this study, HSPA9 and HSPA2 were found to be ubiquitously expressed. HSPA9, found in sperm mitochondria, significantly reduces the effects of heat stress in tropical cattle (33, 34). It is also known to regulate sperm motility through metabolic processes, ATP binding, and folding of mitochondrial proteins (35).

HSPA2 (also known as HSP70-2) was discovered in this study, which is consistent with previous findings in Australian Brahman bulls (*Bos indicus*) (36) and Zebu (37). HSP70-2 (heat shock protein 70–2) was found to be expressed in the sperm nucleus as well as in various other sperm organelles, including the mitochondria and flagellum (Figure 5). HSP70-2 is a component of the chromatin structure and contributes to gene regulation, particularly in the folding and unfolding of proteins. This HSPA2 sperm protein may also be crucial in protecting the sperm of tropical acclimated bulls such as Madura from heat or environmental stress. HSP70 is also an excellent predictor of thermotolerance and thermoresistance (38). The potential use of HSP70 as a biomarker for animal fertility and thermotolerance has been suggested (39). Therefore, HSP70-2 is recognised as a dual-functioning gene (3).

## 4.3 Pathway enrichment

The enriched pathways were also analysed for the identified proteins in the Madura bull spermatozoa. The present study revealed that 16% of the proteins were involved in the biosynthesis of precursor metabolites and energy. Only metabolic pathways were included in the list of the significant ways enriched by the proteins. Energy metabolism is essential for sperm development. Sperm require ATP, which is most likely to maintain morphological changes during the spermatid stage, for the degradation and synthesis of active proteins (40). ATP also serves as the molecular motor that provides the necessary energy for flagellar movement in all kinetics-related biochemical activities. The idea that mitochondria in the sperm midpiece drive mammalian sperm is evolving, as evidence suggests that glycolysis is the preferred

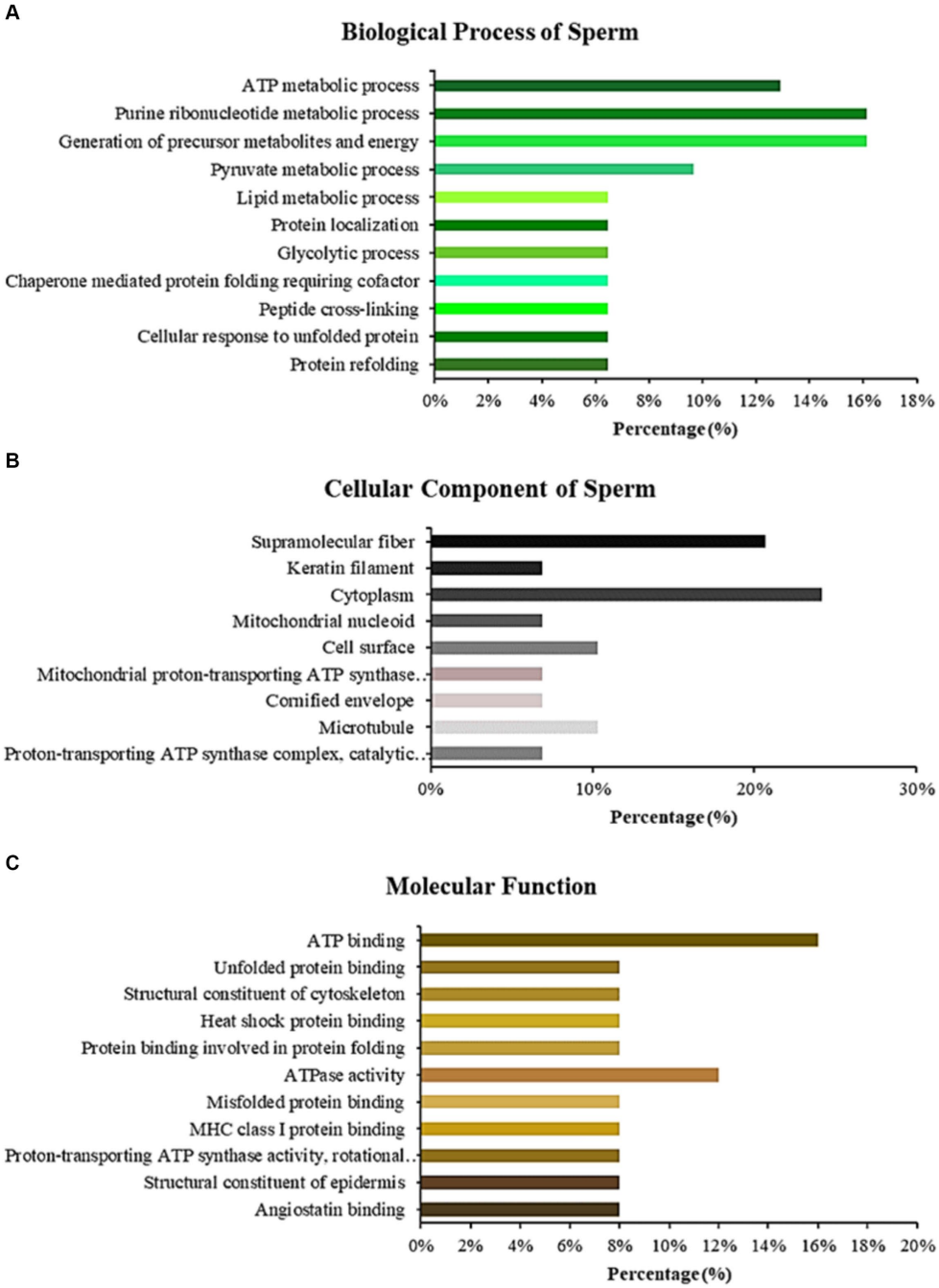



FIGURE 2
Bar chart representing the gene ontology annotations of proteins identified in Madura bull spermatozoa according to biological processes **(A)**, cellular components **(B)**, and molecular functions **(C)**.

metabolic pathway to maintain sperm motility in many species (41). Thus, the current study's results are in line with earlier research showing that ATP synthases are hub proteins found in the sperm of rams that are comparatively less fertile and may regulate how much ATP is used by sperm throughout the fertilization process (42). Existing data indicate that ATP production in sperm is facilitated by both glycolysis and mitochondrial respiration. These processes are interdependent and regulate sperm function based on the availability of energy substrates in the surrounding environment. An alternative glycolysis pathway for ATP production is also found in stallion spermatozoa (43). This study presents novel findings that suggest that Madura sperm are dependent on mitochondrial ATP synthesis, specifically regarding the roles of ATPF51A and ATP5B in sperm function and fertility.

Further, glucose utilisation is restricted during spermatogenesis, and lactate and pyruvate are favoured as substrates for energy synthesis (19). The generation of ATP in mammalian sperm is attributed to the energy metabolism pathway. This pathway exhibits subcellular compartmentalisation, with oxidative phosphorylation (OXPHOS) predominantly occurring in the sperm midpiece and

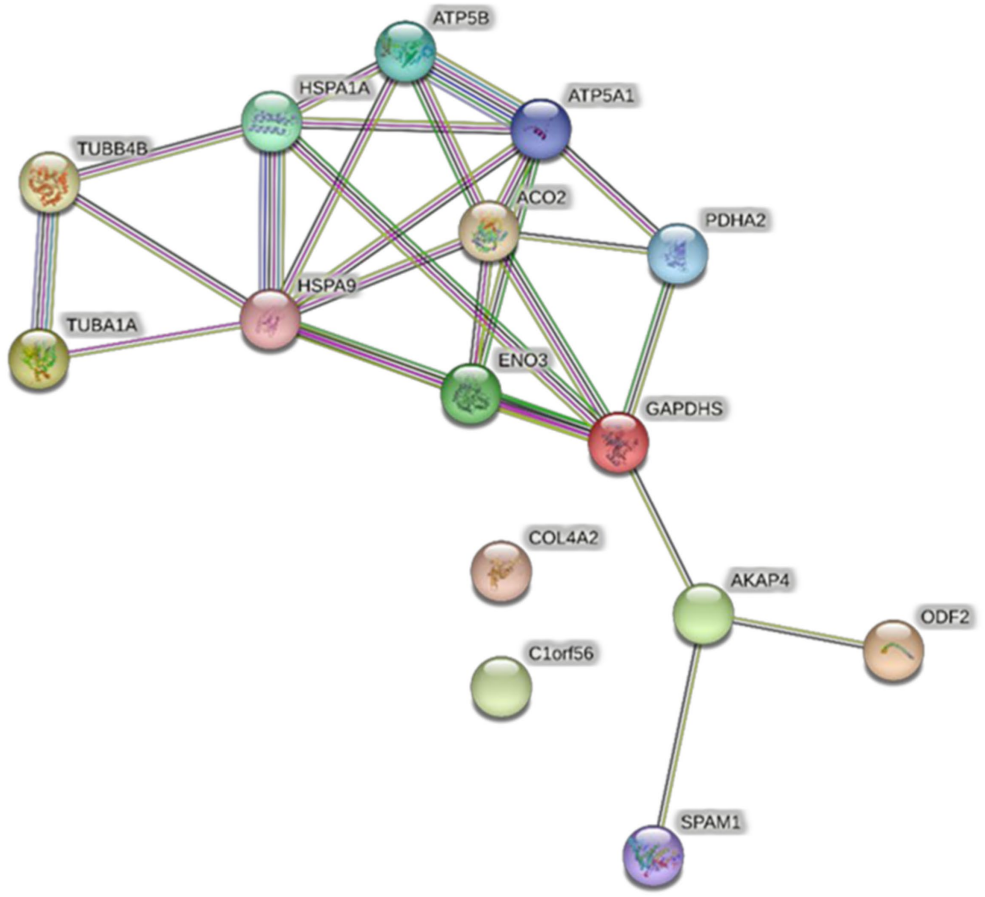



FIGURE 3
STRING protein–protein interaction network showing the interactions of the spermatozoa proteins identified in Madura bull sperm.

glycolysis in the principal piece (40). Glycolysis is a significant mechanism for facilitating the transport of ATP along the flagellum. Westhoff and Kamp (44), as well as Welch et al. (45), have reported the presence of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) in significant quantities within the fibrous sheath of sperm from different mammalian species, including humans. GAPDH is a NAD-dependent glycolytic enzyme responsible for facilitating the conversion of glyceraldehyde 3-phosphate (GAP) to 1,3-biphosphoglycerate (1,3 BPG). Notably, one of the isoforms of GAPDH, known as GAPDHS, has been identified in this study.

Moreover, Zini et al. (46) found that an augmentation in the activity of pyruvate dehydrogenase complexes can increase pyruvate-lactate usage, thus producing a more significant amount of NADH. NADH oxidase can utilise the NADH produced to generate reactive oxygen species (ROS) necessary for the capacitation process. NADH oxidase activity and ROS formation have also been associated with the motility of human spermatozoa. In this way, GAPDHS contributes significantly to energy metabolism. Elkina et al. (47) found GAPDHS in the testes of humans and rats; the enzyme is primarily present in the cytoplasm of all spermatogenic cells, whereas it is localised in the sperm tail in the epididymis. This finding provides support for this theory. This previous study by Kumar et al. (48) also established the involvement of extramitochondrial localised pyruvate dehydrogenase complex (PDHA) in the process of sperm capacitation, as well as the significance of pyruvate in the overall energy metabolism of mammalian sperm. The metabolic pathway encompassed the GAPDHS and PDHA2 proteins, both detected in the sperm of Madura bulls in the present study.

In agreement, energy metabolism plays a critical role in facilitating sperm function. The TCA cycle is the primary energy source for spermatids, although glycolytic and pentose phosphate pathways also play a role in energy synthesis in spermatozoa. The TCA cycle generates adenine, which is converted to ATP by the electron transport chain of the OXPHOS pathway. These proteins may be involved in creating the acrosome and activities there, both of which require energy provided by oxidative phosphorylation. ATP is then delivered to the microtubules responsible for supporting various sperm functions (49). In this study, two candidate proteins related to the tricarboxylic acid cycle pathway were PDHA2 and ACO2.

It was intriguing to discover that specific proteins are involved in so many energy metabolism processes, suggesting that a single protein may not perform glycolysis or OXPHOS as the primary actor in maintaining sperm functionality. The current study shows that proteomic methods may provide a reliable source for sperm protein detection and improve gene ontology comprehension. Therefore, we used high-throughput LC–MS/MS to develop a proteomic profile of Madura bull sperm connected to reproductive characteristics. Preserving superior native germplasm and

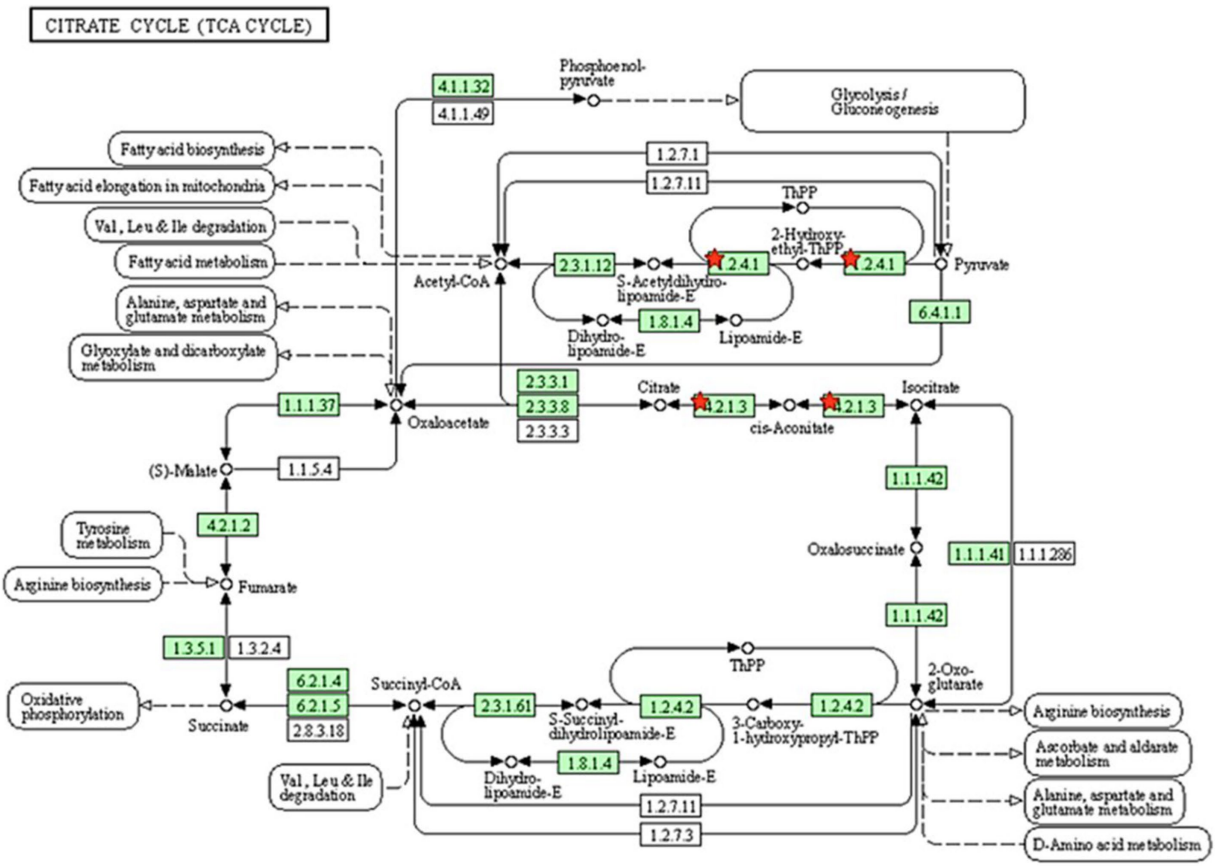



FIGURE 4
Pathway analysis using the KEGG pathway database identified proteins involved in the TCA cycle. The red star denotes the detected proteins.

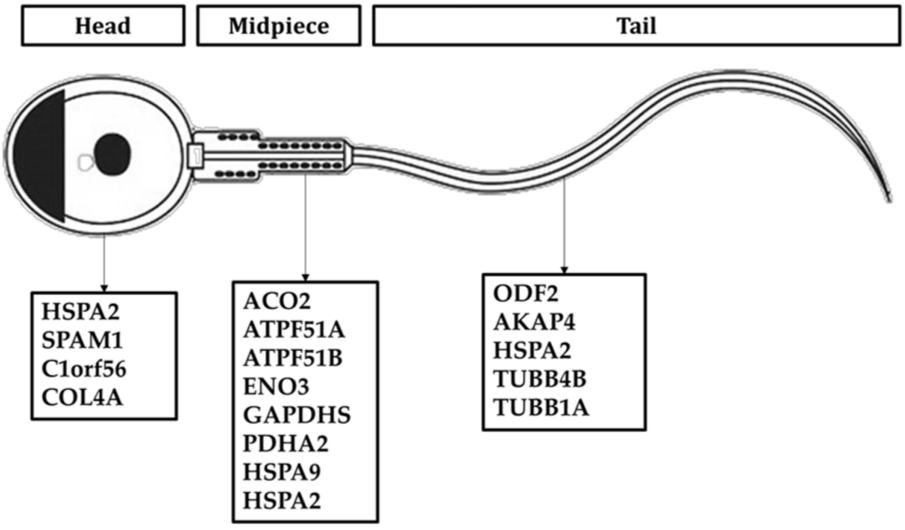



FIGURE 5
Distribution of sperm protein of Madura bull (*Bos indicus* × *Bos javanicus*).

preventing gene pool depletion requires studying protein differences in different breeds. Thus, investigating Madura bull protein composition can help us comprehend native Indonesian bull sperm biology. Finally, knowing the proteome of ejaculated spermatozoa can help us understand the reproductive process in Madura bull and bovine in general.

## 5 Conclusion

We examined the sperm proteomes of Madura cows (*Bos indicus* × *Bos javanicus*). The use of several methods to prepare samples has made it easier to identify proteins in sperm. Most of the proteins identified in this study are essential for bull breeding. We believe that information about the sperm proteome of the superior Madura bull, a breed with good reproductive traits, will speed up future studies on bull fertility. Additionally, this would enable the creation of molecular tools for accurately selecting bulls to conserve indigenous Indonesian cattle breeds and bovines.

## 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 approval was not required for the studies on animals in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used.

## Author contributions

ZR: Conceptualization, Data curation, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Methodology. BP: Data curation, Funding acquisition, Project administration, Validation, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Methodology, Visualization. EK: Writing – review & editing. MG: Writing – review & editing. TM: Writing – review & editing. LT: Writing – review & editing, Conceptualization, Supervision. DS: Conceptualization, Supervision, Writing – review & editing. MU: Conceptualization, Supervision, Writing – review & editing. BP: Conceptualization, Data curation, Funding acquisition, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. SS: Methodology, Supervision, Formal analysis, Validation, Investigation, Funding acquisition, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was financially supported by the Indonesian Ministry of Education, Culture, Research, and Technology through Fundamental Research grant number 18789/IT3.D10/PT.01.02/M/T/2023.

## Acknowledgments

The authors sincerely thank the Singosari and Lembang AIC for providing bull data. We also thank Wiwit for his invaluable help in the laboratory.

## 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



## References

1. Peddinti D, Nanduri B, Kaya A, Feugang JM, Burgess SC, Memili E. Comprehensive proteomic analysis of bovine spermatozoa of varying fertility rates and identification of biomarkers associated with fertility. *BMC Syst Biol.* (2008) 2:1–13. doi: 10.1186/1752-0509-2-19

2. Thompson SK, Kutchy NA, Kwok S, Rosyada ZNA, Imumorin IG, Purwantara B, et al. Review: sperm: comparative morphology and function related to altered reproductive strategies and mammalian fertil. Professional animal. *Scientist.* (2018) 34:558–65. doi: 10.15232/pas.2018-01748

3. Rosyada ZNA, Ulum MF, Tumbelaka LITA, Solihin DD, Purwantara B, Memili E. Implications of sperm heat shock protein 70-2 in bull fertility. *Vet World.* (2022) 15:1456–66. doi: 10.14202/vetworld.2022.1456-1466

4. Nowicka-Bauer K, Malcher A, Włoczkowska O, Kamieniczna M, Olszewska M, Kurpisz MK. Evaluation of seminal plasma HSPA2 protein as a biomarker of human spermatogenesis status. *Reprod Biol.* (2022) 22:100597. doi: 10.1016/j.repbio.2021.100597

5. Rahman MS, Kwon WS, Pang MG. Prediction of male fertility using capacitation-associated proteins in spermatozoa. *Mol Reprod Dev.* (2017) 84:749–59. doi: 10.1002/mrd.22810

6. Somashekar L, Selvaraju S, Parthipan S, Ravindra JP. Profiling of sperm proteins and association of sperm PDC-109 with bull fertility. *Syst Biol Reprod Med.* (2015) 61:376–87. doi: 10.3109/19396368.2015.1094837

7. Manabi Diansyah A, Yusuf M, Latief Toleng A, Ihsan Andi Dagong M, Maulana T, Hasrin H, et al. The sperms post-thawing quality and proteomic seminal plasma on fertility performance of Bali-polled bull. *Adv Anim Vet Sci.* (2022) 11:517–525. doi: 10.17582/journal.aavs/2023/11.4.517.525

8. Karanwal S, Pal A, Chera JS, Batra V, Kumaresan A, Datta TK, et al. Identification of protein candidates in spermatozoa of water buffalo (*Bubalus bubalis*) bulls helps in predicting their fertility status. *Front cell. Dev Biol.* (2023) 11:11. doi: 10.3389/fcell.2023.1119220

9. Monaco E, Gasparrini B, Boccia L, De Rosa A, Attanasio L, Zicarelli L, et al. Effect of osteopontin (OPN) on in vitro embryo development in cattle. *Theriogenology.* (2009) 71:450–7. doi: 10.1016/j.theriogenology.2008.08.012

10. Luongo C, González-Brusi L, Cots-Rodríguez P, Izquierdo-Rico MJ, Avilés M, García-Vázquez FA. Sperm proteome after interaction with reproductive fluids in porcine: from the ejaculation to the fertilization site. *Int J Mol Sci.* (2020) 21:1–27. doi: 10.3390/ijms21176060

11. Guasti PN, Souza FF, Scott C, Papa PM, Camargo LS, Schmith RA, et al. Equine seminal plasma and sperm membrane: functional proteomic assessment. *Theriogenology.* (2020) 156:70–81. doi: 10.1016/j.theriogenology.2020.06.014

12. Pini T, Leahy T, Soleilhavoup C, Tsikis G, Labas V, Combes-Soia L, et al. Proteomic investigation of ram spermatozoa and the proteins conferred by seminal plasma. *J Proteome Res.* (2016) 15:3700–11. doi: 10.1021/acs.jproteome.6b00530

13. Rosyada ZNA, Tumbelaka LI, Ulum MF, Solihin DD, Kaiin EM, Gunawan M, et al. Meta data analysis of conception rate in relation to sperm motility in Madura superior bulls. *IOP Conf Ser Earth Environ Sci.* (2021) 902:012048. doi: 10.1088/1755-1315/902/1/012048

14. Rosyada ZNA, Yoelinda VT, Kaiin EM, Gunawan M, Ulum MF, Tumbelaka LITA, et al. Sperm osteopontin mRNA expression levels and its correlation on semen quality and fertility in Madura bulls. *Biodiversitas*. (2023) 24:563–70. doi: 10.13057/biodiv/d240165

15. Ramesha KP, Mol P, Kannegundla U, Thota LN, Gopalakrishnan L, Rana E, et al. Deep proteome profiling of semen of Indian indigenous Malnad Gidda (*Bos indicus*) cattle. *J Proteome Res*. (2020) 19:3364–76. doi: 10.1021/acs.jproteome.0c00237

16. Druart X, De GS. Seminal plasma proteomes and sperm fertility seminal plasma proteomes and sperm fertility. *Anim Reprod Sci*. (2018) 194:33–40. doi: 10.1016/j.anireprosci.2018.04.061

17. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*. (1976) 72:248–54. doi: 10.1016/0003-2697(76)90527-3

18. Prasad TSK, Harsha HC, Keerthikumar S, Sekhar NR, Selvan LDN, Kumar P, et al. Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. *J Proteome Res*. (2012) 11:247–60. doi: 10.1021/pr200827k

19. Fu Q, Pan L, Huang D, Wang Z, Hou Z, Zhang M. Proteomic profiles of buffalo spermatozoa and seminal plasma. *Theriogenology*. (2019) 134:74–82. doi: 10.1016/j.theriogenology.2019.05.013

20. Martin-Deleon PA. Germ-cell hyaluronidases: their roles in sperm function. *Int J Androl*. (2011) 34:e306–18. doi: 10.1111/j.1365-2605.2010.01138.x

21. Gómez-Torres MJ, Sáez-Espinosa P, Manzano-Santiago P, Robles-Gómez L, Huerta-Retamal N, Aizpurua J. Sperm adhesion molecule 1 (SPAM1) distribution in selected human sperm by hyaluronic acid test. *Biomedicine*. (2022) 10:10102553. doi: 10.3390/biomedicines10102553

22. Zhao W, Li Z, Ping P, Wang G, Yuan X, Sun F. Outer dense fibers stabilize the axoneme to maintain sperm motility. *J Cell Mol Med*. (2018) 22:1755–68. doi: 10.1111/jcmm.13457

23. Ito C, Akutsu H, Yao R, Yoshida K, Yamatoya K, Mutoh T, et al. Odf2 haploinsufficiency causes a new type of decapitated and decaudated spermatozoa, Odf2-DDS, in mice. *Sci Rep*. (2019) 9:14249. doi: 10.1038/s41598-019-50516-2

24. Sumarsono T, Supriatna I, Setiadi MA, Agil M, Purwantara B. Detection of plasma membrane alpha-enolase (ENO1) and its relationship with sperm quality of Bali cattle. *Trop Anim Sci J*. (2023) 46:36–42. doi: 10.5398/tasj.2023.46.1.36

25. He Y, Li H, He J, Zhao X. Heavy ion radiation can promote greater motility and enolase protein expression in ram sperm in in vitro liquid storage. *Anim Reprod Sci*. (2014) 148:260–6. doi: 10.1016/j.anireprosci.2014.06.019

26. Xi MD, Li P, Du H, Qiao XM, Liu ZG, Wei QW. Disaccharide combinations and the expression of enolase3 and plasma membrane Ca2+ ATPase isoform in sturgeon sperm cryopreservation. *Reprod Domest Anim*. (2018) 53:472–83. doi: 10.1111/rda.13134

27. Blommaert D, Sergeant N, Delehedde M, Jouy N, Mitchell V, Franck T, et al. Expression, localization, and concentration of A-kinase anchor protein 4 (AKAP4) and its precursor (proAKAP4) in equine semen: promising marker correlated to the total and progressive motility in thawed spermatozoa. *Theriogenology*. (2019) 131:52–60. doi: 10.1016/j.theriogenology.2019.03.011

28. Carrera A, Gerton G, Moss SB. The major fibrous sheath polypeptide of mouse sperm: structural and functional similarities to the A-kinase anchoring proteins. *Dev Biol*. (1994) 165:272–84. doi: 10.1006/dbio.1994.1252

29. Moss SB, Turner RMO, Burkert KL, Butt HV, Gerton GL. Conservation and function of a bovine sperm A-kinase anchor protein homologous to mouse AKAP82 1. *Biol Reprod*. (1999) 61:335–42. doi: 10.1095/biolreprod61.2.335

30. Kierszenbaum AL. Genomic imprinting and epigenetic reprogramming: unearthing the garden of forking paths. *Mol Reprod Dev*. (2002) 63:269–72. doi: 10.1002/mrd.90011

31. Kutsiyah F. Population dynamic and productivity of Madura cattle in Corservation area of Sapudi Island. *Sains Peternakan*. (2017) 15:70–7. doi: 10.20961/sainspet.v15i2.13160

32. Purandhar K, Jena PK, Prajapati B, Rajput P, Seshadri S. Understanding the role of heat shock protein isoforms in male fertility, aging and apoptosis. *World J Mens Health*. (2014) 32:123. doi: 10.5534/wjmh.2014.32.3.123

33. Aarif O, Mir N, Shergojry S, Ramesha KP, Mir A. Genetic polymorphisms within exon 8, 9, and 10 of heat shock protein (HSP) 90AA1 in deoni cattle. *Anim Sci Report*. (2014) 8:26–30.

34. Rosyada Z, Pardede B, Kaiin E, Tumbelaka L, Solihin D, Purwantara B, et al. Identification of HSP70-2 and PRM-1 mRNA, proteins, and analyses of their association with fertility using frozen-thawed sperm in Madura bulls. Animal. *Bioscience*. (2023) 36:1796–805. doi: 10.5713/ab.23.0142

35. Liang J, Zheng Y, Zeng W, Chen L, Yang S, Du P, et al. Proteomic profile of sperm in infertile males reveals changes in metabolic pathways. *Protein J*. (2021) 40:929–39. Available from:. doi: 10.1007/s10930-021-10013-w

36. Rego JPA, Crisp JM, Moura AA, Nouwens AS, Li Y, Venus B, et al. Seminal plasma proteome of electroejaculated *Bos indicus* bulls. *Anim Reprod Sci*. (2014) 148:1–17. doi: 10.1016/j.anireprosci.2014.04.016

37. Zhang XG, Hong JY, Yan GJ, Wang YF, Li QW, Hu JH. Association of heat shock protein 70 with motility of frozen-thawed sperm in bulls. *Czech JAnim*. (2015) 60:256–62. doi: 10.17221/8239-CJAS

38. Ul HF, Nawaz A, Rehman MS, Ali MA, SMR D, Yang C. Prospects of HSP70 as a genetic marker for thermo-tolerance and immuno-modulation in animals under climate change scenario. Animal. *Nutrition*. (2019) 5:340–50. doi: 10.1016/j.aninu.2019.06.005

39. Wang J, Que SQ, Liu X, Jin M, Xin TR, Zou ZW, et al. Characteristic and expression of Hsp70 and Hsp90 genes from Tyrophagus putrescentiae and their response to thermal stress. *Sci Rep [Internet]*. (2021) 11:1–12. Available from:. doi: 10.1038/s41598-021-91206-2

40. Amaral A. Energy metabolism in mammalian sperm motility. *WIREs Mechanisms of Disease*. (2022) 14:1569. doi: 10.1002/wsbm.1569

41. Castellini C, D'andrea S, Cordeschi G, Totaro M, Parisi A, Di Emidio G, et al. Pathophysiology of mitochondrial dysfunction in human spermatozoa: Focus on energetic metabolism, oxidative stress and apoptosis. *Antioxidants MDPI*. (2021) 10. doi: 10.3390/antiox10050695

42. Hitit M, Özbek M, Ayaz-Guner S, Guner H, Oztug M, Bodu M, et al. Proteomic fertility markers in ram sperm. *Anim Reprod Sci*. (2021) 235:106882. doi: 10.1016/j.anireprosci.2021.106882

43. Davila MP, Muñoz PM, Bolaños JMG, Stout TAE, Gadella BM, Tapia JA, et al. Mitochondrial ATP is required for the maintenance of membrane integrity in stallion spermatozoa, whereas motility requires both glycolysis and oxidative phosphorylation. *Reproduction*. (2016) 152:683–94. doi: 10.1530/REP-16-0409

44. Westhoff D, Kamp G. Glyceraldehyde 3-phosphate dehydrogenase is bound to the fibrous sheath of mammalian spermatozoa. *J Cell Sci*. (1997) 110:1821–9. doi: 10.1242/jcs.110.15.1821

45. Welch JE, Brown PL, O'Brien DA, Magyar PL, Bunch DO, Mori C, et al. Human glyceraldehyde 3-phosphate dehydrogenase-2 gene is expressed specifically in spermatogenic cells. *J Androl*. (2000) 21:328–38. doi: 10.1002/j.1939-4640.2000.tb02111.x

46. Zini A, O'Bryan MK, Israel L, Schlegel PN. Human sperm NADH and NADPH diaphorase cytochemistry: correlation with sperm motility. *Urology*. (1998) 51:464–8. doi: 10.1016/S0090-4295(97)00631-6

47. Elkina YL, Atroshchenko MM, Bragina EE, Muronetz VI, Schmalhausen EV. Oxidation of glyceraldehyde-3-phosphate dehydrogenase decreases sperm motility. *Biochemistry (Mosc)*. (2011) 76:268–72. doi: 10.1134/S0006297911020143

48. Kumar V, Rangaraj N, Shivaji S. Activity of pyruvate dehydrogenase a (PDHA) in Hamster spermatozoa correlates positively with Hyperactivation and is associated with sperm capacitation. *Biol Reprod*. (2006) 75:767–77. doi: 10.1095/biolreprod.106.053587

49. Magdanz V, Boryshpolets S, Ridzewski C, Eckel B, Reinhardt K. The motility-based swim-up technique separates bull sperm based on differences in metabolic rates and tail length. *PLoS One*. (2019) 14:e0223576. doi: 10.1371/journal.pone.0223576

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 12 December 2023
DOI 10.3389/fvets.2023.1269904



OPEN ACCESS

EDITED BY
Ana Fabrícia Braga Magalhães,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Krishnamoorthy Srikanth,
Cornell University, United States
Marta Martins,
Brazilian Agricultural Research Corporation (EMBRAPA), Brazil
Maria Malane M. Muniz,
University of Guelph, Canada

*CORRESPONDENCE
Xiaoyan Kong
kongyan988@163.com
Dawei Yan
1995027@ynau.edu.cn

RECEIVED 31 July 2023
ACCEPTED 27 November 2023
PUBLISHED 12 December 2023

CITATION
Li X, Xin A, Ma L, Gou X, Fang S, Dong X, Ni B, Tang L, Zhu L, Yan D and Kong X (2023) Molecular genetic characterization and meat-use functional gene identification in Jianshui yellow–brown ducks through combined resequencing and transcriptome analysis.
*Front. Vet. Sci.* 10:1269904.
doi: 10.3389/fvets.2023.1269904



# Molecular genetic characterization and meat-use functional gene identification in Jianshui yellow–brown ducks through combined resequencing and transcriptome analysis

Xinpeng Li[1], Aiguo Xin[2], Li Ma[3], Xiao Gou[4], Suyun Fang[1], Xinxing Dong[1], Bin Ni[4], Lin Tang[1], Li Zhu[1], Dawei Yan[1]* and Xiaoyan Kong[1]*

[1]Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming, China, [2]Poultry Husbandry and Disease Research Institute, Yunnan Academy of Animal Husbandry and Veterinary Sciences, Kunming, China, [3]Animal Husbandry and Veterinary College, Yunnan Vocational and Technical College of Agriculture, Kunming, China, [4]School of Life Science and Engineering, Foshan University, Foshan, China

The Jianshui yellow–brown duck is a unique country-specific waterfowl species in Yunnan Province, well known for its tender meat. However, there is a lack of comprehensive systematic research on the molecular genetic characteristics, especially germplasm resources and economic traits, of the Jianshui yellow–brown ducks. This study investigated the molecular genetic characteristics of Jianshui yellow–brown ducks, compared their selection signals with those of ancestral mallard and meat-type Pekin ducks, and identified genes specific to their meat-use performance. Furthermore, this study also evaluated the breeding potential for its meat performance. In this study, phylogenetic trees, PCA and Admixture analysis were used to investigate the population genetic structure among local duck breeds in China; population genetic differentiation index (Fst), nucleotide diversity and Tajima's D were used to detect selected loci and genes in the population of Jianshui yellow–brown ducks; and transcriptome technology was used to screen for differentially expressed genes in the liver, sebum and breast muscle tissues, and finally, the results of the genome selection signals and transcriptome data were integrated to excavate functional genes affecting the meat performance of the Jianshui yellow–brown ducks. The results of the genetic structure of the population showed that Jianshui yellow–brown ducks were clustered into a separate group. Selection signal analysis indicated significant selection pressure on certain genes related to meat characteristics (*ELOVL2*, *ELOVL3*, *GDF10*, *VSTM2A*, *PHOSPHO1*, and *IGF2BP1*) in both Jianshui yellow–brown ducks and mallards. Transcriptomic data analysis suggested that *ELOVL3*, *PHOSPHO1*, and *GDF10* are vital candidate genes influencing meat production and quality in Jianshui yellow–brown ducks. A comparison of selection signals between Jianshui yellow–brown ducks and Pekin ducks revealed only 21 selected genes in the Jianshui yellow–brown duck population, and no significant genes were related to meat traits. Moreover, whole-genome resequencing data suggested that the Jianshui yellow–brown duck represents a unique category with distinct genetic mechanisms. Through selection signaling and transcriptomic approaches, we successfully screened and identified important candidate genes affecting meat traits in Jianshui yellow–brown ducks. Furthermore, the Jianshui

yellow–brown duck has good potential for improved meat performance, highlighting the need for further improvement.

KEYWORDS

Jianshui yellow–brown duck, meat-use functional gene, molecular genetics, resequencing, transcriptome

# 1 Introduction

China, has an rich extensive history of duck domestication and holds a rich abundance of waterfowl resources. China is home to nearly half of the world's existing duck varieties (1). However, despite this vast array of varieties, the full potential of the abundant genetic resources remains untapped because of the delayed initiation of modernized breeding practices for domestic ducks in the country. One notable example is the development of Cherry Valley Peking ducks by the British Cherry Valley Company through systematic and scientific breeding of Chinese Peking ducks. In contrast to native Peking ducks, Cherry Valley Peking ducks stand out due to their large size and rapid growth rate. Upon entering the Chinese market, Cherry Valley Peking ducks quickly gained favor among duck farmers and swiftly dominated the Chinese duck meat market. This, in turn, resulted in a significant reduction in the population of native duck breeds and placed substantial pressure on their living spaces (2, 3). Therefore, the rational development and utilization of resources, guided by breed characteristics, are of paramount importance in breed selection and the sustainable development of related industries.

Sequencing technology has been widely used in studying species origin, evaluating germplasm resources, and understanding population genetic diversity. Feng et al. calculated and compared genetic diversity in five duck varieties using genome-wide single-nucleotide polymorphism (SNP) loci. They discovered that domestic duck varieties exhibited lower genetic diversity than wild duck varieties (4). In evolutionary tree analysis, individuals from domestic duck varieties were clustered into one category, while those from wild duck varieties were clustered into another category, suggesting a common wild ancestor for domestic ducks (5). Hence, assessment of the genetic diversity of duck varieties and accurate identification and evaluation of germplasm resources form an important basis for identifying excellent candidate genes and promoting the development, utilization, and innovation of germplasm resources. With the widespread application of high-throughput sequencing technology, the integration of multi-omics data enables a deeper understanding of the genetic basis and molecular mechanisms underlying complex traits (6–8). The lack of systematic breeding conservation measures and scientific selective breeding programs for Jianshui yellow–brown ducks has resulted in issues such as varietal complexity and degradation, thereby affecting its genetic resources adversely.

Jianshui yellow–brown ducks are native to Jianshui County, Yunnan Province, and have excellent traits such as tender meat, docile temperament and strong adaptability after selective breeding. However, as a local breed for meat and eggs, Jianshui yellow–brown ducks, which are less selectively bred, have the potential to be bred into specialized meat breeds (e.g., Pekin ducks for meat breeds) in the future. However, the molecular mechanisms influencing the meat production and quality of Jianshui yellow–brown ducks have not been investigated in depth. This study investigated the population genetic structure and genetic diversity of the Jianshui yellow–brown duck population using resequencing technology. Subsequently, based on the results of the population genetic structure analysis, this study performed a comparison between this duck population and the mallard population to identify candidate genes associated with meat production and fat deposition in Jianshui yellow–brown ducks through combined analysis of selection signals and transcriptomes. Furthermore, a selection signal analysis was performed on Jianshui yellow–brown ducks and meat-type Pekin ducks to explore genomic-level genetic differences between the two varieties resulting from different domestication and breeding processes. Finally, the meat-trait breeding status in Jianshui yellow–brown ducks was evaluated.

# 2 Materials and methods

## 2.1 Sample collection

We randomly selected 18 healthy Jianshui yellow–brown ducks (YB), including 7 males and 11 females, at the Jianshui yellow and brown duck breeding farm in Jianshui County, Yunnan Province. During the collection process, 5 mL of blood samples were taken from the wing veins of these ducks using vacuum blood collection tubes with EDTA anticoagulant. Ice packs were cryopreserved and transported back to the laboratory and stored in a −20°C refrigerator for subsequent extraction of whole gene DNA. In addition, 6 Jianshui yellow–brown ducks (3 males and 3 females) were randomly selected for slaughter, and 6 liver, 6 sebum, and 3 breast muscle tissue samples were collected, immediately place the tissue into a pre-prepared cryopreservation tube containing RNA preservation solution (the RNA preservation solution is RNA solid Stable Preservation Solution, Wuhan Servicebio Company), place it in liquid nitrogen, take it back to the laboratory and store it in a −80°C ultra-low temperature refrigerator for subsequent extraction of total tissue RNA. For detailed information on the collected samples of Jianshui yellow–brown ducks, please refer to Supplementary Table S1. All experimental procedures adhered to the regulations of the Administration of Laboratory Animal Affairs and were approved by the Ethics Committee of Yunnan Agricultural University (No. 202103035).

## 2.2 Whole-genome resequencing and bioinformatics analysis

### 2.2.1 Population variation detection

The quality and concentration of the extracted DNA were assessed using 1% agarose gel electrophoresis and ultraviolet spectrophotometry. After passing the quality control, the samples will be sent to

BerryGenomics Corporation (Beijing, China) for whole genome resequencing. The sequencing will be performed on Illumina novaseq 6,000 sequencing platform, using paired-end sequencing, and the sequencing depth is 15×. Additionally, resequencing data for a few duck varieties were downloaded from the NCBI database (Bioproject ID: PRJNA450892, PRJNA599025, PRJNA645648, and PRJNA419832), including wild varieties (mallard, ML, $n = 10$; spot-billed duck, SB, $n = 10$), meat-type varieties (Pekin duck, PK, $n = 16$; Cherry Valley duck, CV, $n = 10$), and meat-egg-type varieties (Fenghua duck, FH, $n = 10$; Shanma duck, SM, $n = 10$; Shaoxing duck, SX, $n = 10$; Gaoyou duck, GY, $n = 8$; Jinding duck, JD, $n = 8$). Information on some of the Chinese local duck breeds included in this study is shown in Supplementary Table S2. The resequencing data of 18 Jianshui yellow–brown ducks (from a total of 10 populations and 110 individuals) obtained in this study were subjected to population structure and genetic diversity analyses. The raw data underwent quality control and filtering using FASTP (9) software with default parameters. After quality control, the clean data were compared with the Pekin duck reference genome (10) by aligning both sequences using bwa (11) software (version CAU_duck1.0: https://ftp.ensembl.org/pub/release-110/fasta/anas_platyrhynchos_platyrhynchos/dna/Anas_platyrhynchos_platyrhynchos.CAU_duck1.0.dna.toplevel.fa.gz). The sam files were converted to bam files, and sorted using the "-sort" command in SAMTOOLS (12) software. The commands "flagstat" and "coverage" were used to count the alignment rate, coverage rate, and coverage depth in each sample. PICARD software was used to label repetitive sequences resulting from amplification through polymerase chain reaction during sequencing and remove them. Alignment in separate samples and separate chromosomes was conducted using the "HaplotypeCaller," "CombineGVCFs," "GenotypeGVCFs," "MergeVcfs," and "SelectVariants" commands in the GATK (13) software to obtain the individual variation detection results SNP loci were extracted from the individual results, and population-level SNP variation results were obtained. The population variation detection results were initially filtered using the "VariantFiltration" command with the following specific criteria: QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0. All SNPs were further filtered to improve the confidence of the results. Those SNPs that did not meet the requirements of deletion rate and minimum allele frequency in the population were further filtered using VCFTOOLS (14) software according to the criteria of "--geno 0.1 ---maf 0.01." Finally, the high-confidence population SNP variation detection results were saved in the vcf format and used for subsequent downstream analysis.

### 2.2.2 Population structure, genetic differentiation, and genetic diversity analyses

Based on the VCF files obtained in the previous step, the population was subjected to PCA analysis using PLINK software, followed by visualization of the PCA results using R. The vcf file of the population SNP variation results was converted to the phylip format using the run_pipeline.pl. program in TASSEL (15) software. A neighbor-joining (NJ) evolutionary tree was constructed using PHYLIP (16) software, and the phylogenetic tree was beautified using the iTOL online website (17). The genetic differentiation index Fst between two populations was calculated using VCFTOOLS software with a sliding window approach, utilizing a window size of 100 kb and a sliding step size of 10 kb. Cross-validation error rates were calculated using ADMIXTURE (18) software for different $K$ values, and the optimal $K$ value was determined. Stacked plots were created using the pophelper software package in R. Nucleotide diversity values for each of the 10 populations were calculated using the vcftools software "--window-pi 10 M" command, and boxplots were generated using ggplot2 for visualization and comparison [In the study of population genetic differentiation and comparison of population nucleotide diversity, we used the parameters of window size and sliding step size with reference to Wang et al. (19)].

### 2.2.3 Population selection analysis

Nucleotide diversity ratio θπ and genetic differentiation index Fst were calculated using Vcftools software with a sliding window size of "40 kb" and a sliding step size of "10 kb." [In the study of selection signal analysis of populations based on population genetic differentiation and population nucleotide diversity, we used the parameters of window size and sliding step size with reference to Zhou et al. (20)]. The top 1% regions of the joint region of θπ and Fst were identified as candidate regions for selection. Duplicate regions were merged using bedtools software, and the genes within the filtered regions were annotated using the species gene annotation file.

## 2.3 Transcriptome sequencing and bioinformatics analysis

The Tiangen Total RNA Extraction Kit was used to extract total RNA from the liver, sebum and breast muscle tissue according to the manufacturer's instructions. The quality-controlled samples were sent to Berry Genomics (Beijing, China) for sequencing, which will be performed on the Illumina novaseq 6,000 sequencing platform using paired-end sequencing (Paired-end sequencing, measuring 150 bp at each end, and the length of each read is 300 bp) with 10 G per sample (i.e., 1 billion base pairs per sample by transcriptome sequencing to ensure sufficient data for accurate gene expression analysis). The liver, sebum and breast muscle transcriptomic data of mallard and Pekin ducks used in this study were downloaded from public databases (NCBI and BIG Data Center, Bioproject ID: PRJNA645648, BIG accession codes PRJCA001307), transcriptome sequencing data downloaded from public databases also use paired-end sequencing, with 150 bp at each end and a read length of 300 bp. The same tissues from Jianshui yellow–brown ducks, mallards, and Pekin ducks were compared to identify differentially expressed genes (DEGs). The raw data underwent filtering and quality control using FASTP software with default parameters. The resulting clean data were aligned to the Pekin duck reference genes using HISAT2 (21) software (version: CAU_duck1.0). The parameters for mapping sequenced reads back to the reference genome using HISAT2 are as follows: --rna-strandness --new-summary -x genome −1 read1.fq.gz −2 reads2.fq.gz -S sample.sam. We set the parameter "--rna-strandness" mainly because we adopted the strand-specific library construction method when we sequenced the transcriptome of Jianshui yellow–brown duck tissues. In this way, we were able to obtain the orientation information of the RNA fragments to estimate the gene expression level more accurately. At the same time, we used the "-x" parameter to specify the reference gene of the selected species. The aim of this study was to look for differences between species by performing transcriptome analyses of the same tissue from different species. In order to control a single variable as much as possible and reduce the influence of sex factors on differentially expressed genes,

we first removed the chromosomal information related to the sex of the species on the reference genome, and only aligned the reads back to the autosomes on the reference genome of the species. Expression quantification was performed on the alignment results using the "featurecounts" command of SUBREAD (22) software to obtain the raw expression matrix. Using the TPM (Transcripts Per Million) approach to normalize the raw expression matrix of transcriptomic data. TPM correction of raw counts prior to correlation analysis of samples is required in order to eliminate variability due to differences in gene length and sequencing depth. This ensures that comparisons of gene expression between samples are accurate and fair. Such standardization is essential for subsequent sample correlation analysis. Differential expression analysis was conducted on the raw expression matrix using DESEQ2 (23). We will use the raw expression matrix generated by featurecounts as the input file for DESEQ2. In this process, we set DESEQ2's comparison grouping information, i.e., grouping according to varieties. After that DESEQ2 will calculate key information such as $\log_2$|Foldchange|, value of $p$ and padj for each gene. Significant DEGs were identified based on the criteria of |log2 Fold Change| > 1.5 and corrected $p$-values (P.adj) less than 0.001. The correction for false-positive detection due to multiple comparison tests was performed using the Benjamini–Hochberg (BH) method. Screening for differentially expressed genes (DEGs) by adjusted $p$-values reduces the rate of false positives and enhances the reliability of the study results. The number of DEGs in each population was then counted, and volcano plots were generated using R.

Joint analysis of selection signal and transcriptome

To further narrow down the candidate gene list and increase the robustness of the results, a joint analysis of selection signals and transcriptome data was performed. First, we screened the candidate genes related to the target trait (i.e., meat performance) from the selection signal results, and intersected the candidate genes from the selection signal with the differentially expressed genes obtained from the transcriptome analysis, and the intersected portion of the results was the result of the joint analysis.

# 3 Results

## 3.1 Genome sequencing results and variation detection

Eighteen Jianshui yellow–brown ducks underwent resequencing, and the resequencing data of 92 ducks from nine different varieties were downloaded from the NCBI database, totaling 110 individuals from 10 populations. The quality of the sequencing data was assessed, and the 18 Jianshui yellow–brown ducks exhibited a base population quality with Q20 and Q30 values above 90%, normal CG distribution, an average sample alignment rate of 98.22%, an average sequencing depth of 18.94×, and an average sample coverage rate of 92.37% (Supplementary Table S3). The Q20 and Q30 values for the 92 resequencing data from the NCBI database also exceeded 90%, with an average alignment rate of 97.69%, an average sequencing depth of 8.14×, and an average coverage rate of 90.66% (Supplementary Table S4). Thus, the sequencing data demonstrated reliable quality and met the requirements for resequencing analysis. After the detection and filtering of population SNPs, a total of 19,795,912 high-quality SNPs were obtained from 110 individuals. The subsequent analysis focused on population structure, population genetic diversity, and population selection based on these population SNPs.

## 3.2 Population genetic structure and genetic differentiation

To examine the presence of outlier samples and inter-population genetic structure and relationships, PCA, phylogenetic tree construction, and population structure analysis were performed using whole-genome SNPs from 110 individuals representing 10 populations (Pekin, Cherry Valley Pekin, mallard, spot-billed, Fenghua, Shanma, Shaoxing, Jinding, Gaoyou, and Jianshui yellow–brown ducks). PCA results demonstrated that individuals within the same population tended to cluster together (Figure 1A). The 10 populations could be categorized into four groups: Gaoyou duck, Jinding duck, Shanma duck, and Shaoxing duck formed one group; mallard, spot-billed duck, and Fenghua duck formed another group; Pekin duck and Cherry Valley Pekin duck clustered together; and Jianshui yellow–brown duck formed a distinct individual population. The results from the NJ evolutionary tree (Figure 1B) were consistent with the PCA results, with the Jianshui yellow–brown duck population forming a separate cluster. Structure analysis indicated that the cross-validation error rate was minimized when $K = 4$ (Supplementary Figure S1). Figure 1C depicts the structure results when $K = 2$–4. When $K = 2$, the 10 populations were divided into two groups: one comprising domesticated varieties and the other consisting of wild varieties. Notably, the Fenghua duck and mallard populations showed signs of hybridization. When $K = 3$, Pekin and Cherry Valley Pekin ducks were separated from the domesticated varieties, resulting in three populations: the shelduck population (Shaoxing ducks, Gaoyou ducks, Jinding ducks, Shanma ducks, and Jianshui yellow–brown ducks), white-feathered meat-type duck population (Pekin ducks and Cherry Valley Pekin ducks), and wild duck population (mallard, spot-billed ducks, and Fenghua ducks). However, when $K = 4$, Jianshui yellow–brown ducks were further separated from the shelduck population. The results of genetic differentiation indicated that, among the 10 populations, the highest level of genetic differentiation was observed between Jinding ducks and Gaoyou ducks (Fst = 0.217), whereas the lowest level was observed between mallard and spot-billed ducks (Fst = 0.056; Supplementary Table S5). The Jianshui yellow–brown duck population exhibited the greatest degree of differentiation from the Gaoyou duck population (Fst = 0.209) and the smallest degree of differentiation from the Fenghua duck population (Fst = 0.139).

## 3.3 Population genetic diversity

Nucleotide diversity values are indicative of the genetic diversity within a population. Supplementary Figure S2 illustrates the genetic diversity levels of the 10 populations. Mallard exhibited the highest genetic diversity, while Gaoyou ducks displayed the lowest. The two wild duck populations showed relatively rich genetic diversity. The genetic diversity of the Jianshui yellow–brown duck population was lower than that of mallard, spot-billed ducks, Shaoxing ducks, and Fenghua ducks but higher than that of Pekin ducks, Shanma ducks, Jinding ducks, Cherry Valley Pekin ducks, and Gaoyou ducks.

10.3389/fvets.2023.1269904

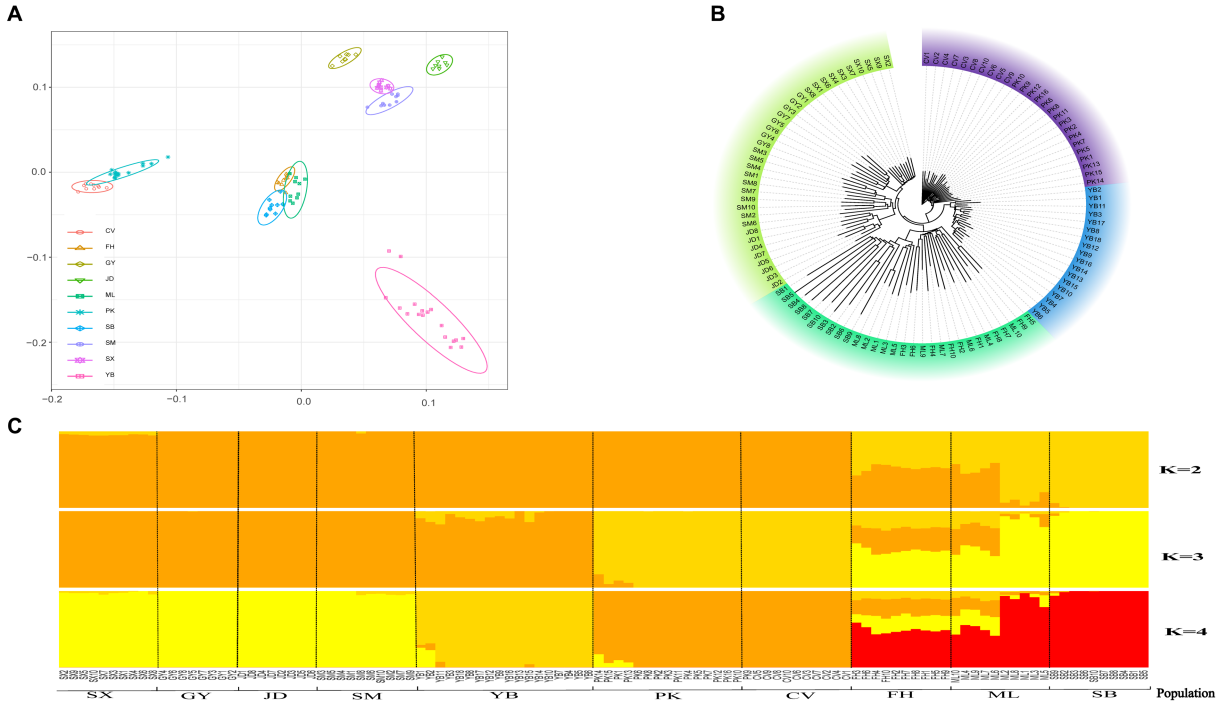



FIGURE 1
Population stratification of Jianshui yellow–brown duck. **(A)** Principal Component Analysis [Cherry Valley Pekin duck (CV), Fenghua duck (FH), Gaoyou duck (GY), Jinding duck (JD), Mallard (ML), Pekin duck (PK), Spot-billed duck (SB), Shanma duck (SM), Shaoxing duck (SX), Jianshui Yellow–brown duck (YB)]. **(B)** Neighbor-joining phylogenetic of eleven populations. **(C)** Population structure analysis.

Consequently, Jianshui yellow–brown ducks ranked fifth among the 10 populations in terms of genetic diversity.

In descending order, the values of genetic diversity were as follows: mallard, spot-billed ducks, Shaoxing ducks, Fenghua ducks, Jianshui yellow–brown ducks, Pekin ducks, Shanma ducks, Jinding ducks, Cherry Valley Pekin ducks, Gaoyou ducks. Two wild duck populations showed relatively rich genetic diversity. Among the eight domesticated populations, Jianshui yellow–brown ducks ranked third in terms of genetic diversity, after Fenghua ducks and Shaoxing ducks.

## 3.4 Selection signals

Meat performance is a crucial economic trait in the domestication of ducks. To investigate the selection of functional genes related to meat traits in varieties under different degrees of selection, the selection signals of Jianshui yellow–brown ducks, which undergo low selection, and Pekin ducks, which experience high selection, were analyzed with wild mallard as the control. Fst (population fixation index) and π (nucleotide diversity) were calculated for the whole genome regions of Jianshui yellow–brown duck, Pekin duck and mallards using the sliding window method with VCFTOOLS software. The window size and step size used in the calculation were in accordance with the parameters set in the section "2.2.3 Population selection analysis." Based on the above method, the obtained Fst values were converted into *Z*-scores, and $\log_2\theta\pi$ values were calculated for the compared populations. Figure 2 shows the results of the selection signals for Jianshui yellow–brown ducks and mallard. In performing the selection signal analysis, we screened genomic regions with ZFst and θπ values ranked in the top 1% as candidate regions that might be subject to selection. Specifically, we screened out regions in the genome of the Jianshui yellow–brown duck population that might be subjected to selection by setting a threshold of ZFst value >3.42 and $\log2(\pi_{\text{Mallard}}/\pi_{\text{Jianshui yellow-brown duck}}) \geq -0.55$. BEDTOOLS software was applied to merge the overlapping candidate regions, and in the results we identified 62 potential candidate genomic regions. After detailed annotation of genes within these regions, we identified a total of 136 genes that were subject to selection in Jianshui yellow–brown ducks.

To identify the selected genes in the Pekin duck population (Pekin duck vs. mallard), by calculating the thresholds for the ZFst and $\log_2\theta\pi$ values, we considered the regions with ZFst values >3.26 and $\log2(\pi_{\text{Mallard}}/\pi_{\text{Pekin duck}}) \geq -0.37$ as the regions that might be subject to selective action in the Pekin duck population (Figure 3), 90 selected genes were annotated within 63 candidate regions for Pekin ducks. Pekin ducks are globally recognized for their meat characteristics, while Jianshui yellow–brown ducks are a meat-egg-type variety with a lower degree of selective breeding. To compare the genomic differences between these two varieties, selection signal analysis was conducted on Jianshui yellow–brown ducks and Pekin ducks to identify selected genes related to meat performance in the Jianshui yellow–brown ducks population and evaluate the selective breeding of Jianshui yellow–brown ducks for meat traits. Through gene function query and Tajima's D calculation, 6 genes associated with growth and development and lipid metabolism – *ELOVL2* (ELOVL Fatty Acid Elongase 2), *ELOVL3* (ELOVL Fatty Acid Elongase 3), *GDF10* (Growth Differentiation Factor 10), *VSTM2A* (V-Set And Transmembrane Domain Containing 2A), *PHOSPHO1* (Phosphocholine Phosphatase 1), and *IGF2BP1* (Insulin Like Growth

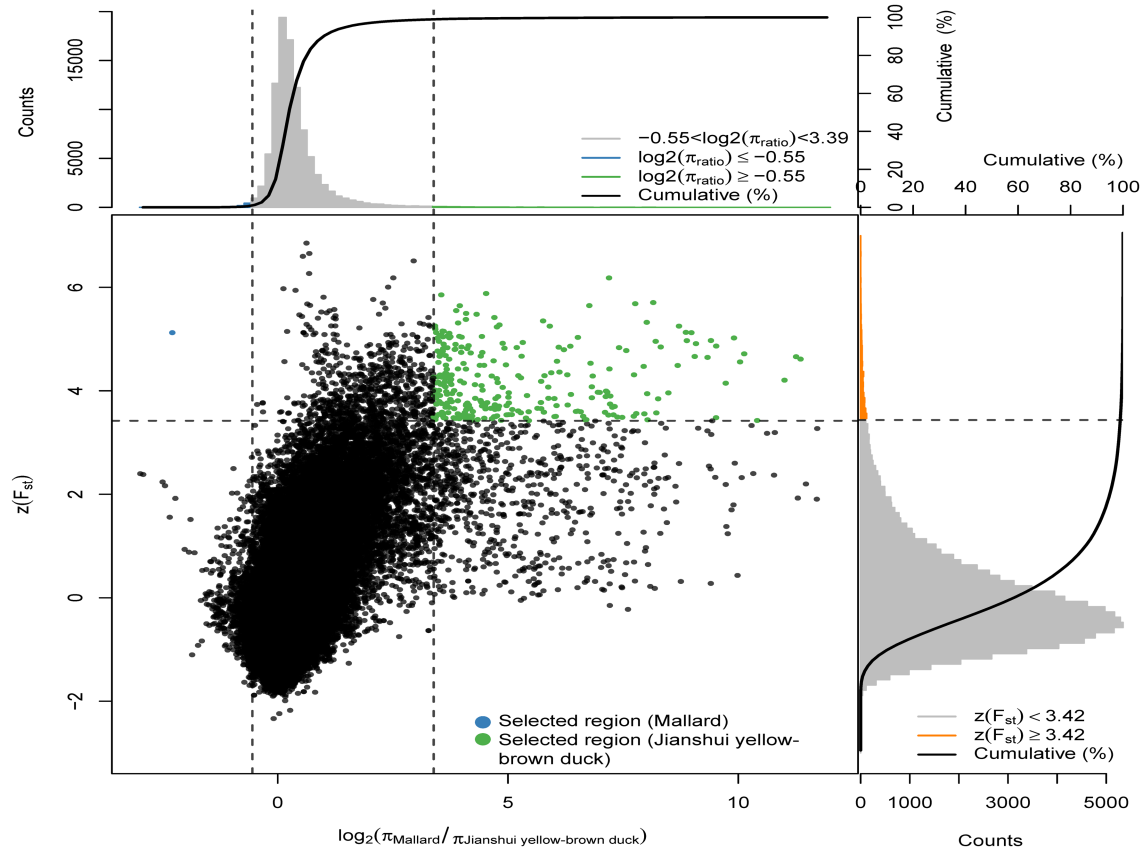



FIGURE 2
Distribution of *Z*(FST) values and log2(πMallard/Jianshui Yellow–brown duck) calculated in 40 kb sliding windows with 10 kb overlap between Jianshui yellow–brown duck and Mallard. Green dots represent selected regions on the genome of Jianshui yellow–brown duck, blue dots represent selected regions on the genome of Mallard.

Factor 2 MRNA Binding Protein 1) – were subjected to significant selection in the genome of Jianshui yellow–brown ducks (Supplementary Figures S3–S8). Furthermore, compared with Pekin and mallard, 6 genes related to muscle growth and fat deposition – *GDF10*, *IGF2BP1*, *SCD* (Stearoyl-CoA Desaturase), *MAP3K20* (Mitogen-Activated Protein Kinase Kinase Kinase 20), *TGFB3* (Transforming Growth Factor Beta 3), and *WNT8B* (Wnt Family Member 8B) – were identified (Supplementary Figures S9–S14). Compared with the Pekin duck population, 21 genes subjected to significant selection were observed in the Jianshui yellow–brown duck population (Supplementary Figure S15). However, based on functional annotation of these genes and comparison with the previously screened genes related to meat traits, none of these 21 genes were identified as candidate genes associated with muscle development and fat deposition.

## 3.5 Quality control of transcriptome raw data and screening of differentially expressed genes

Transcriptome sequencing technology facilitates the detection of mRNA expression in the same tissue across different duck varieties. This allows for the annotation and enrichment analysis of the functions of Differently Expressed genes (DEGs), which facilitates the identification of important functional genes related to characteristic traits in various varieties. In this study, transcriptome sequencing analysis was conducted on the liver, sebum and breast muscle tissues of Jianshui yellow–brown ducks and mallards. Additionally, DEGs were screened in the liver, sebum and breast muscle tissues between Pekin ducks and mallards. After screening and quality control of the raw data of transcriptome sequencing, and then comparing the high-quality reads back to the reference genome of ducks (CAU_duck1.0: https://ftp.ensembl.org/pub/release-110/fasta/anas_platyrhynchos_platyrhynchos/dna/Anas_platyrhynchos_platyrhynchos.CAU_duck1.0.dna.toplevel.fa.gz), the results showed that the transcriptome sequencing results of the tissue samples from the Jianshui yellow and brown duck population were of good quality, with an average Q20, Q30, and comparison rate of 92.61, 97.11, and 78.92%, respectively (Supplementary Table S6). The transcriptome sequencing results of Pekin duck and mallard duck downloaded from public databases also showed a high quality level, with average Q20, Q30 and alignment rates of 94.17, 97.48, and 81.01%, respectively (Supplementary Table S7). After quantification of gene expression, sample correlation analysis was performed on the normalized data using the TPM correction method. Subsequently, sequencing quality and sample data quality were preliminarily

10.3389/fvets.2023.1269904

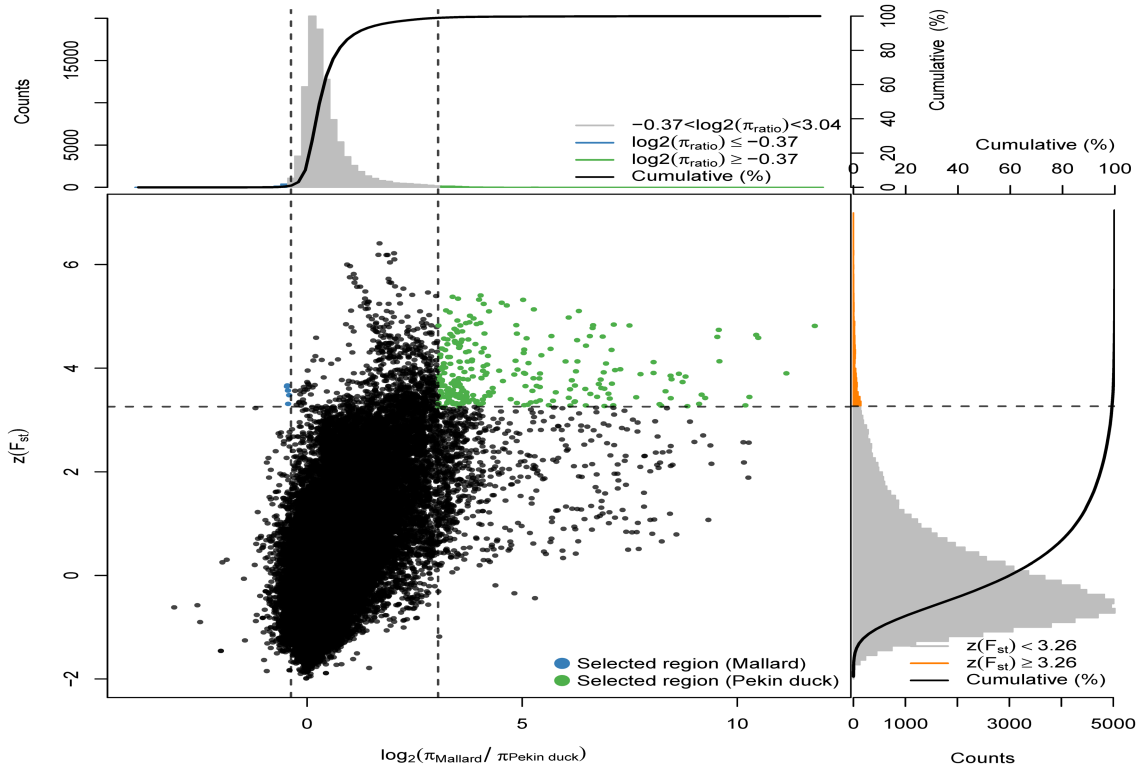



FIGURE 3
Distribution of *Z*(FST) values and log2(πMallard/rPekin duck) calculated in 40 kb sliding windows with 10 kb overlap between Pekin duck and Mallard. Green dots represent selected regions on the genome of Pekin duck, blue dots represent selected regions on the genome of Mallard.

examined through analyses of sequencing base quality values, alignment rates, and sample correlation. Sample correlation analysis was performed with six sets of transcriptome data obtained from the same tissue among different duck varieties. The results revealed the presence of abnormal data (Supplementary Figures S16–S17). To improve the credibility of the results, sample correlation analysis was repeated after removing the abnormal data, and the results of this analysis suggested that the intra-population correlation was higher than the inter-population correlation, and samples from the same variety could be clustered together (Figures 4, 5). Samples with high confidence were then used to screen for DEGs, using threshold values of $|\log2FoldChange| > 1.5$ and $Padj < 0.001$. According to the presentation in Supplementary Figures S18A–C, the transcriptomic data of Jianshui yellow–brown ducks and mallards showed that these two breeds exhibited different gene expression profiles in liver, sebum and breast muscle tissues. Specifically, there were 464 differentially expressed genes in liver tissues (251 up-regulated and 213 down-regulated), 1,435 differentially expressed genes in sebum tissues (958 up-regulated and 477 down-regulated), and 1,290 differentially expressed genes in breast muscle tissues (439 up-regulated and 851 down-regulated genes). After another transcriptome analysis of Pekin and mallard duck tissues (Supplementary Figures S18D–F), we identified 2,050 differentially expressed genes in liver tissues (532 genes were up-regulated and 1,518 genes were down-regulated), 473 differentially expressed genes in subcutaneous adipose tissues (including 295 genes up-regulated and 178 genes down-regulated), and 475 differentially expressed genes were screened out of breast muscle tissues (185 genes were up-regulated and 290 genes were down-regulated), differentially expressed genes between groups can be viewed in Supplementary Tables S8–S13.

## 3.6 Joint analysis results

Genotypic differences between populations within the region under selection may trigger changes in the expression of the gene. Multi-omics data were utilized to identify reliable artificial selection loci and functional genes associated with economically important traits. Five functional genes (*ELOVL2*, *ELOVL3*, *GDF10*, *VSTM2A*, and *IGF2BP1*) that may affect meat production and quality in Jianshui yellow–brown ducks were identified from genome-wide signal selection analysis. Transcriptome analysis of DEGs revealed significant differences in *PHOSPHO1* gene expression in the breast muscle tissues of Jianshui yellow–brown ducks and mallards (Figure 6). In addition, the expression levels of *ELOVL3* and *GDF10* genes also showed significant differences in sebum tissue (Figures 7, 8). To identify the functional genes related to muscle growth and fat deposition traits in Pekin ducks, and provide references for the selective breeding of meat-type duck varieties, genomic selection signal detection was performed on Pekin ducks and mallards. This analysis led to the identification of six functional genes related to meat traits (*GDF10*, *IGF2BP1*, *SCD*, *MAP3K20*, *TGFB3*, and *WNT8B*). Data on DEGs indicated significant differences in *TGFB3* gene expression in the liver tissue and *SCD* gene expression in both liver and sebum tissues (Figures 9, 10).

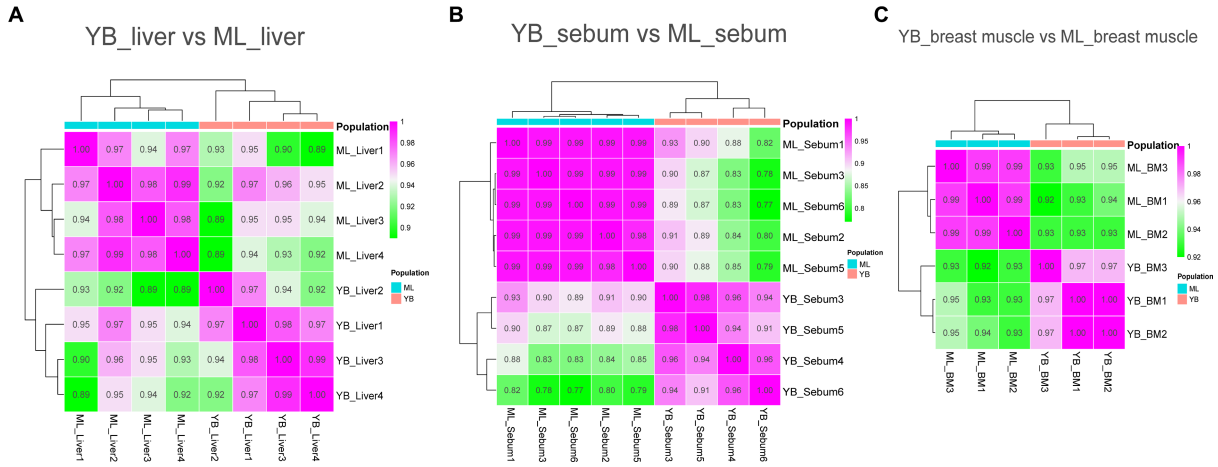



FIGURE 4
Sample correlation analysis of liver, sebum, and breast muscle transcriptome data of Jianshui yellow–brown ducks and Mallards after excluding further outlier samples. **(A)** Sample correlation thermograms of liver tissues from Jianshui yellow–brown ducks and mallards. **(B)** Sample correlation thermograms of sebum tissues from Jianshui yellow–brown ducks and Mallards. **(C)** Sample correlation thermograms of breast muscle tissues from Jianshui yellow–brown ducks and Mallards.

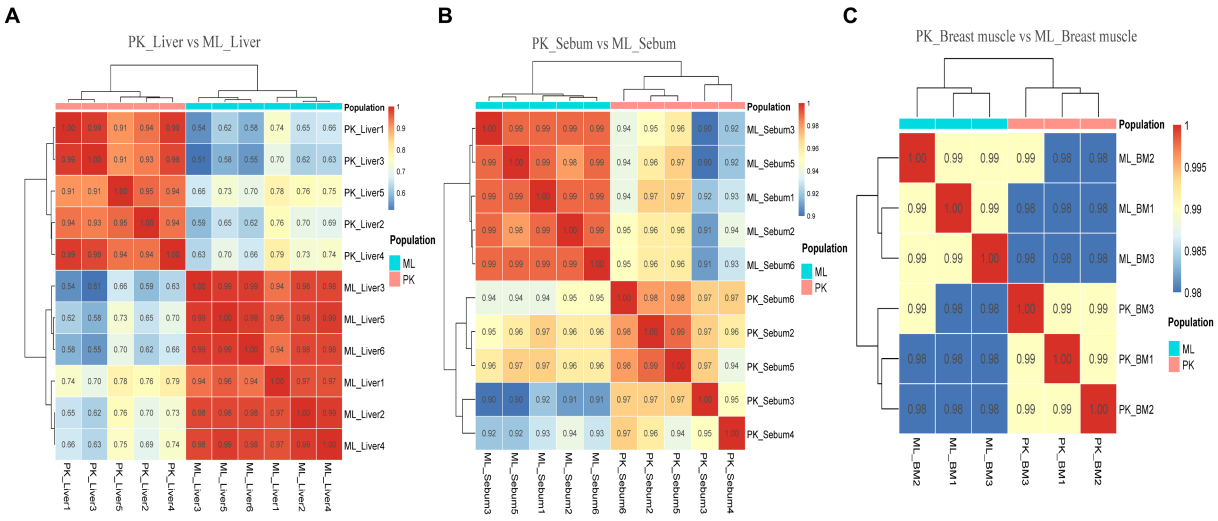



FIGURE 5
Sample correlation analysis of liver, sebum and breast muscle transcriptome data of Pekin ducks and Mallards after excluding further outlier samples. **(A)** Sample correlation thermograms of liver tissues from Pekin ducks and Mallards. **(B)** Sample correlation thermograms of sebum tissues from Pekin ducks and Mallards. **(C)** Sample correlation thermograms of breast muscle tissues from Pekin ducks and Mallards.

# 4 Discussion

## 4.1 Molecular genetic characterization of Jianshui yellow–brown ducks

To comprehensively and accurately assess the germplasm resources of Jianshui yellow–brown ducks, resequencing was performed on 18 Jianshui yellow–brown ducks. The resequencing data from the ducks in the NCBI database were also used for population structure analysis. PCA was performed based on genome-wide SNPs. The results revealed that the Jianshui yellow–brown duck population was clustered into a single category, and individuals of the same species could be clustered together without any outlier samples in the PCA plot. In the evolutionary tree, 110 individuals from 10 populations were categorized into four groups. Among them, Shaoxing ducks, Gaoyou ducks, Shanma ducks, and Jinding ducks were clustered into one category, and all four groups were shelduck varieties, originating from East China. Thus, they were implicated to have a closer genetic relationship than the other six populations. Mallards, spot-billed ducks, and Fenghua ducks were clustered into a single category. Although differences in body size and appearance existed between mallards and spot-billed ducks, the degree of genetic differentiation between the two varieties was small. Therefore, no reproductive isolation occurred after mating, and separating the two populations in the evolutionary tree was challenging (5, 24). In the relationship tree, Fenghua ducks and mallards showed a closer genetic

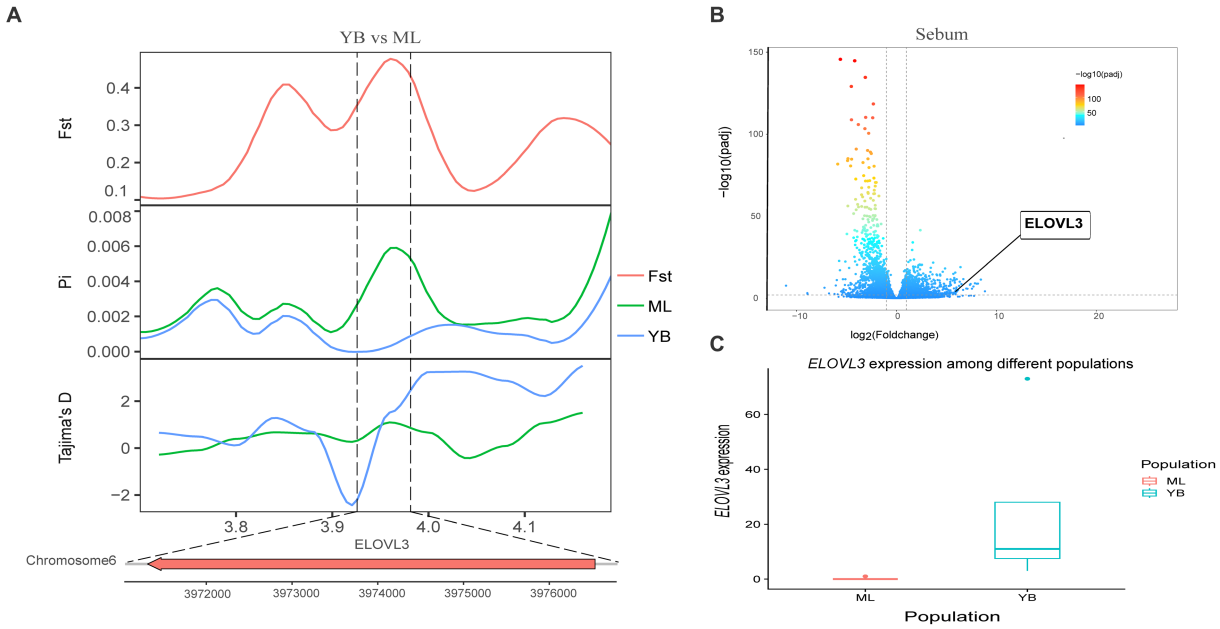



FIGURE 6
Combined analysis of genomic selection signal candidate genes and transcriptome differentially expressed genes in Jianshui yellow–brown duck and Mallard. **(A)** Fst Pi and Tajima's D value of *ELOVL3* between Jianshui Yellow–brown duck and Mallard. **(B)** Differentially expressed genes in sebum of Jianshui yellow–brown and Mallard. **(C)** Gene expression of *ELOVL3* in sebum tissue of Jianshui yellow–brown ducks and Mallards.

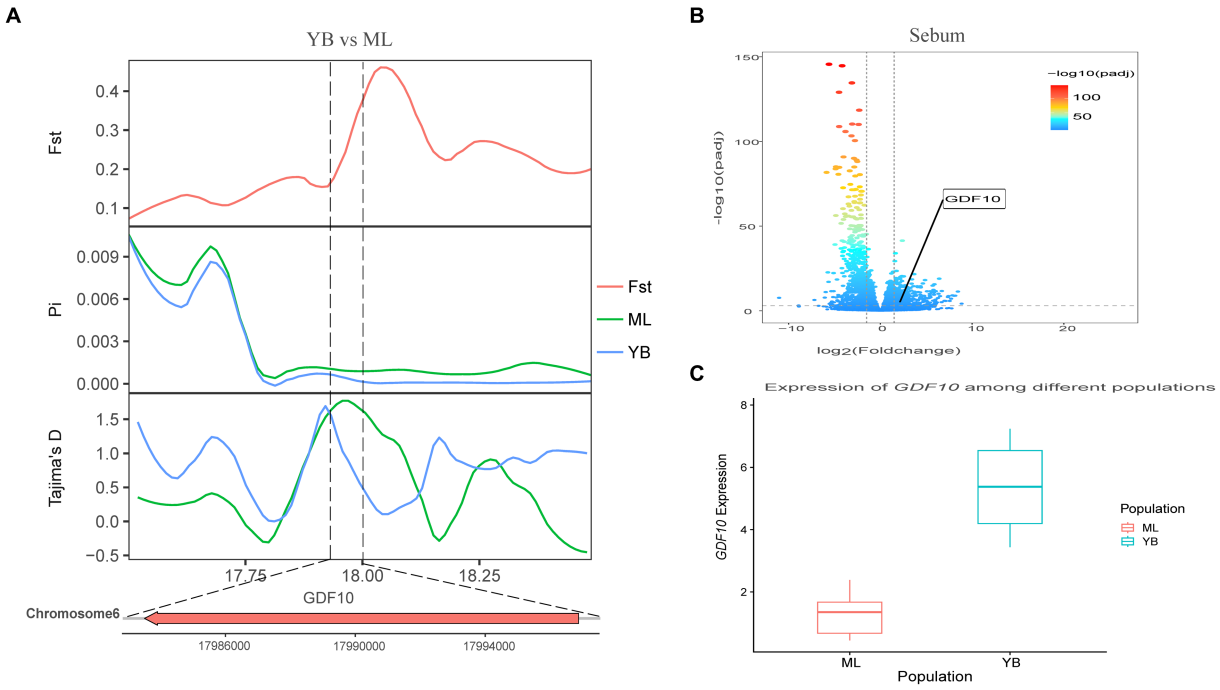



FIGURE 7
Combined analysis of genomic selection signal candidate genes and transcriptome differentially expressed genes in Jianshui yellow–brown duck and Mallard. **(A)** Fst Pi and Tajima's D value of *GDF10* between Jianshui Yellow–brown duck and Mallard. **(B)** Differentially expressed genes in sebum of Jianshui yellow–brown and Mallard. **(C)** Gene expression of *GDF10* in sebum tissue of Jianshui yellow–brown ducks and Mallards.

relationship. Fenghua ducks are directly domesticated from mallard ducks. Because of the short period of domestication, Fenghua ducks still retain some of the habits and characteristics of mallards (4, 25). Two white-feathered varieties, namely, Pekin duck and Cherry Valley Pekin duck, were clustered into a single category owing to their close genetic relationship. The Cherry Valley Pekin duck is a duck variety

10.3389/fvets.2023.1269904

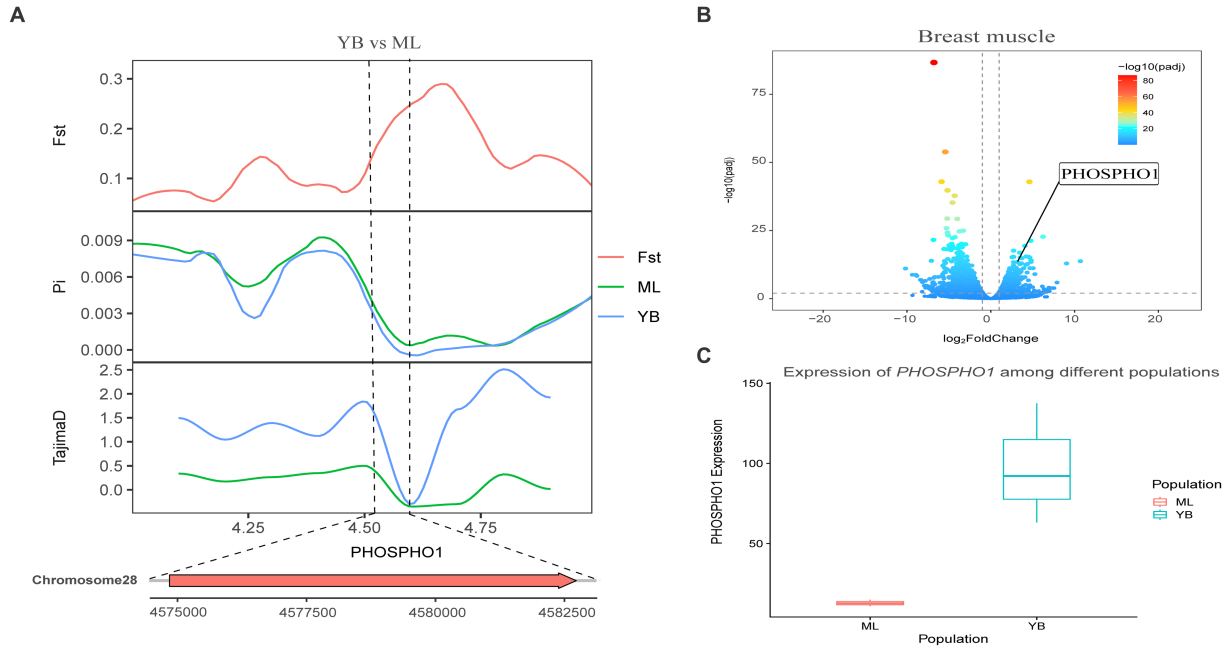



FIGURE 8
Combined analysis of genomic selection signal candidate genes and transcriptome differentially expressed genes in Jianshui yellow–brown duck and Mallard. **(A)** Fst Pi and Tajima's D value of *PHOSPHOI* between Jianshui Yellow–brown duck and Mallard. **(B)** Differentially expressed genes in breast muscle of Jianshui yellow–brown and Mallard. **(C)** Gene expression of *PHOSPHOI* in breast muscle tissue of Jianshui yellow–brown ducks and Mallards.

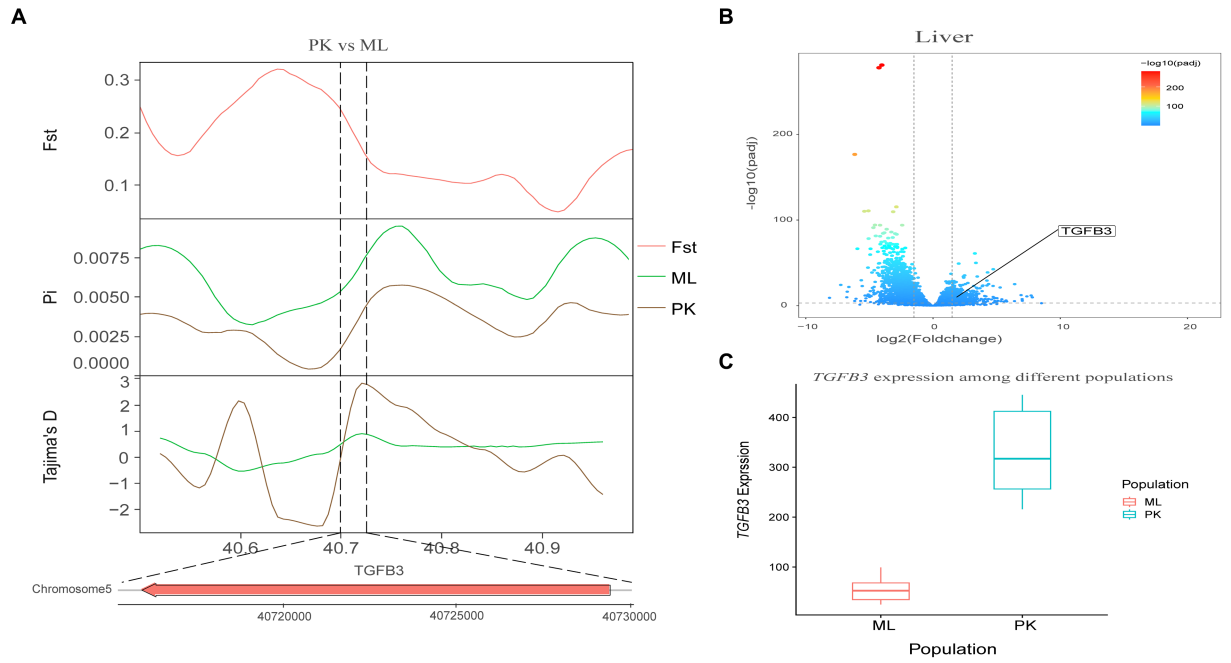



FIGURE 9
Combined analysis of genomic selection signal candidate genes and transcriptome differentially expressed genes in Pekin duck and Mallard. **(A)** Fst Pi and Tajima's D value of *TGFB3* between Pekin duck and Mallard. **(B)** Differentially expressed genes in liver of Pekin duck and Mallard. **(C)** Gene expression of *TGFB3* in liver tissue of Pekin ducks and Mallards.

known for its high-quality meat, and this variety was improved and bred on the Chinese native Pekin duck by the British Cherry Valley Company. Therefore, the Cherry Valley Pekin duck and the Pekin duck exhibited a close genetic relationship (3, 26). In addition, the Jianshui yellow–brown duck population was clustered into a separate category, a finding similar to those of the distribution of the Jianshui

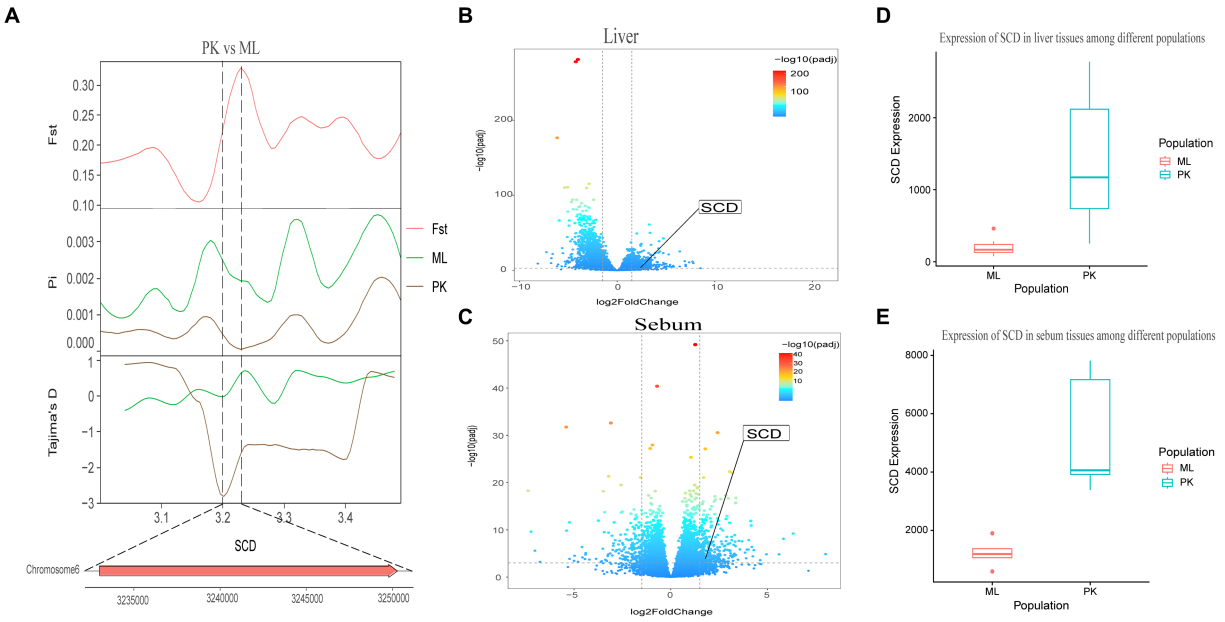



FIGURE 10
Combined analysis of genomic selection signal candidate genes and transcriptome differentially expressed genes in Pekin duck and Jianshui yellow–brown duck. **(A)** Fst Pi and Tajima's D value of *SCD* between Pekin duck and Mallard. **(B)** Differentially expressed genes in liver of Pekin duck and Mallard. **(C)** Differentially expressed genes in sebum of Pekin duck and Mallard. **(D)** Gene expression of *SCD* in liver tissue of Pekin ducks and Mallards. **(E)** Gene expression of *SCD* in sebum tissue of Pekin ducks and Mallards.

yellow–brown duck population in PCA. Compared with PCA and evolutionary tree analysis, structure analysis can be used to determine the reasonableness of grouping and gene exchange among different varieties. In structure analysis, when the cross-validation error rate was minimum (i.e., $K=4$), 10 populations were categorized into four groups. Jianshui yellow–brown ducks were separated from the shelduck population, forming a distinct population. This finding validated the results from PCA and evolutionary tree analysis, highlighting the differences in molecular genetic characteristics between Jianshui yellow–brown ducks and wild ducks, white-feathered meat-type ducks, and shelducks in East China. We hypothesized that the Jianshui yellow–brown ducks have developed a unique molecular genetic mechanism owing to the influence of natural and artificial selection.

## 4.2 Selection signal analysis of Jianshui yellow–brown ducks, Pekin ducks, and mallard

Jianshui yellow–brown ducks, known for their tender meat, have become the key ingredient for the specialty dish "Qujiang Roast Duck" in Jianshui County, Yunnan Province, China. Identifying candidate genes that affect the traits of these varieties is crucial for understanding the genetic basis of their characteristics and conserving germplasm resources. In the selective breeding process, selection signal analysis was conducted on Jianshui yellow–brown ducks and mallard to explore the functional genes influencing meat production and meat quality traits. Several candidate genes related to muscle quality were identified, including members of the elongase of the very-long-chain fatty acid (ELOVL) gene family, such as *ELOVL2* and *ELOVL3*, both of which were subjected to significant selection. The ELOVL family, as initiation and rate-limiting enzymes that catalyze fatty acid synthesis, plays an important role in fatty acid chain elongation (27). *VSTM2A* plays an essential role in regulating preadipocyte differentiation. *VSTM2A* overexpression increased the efficiency of fat synthesis, whereas *VSTM2A* knockdown inhibited fat formation (28, 29). Growth-related genes, including *GDF10*, *IGF2BP1*, and *PHOSPHO1* were also identified. Gene polymorphisms in *GDF10*, also known as bone morphogenetic protein-3b (BMP-3b), have shown a significant correlation with some growth traits in cattle and have a possible involvement in skeletal muscle formation and development. The *GDF10* gene is also involved in lipogenesis and metabolism (30). Based on the results of GWAS analysis, Zhou et al. identified *IGF2BP1* as a candidate gene affecting the growth and carcass of domestic ducks (20). Moreover, the study by Wang et al. on domestic chickens, which was based on the pan-genomic data of domestic ckickens, showed that high expression of the *IGF2BP1* gene influences the body size of ckickens (31). The *PHOSPHO1* is a member of the phosphate dehydrogenase family and encodes an enzyme belonging to the acid phosphatase class that catalyzes the hydrolysis of inorganic phosphate esters. This gene is well known for its key role in bone growth and bone mineralization in organisms (32, 33). Recently, experimental results from Peng et al. have also shown that the *PHOSPHO1* gene also plays an important role in the regulation of myogenic cell differentiation (34). These findings contribute to understanding the genetic mechanisms underlying the meat production and quality characteristics of Jianshui yellow–brown ducks.

Based on the analysis of population structure, it was found that the Jianshui yellow–brown duck and Pekin duck, two domesticated varieties, have significantly diverged from the wild mallard through domestication and selective breeding. In the selection signal analysis

between the Jianshui yellow–brown duck and mallard, candidate genes related to meat traits of the Jianshui yellow–brown duck were identified. However, as a meat-egg-type local variety, the meat performance of Jianshui yellow–brown ducks still requires further improvement. Therefore, the meat-type Pekin duck was chosen as the research object and the mallard as the control, and the selection signaling was used to identify candidate genes affecting muscle growth and fat deposition traits in Pekin duck. Six functional genes (*GDF10*, *IGF2BP1*, *SCD*, *MAP3K20*, *TGFB3*, and *WNT8B*) related to meat traits were identified. Among them, *SCD*, a key enzyme required for the synthesis of monounsaturated fatty acids, plays a crucial role in the regulation of lipogenesis. Mice with *SCD1* gene knockout have increased energy expenditure and fat consumption and decreased amounts of fat deposition (35). Bioinformatics approaches have been used for studying meat quality in livestock, and the results have suggested that the *SCD* gene is an important functional gene affecting lipogenesis and fat deposition (36, 37). *TGFB3*, a candidate gene affecting variation in rib count in pigs, also plays a role in avian lipogenesis and lipometabolism. Based on the GWAS-derived SNP association results affecting abdominal fat percentage in chickens, *TGFB3* was identified as a candidate functional gene affecting fat deposition (38). Wang et al. identified *TGFB3* as a target gene of miR-122, which is a regulator of adipose metabolism (39, 40). *MAP3K20* and *WNT8B*, belonging to the MAPK and Wnt signaling pathways (41, 42), respectively, are involved in cellular signaling pathways crucial for muscle growth and development. These genes may have a regulatory role in the muscle growth of Pekin ducks.

Jianshui yellow–brown duck, as a meat-egg-type local variety, has undergone limited selective breeding for meat traits. In contrast, the Pekin duck is a meat-type variety undergoing extensive selective breeding. To explore the genetic differences between these two varieties and assess the selection potential of Jianshui yellow–brown duck for meat performance, a selection signal analysis was performed. Only 21 genes were subjected to selection in the Jianshui yellow–brown duck population compared with those in the Pekin duck. When using mallard as a control, 136 and 90 genes were selected in the Jianshui yellow–brown duck and Pekin duck populations, respectively. It is worth noting that during the functional annotation of these 21 genes (YB vs. PK) selected from the Jianshui yellow–brown duck population, we also found an interesting gene, *ITGA4* (Integrin Subunit Alpha 4), which encodes the α4 integrin, which has been shown to play an important role in the regulation of inflammation and immune responses (43, 44). Jianshui yellow–brown ducks have always lived at relatively high altitudes and, due to their relatively short history of domestication, they exhibit strong environmental adaptations and disease resistance. Compared with the Pekin duck population, we observed that the *ITGA4* was strongly selected for in the genome, and hypothesized that the *ITGA4* plays an important role in both environmental adaptation and disease resistance in Jianshui yellow–brown ducks.

Among these selected genes, 6 genes were associated with muscle growth and fat deposition in Jianshui yellow–brown ducks, and 6 genes were related to meat traits in Pekin ducks. However, these genes were not detected in the Jianshui yellow–brown duck population when comparing selection signals between Jianshui yellow–brown ducks and Pekin ducks. Although the Jianshui yellow–brown duck variety has been bred for its meat characteristics for some time, the degree of its improvement in meat traits was not adequately high because of the short duration and low intensity of selective breeding. Therefore, further selective breeding efforts are needed to enhance meat yield and quality in Jianshui yellow–brown ducks.

### 4.3 Joint analysis of genomic selection signals and transcriptome

The selection signaling method has gained attention for identifying important functional genes related to economic traits in ducks (6, 45–47). By screening genes within regions subject to selection, which exhibit reduced polymorphism, higher inter-population differentiation, and changes in allele frequency, the accuracy of results can be improved (48). Combining multiple selection signal assays enhances the reliability of the findings. Additionally, the availability of upgraded reference genomes facilitates gene annotation within selected regions (49–51). Transcription of DNA into mRNA is a key event for proteins to exert their functions. However, gene expression during the transcription process is spatiotemporal specific, wherein different growth periods and feeding conditions affect the degree of gene expression (52, 53). Therefore, joint analysis of selection signals and transcriptome data provides a preliminary understanding of the effect of allelic changes in the selected region on gene expression and facilitates the identification of reliable selection sites and functional genes (1, 20). In this study, *ELOVL3*, *GDF10*, and *PHOSPHO1* were identified as important functional genes affecting muscle growth and fat deposition in Jianshui yellow–brown ducks through a joint analysis of selection signals and transcriptome data. Moreover, the selection signal analysis between Jianshui yellow–brown ducks and Pekin ducks revealed that genes related to meat traits in the Jianshui yellow–brown duck population were not strongly selected, indicating the need for further emphasis on selective breeding for meat traits. The *TGFB3* and *SCD* genes identified in Pekin ducks, associated with meat performance, may serve as potential markers for the molecular breeding of meat-type Jianshui yellow–brown ducks in the future.

## 5 Conclusion

The analysis of genetic structure based on whole-genome resequencing data revealed that the Jianshui yellow–brown duck represents a distinct population category with unique genetic mechanisms, possibly attributed to natural adaptation and artificial breeding. The integration of selection signaling and transcriptomic methods led to the identification of *ELOVL3*, *GDF10*, and *PHOSPHO1* as important functional genes influencing meat traits in Jianshui yellow–brown ducks. Compared with the selection signals in Pekin duck, Jianshui yellow–brown ducks have undergone less intense selective breeding for meat performance, necessitating further efforts to enhance their meat yield and quality. These findings provide valuable insights for the conservation and utilization of Jianshui yellow–brown duck germplasm resources and guide future selective breeding endeavors.

## Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: National Center for

Biotechnology Information (NCBI) BioProject, https://www.ncbi.nlm.nih.gov/bioproject/, PRJNA1000221 and PRJNA1000222.

## Ethics statement

The animal study was approved by Yunnan Agricultural University Laboratory Animal Welfare Ethics Committee; Acceptance No. 202103035. The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

XL: Formal analysis, Software, Visualization, Writing – original draft. AX: Supervision, Project administration, Writing – original draft. LM: Investigation, Software, Writing – original draft. XG: Conceptualization, Writing – original draft. SF: Methodology, Writing – original draft. XD: Methodology, Writing – original draft. BN: Software, Writing – original draft. LT: Software, Writing – original draft. LZ: Software, Writing – original draft. DY: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. XK: Funding acquisition, Project administration, Supervision, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Key Science and Technology Program of Yunnan Province (no. 202102AE090029) and China Agriculture Research System of MOF and MARA (CARS-42-54).

## Acknowledgments

Thank you for the support from Key Science and Technology Program of Yunnan Province (202102AE090029) and China Agriculture Research System of MOF and MARA (CARS-42-54) in this research. We also extend our gratitude to the members of Gou Xiao's research group at the School of Life Science and Engineering, Foshan University, for their assistance in conducting the experiments. Additionally, we would like to express our appreciation to Zhiyun Company for providing language polishing services for this paper.

## 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



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2023.1269904/full#supplementary-material

## References

1. Zhang ZB, Jia YX, Almeida P, Mank JE, van Tuinen M, Wang Q, et al. Whole-genome resequencing reveals signatures of selection and timing of duck domestication. *Gigascience*. (2018) 7:12. doi: 10.1093/gigascience/giy027

2. Zhu F, Yuan JM, Zhang ZH, Hao JP, Yang YZ, Hu SQ, et al. De novo transcriptome assembly and identification of genes associated with feed conversion ratio and breast muscle yield in domestic ducks. *Anim Genet*. (2015) 46:636–45. doi: 10.1111/age.12361

3. Zhang Z, Jia Y, Chen Y, Wang L, Lv X, Yang F, et al. Genomic variation in Pekin duck populations developed in three different countries as revealed by whole-genome data. *Anim Genet*. (2018) 49:132–6. doi: 10.1111/age.12639

4. Feng PS, Zeng T, Yang H, Chen GH, Du JP, Chen L, et al. Whole-genome resequencing provides insights into the population structure and domestication signatures of ducks in eastern China. *BMC Genomics*. (2021) 22:1. doi: 10.1186/s12864-021-07710-2

5. Guo X, He XX, Chen H, Wang ZC, Li HF, Wang JX, et al. Revisiting the evolutionary history of domestic and wild ducks based on genomic analyses. *Zool Res*. (2021) 42:43–50. doi: 10.24272/j.issn.2095-8137.2020.133

6. Gu H, Zhu T, Li X, Chen Y, Wang L, Lv X, et al. A joint analysis strategy reveals genetic changes associated with artificial selection between egg-type and meat-type ducks. *Anim Genet*. (2020) 51:890–8. doi: 10.1111/age.13014

7. Zhou H, Yang Y, Wang LX, Ye SQ, Liu JJ, Gong P, et al. Integrated multi-Omic data reveal the potential molecular mechanisms of the nutrition and flavor in Liancheng white duck meat. *Front Genet*. (2022) 13:12. doi: 10.3389/fgene.2022.939585

8. Zhang F, Yin ZT, Zhang JF, Zhu F, Hincke M, Yang N, et al. Integrating transcriptome, proteome and Qtl data to discover functionally important genes for duck eggshell and albumen formation. *Genomics*. (2020) 112:3687–95. doi: 10.1016/j.ygeno.2020.04.015

9. Chen SF, Zhou YQ, Chen YR, Gu J. Fastp: an ultra-fast all-in-one Fastq preprocessor. *Bioinformatics*. (2018) 34:i884–90. doi: 10.1093/bioinformatics/bty560

10. Huang YH, Li YR, Burt DW, Chen HL, Zhang Y, Qian WB, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genet*. (2013) 45:776–83. doi: 10.1038/ng.2657

11. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and Samtools. *Bioinformatics*. (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352

13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a Mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. (2010) 20:1297–303. doi: 10.1101/gr.107524.110

14. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and Vcftools. *Bioinformatics*. (2011) 27:2156–8. doi: 10.1093/bioinformatics/btr330

15. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics*. (2007) 23:2633–5. doi: 10.1093/bioinformatics/btm308

16. Saitou N, Nei M. The neighbor-joining method – a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. (1987) 4:406–25. doi: 10.1093/oxfordjournals.molbev.a040454

17. Letunic I, Bork P. Interactive tree of life (Itol) V5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. (2021) 49:W293–6. doi: 10.1093/nar/gkab301

18. 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

19. Wang R, Sun JL, Han H, Huang YF, Chen T, Yang MM, et al. Whole-genome resequencing reveals genetic characteristics of different duck breeds from the Guangxi

region in China. *G3-Genes Genomes Genetics*. (2021) 11:54. doi: 10.1093/g3journal/jkab054

20. Zhou ZK, Li M, Cheng H, Fan WL, Yuan ZR, Gao Q, et al. An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat Commun*. (2018) 9:10. doi: 10.1038/s41467-018-04868-4

21. Kim D, Landmead B, Salzberg SL. Hisat: a fast spliced aligner with low memory requirements. *Nat Methods*. (2015) 12:357–60. doi: 10.1038/nmeth.3317

22. Liao Y, Smyth GK, Shi W. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. (2014) 30:923–30. doi: 10.1093/bioinformatics/btt656

23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for Rna-Seq data with Deseq2. *Genome Biol*. (2014) 15:550. doi: 10.1186/s13059-014-0550-8

24. Zhang Y, Bao Q, Cao Z, Bian YQ, Zhang Y, Cao ZF, et al. Chinese domestic ducks evolved from mallard duck (Anas Platyrhynchos) and spot-billed duck (A. zonorhyncha). *Animals*. (2023) 13:71156. doi: 10.3390/ani13071156

25. Tian Y, Li GQ, Du XZ, Zeng T, Chen L, et al. Integration of Lc-Ms-based and Gc-Ms-based metabolic profiling to reveal the effects of domestication and boiling on the composition of duck egg yolks. *Meta*. (2023) 13:22. doi: 10.3390/metabo13010135

26. Kokoszynski D, Wasilewski R, Steczny K, Kotowicz M, Hrncar C, Arpasova H. Carcass composition and selected meat quality traits of Pekin ducks from genetic resources flocks. *Poult Sci*. (2019) 98:3029–39. doi: 10.3382/ps/pez073

27. Zhang JY, Kothapalli KSD, Brenna JT. Desaturase and Elongase-limiting endogenous Long-chain polyunsaturated fatty acid biosynthesis. *Curr Opin Clin Nutr Metab Care*. (2016) 19:103–10. doi: 10.1097/mco.0000000000000254

28. Al Dow M, Silveira MAD, Poliquin A, Tribouillard L, Fournier E, Trebaol E, et al. Control of Adipogenic commitment by a Stat3-Vstm2a Axis. *Am J Physiol*. (2021) 320:E259–69. doi: 10.1152/ajpendo.00314.2020

29. Secco B, Camire E, Briere MA, Caron A, Billong A, et al. Amplification of Adipogenic commitment by Vstm2a. *Cell Rep*. (2017) 18:93–106. doi: 10.1016/j.celrep.2016.12.015

30. Marti-Pamies I, Thoonen R, Seale P, Vite A, Caplan A, Tamez J, et al. Deficiency of bone morphogenetic protein-3b induces metabolic syndrome and increases adipogenesis. *Am J Physiol*. (2020) 319:E363–75. doi: 10.1152/ajpendo.00362.2019

31. Wang KJ, Hu HF, Tian YD, Li JY, Scheben A, Zhang CX, et al. The chicken Pan-genome reveals gene content variation and a promoter region deletion in Igf2bp1 affecting body size. *Mol Biol Evol*. (2021) 38:5066–81. doi: 10.1093/molbev/msab231

32. MacRae VE, Davey MG, McTeir L, Narisawa S, Yadav MC, Millan JL, et al. Inhibition of Phospho1 activity results in impaired skeletal mineralization during limb development of the Chick. *Bone*. (2010) 46:1146–55. doi: 10.1016/j.bone.2009.12.018

33. Millán JL. The role of phosphatases in the initiation of skeletal mineralization. *Calcif Tissue Int*. (2013) 93:299–306. doi: 10.1007/s00223-012-9672-8

34. Peng Y, Yue F, Chen JJ, Xia W, Huang KL, et al. Phosphatase orphan 1 inhibits myoblast proliferation and promotes myogenic differentiation. *FASEB J*. (2021) 35:e21154. doi: 10.1096/fj.202001672R

35. Ntambi JM, Miyazaki M, Stoehr JP, Lan H, Kendziorski CM, Yandell BS, et al. Loss of Stearoyl-Coa Desaturase-1 function protects mice against adiposity. *Proc Natl Acad Sci U S A*. (2002) 99:11482–6. doi: 10.1073/pnas.132384699

36. Du LL, Chang TP, An BX, Liang M, Deng TY, et al. Transcriptomics and lipid metabolomics analysis of subcutaneous, visceral, and abdominal adipose tissues of beef cattle. *Genes*. (2023) 14:20. doi: 10.3390/genes14010037

37. Zhang YF, Zhang JJ, Gong HF, Cui LL, Zhang WC, Ma JW, et al. Genetic correlation of fatty acid composition with growth, carcass, fat deposition and meat quality traits based on Gwas data in six pig populations. *Meat Sci*. (2019) 150:47–55. doi: 10.1016/j.meatsci.2018.12.008

38. Abasht B, Lamont SJ. Genome-wide association analysis reveals cryptic alleles as an important factor in Heterosis for fatness in chicken F-2 population. *Anim Genet*. (2007) 38:491–8. doi: 10.1111/j.1365-2052.2007.01642.x

39. Wang XG, Shao F, Yu JF, Jiang HL, Gong DQ, Gu ZL. Microrna-122 targets genes related to liver metabolism in chickens. *Comp Biochem Physiol B*. (2015) 184:29–35. doi: 10.1016/j.cbpb.2015.02.002

40. Hicks JA, Trakooljul N, Liu HC. Discovery of chicken microrna associated with lipogenesis and cell proliferation. *Physiol Genomics*. (2010) 41:185–93. doi: 10.1152/physiolgenomics.00156.2009

41. Girardi F, Le Grand F. Wnt signaling in skeletal muscle development and regeneration. J Larrain and G Olivares, (Eds.) *Wnt signaling in health and disease. Progress in Molecular Biology and Translational Science*. *153*. San Diego: Elsevier Academic Press Inc (2018). 157–179

42. Keren A, Tamir Y, Bengal E. The P38 Mapk signaling pathway: a major regulator of skeletal muscle development. *Mol Cell Endocrinol*. (2006) 252:224–30. doi: 10.1016/j.mce.2006.03.017

43. Genbacev O, Larocque N, Ona K, Prakobphol A, Garrido-Gomez T, Kapidzic M, et al. Integrin A4-positive human trophoblast progenitors: functional characterization and transcriptional regulation. *Hum Reprod*. (2016) 31:1300–14. doi: 10.1093/humrep/dew077

44. Bernsmeier C, van der Merwe S, Périanin A. Innate immune cells in cirrhosis. *J Hepatol*. (2020) 73:186–201. doi: 10.1016/j.jhep.2020.03.027

45. Wang L, Guo JZ, Xi Y, Ma SC, Li YY, He H, et al. Understanding the genetic domestication history of the Jianchang duck by genotyping and sequencing of genomic genes under selection. *G3-Genes Genomes Genetics*. (2020) 10:1469–76. doi: 10.1534/g3.119.400893

46. Zhu T, Qi X, Chen Y, Wang L, Lv XZ, Yang WF, et al. Positive selection of skeleton-related genes during duck domestication revealed by whole genome sequencing. *BMC Ecol Evol*. (2021) 21:8. doi: 10.1186/s12862-021-01894-7

47. Wang LY, Li XX, Ma J, Zhang YW, Zhang H. Integrating genome and transcriptome profiling for elucidating the mechanism of muscle growth and lipid deposition in Pekin ducks. *Sci Rep*. (2017) 7:11. doi: 10.1038/s41598-017-04178-7

48. Qanbari S, Simianer H. Mapping signatures of positive selection in the genome of livestock. *Livest Sci*. (2014) 166:133–43. doi: 10.1016/j.livsci.2014.05.003

49. Li R, Fu WW, Su R, Tian XM, Du D, et al. Towards the complete goat Pan-genome by recovering missing genomic segments from the reference genome. *Front Genet*. (2019) 10:11. doi: 10.3389/fgene.2019.01169

50. Tian XM, Li R, Fu WW, Li Y, Wang XH, Li M, et al. Building a sequence map of the pig Pan-genome from multiple De novo assemblies and hi-C data. *Sci China-Life Sci*. (2020) 63:750–63. doi: 10.1007/s11427-019-9551-7

51. Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biol*. (2020) 21:184. doi: 10.1186/s13059-020-02105-0

52. Liu YH, Jia YX, Liu C, Ding LM, Xia ZF. Rna-Seq transcriptome analysis of breast muscle in Pekin ducks supplemented with the dietary probiotic Clostridium Butyricum. *BMC Genomics*. (2018) 19:14. doi: 10.1186/s12864-018-5261-1

53. Hu ZG, Cao JT, Ge LY, Zhang JQ, Zhang HL, Liu XL. Characterization and comparative transcriptomic analysis of skeletal muscle in Pekin duck at different growth stages using Rna-Seq. *Animals*. (2021) 11:20. doi: 10.3390/ani11030834

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 January 2024
DOI 10.3389/fgene.2023.1304638



OPEN ACCESS

EDITED BY
Ana Fabrícia Braga Magalhães,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Xiaolong Kang,
Ningxia University, China
Thaise Melo,
Federal University of Santa Maria, Brazil

*CORRESPONDENCE
Jaemin Kim,
jmkim85@gnu.ac.kr
Woncheoul Park,
wcpark1982@korea.kr

RECEIVED 29 September 2023
ACCEPTED 18 December 2023
PUBLISHED 09 January 2024

CITATION
Jang S, Jang S, Kim J and Park W (2024), Multi-tissue transcriptome analysis to identify candidate genes associated with weight regulation in Hanwoo cattle. *Front. Genet.* 14:1304638. doi: 10.3389/fgene.2023.1304638



# Multi-tissue transcriptome analysis to identify candidate genes associated with weight regulation in Hanwoo cattle

Subin Jang[1,2], Sunsik Jang[3], Jaemin Kim[1,2]* and Woncheoul Park[4]*

[1]Division of Applied Life Science (BK21), Gyeongsang National University, Jinju, Republic of Korea, [2]Institute of Agriculture and Life Sciences, Gyeongsang National University, Jinju, Republic of Korea, [3]Hanwoo Research Institute, National Institute of Animal Science, Rural Development Administration, Pyeongchang, Republic of Korea, [4]Animal Genomics and Bioinformatics Division, National Institute of Animal Science, Rural Development Administration, Wanju, Republic of Korea

While genetic markers related to meat production traits have been identified in many other cattle breeds, research on weight in Hanwoo cattle (Korean native cattle) is still insufficient. In this study, we performed expression quantitative trait loci (eQTL) analysis and differential gene expression analysis to detect candidate genes influencing the weight characteristics of 32 castrated Hanwoo cattle across 22 tissues and, we identified variants that affect gene expression levels. In total, we identified a total of 3,298 differentially expressed genes, among which we discovered key genes such as *UBD*, *RGS2*, *FASN*, and *SCD* that have functions related to adipogenesis, body weight, obesity, and lipid metabolism. Gene-set enrichment analysis revealed that candidate genes in adipose tissue are involved in metabolic pathways linked to obesity-related traits, adipose metabolism, and lipid metabolism. Additionally, we found that decreased expression of *TRIM31* contributes to weight gain which can be explained by the associated candidate *cis*-eQTL genotypes for *TRIM31* and their effect on differential gene expression between the lower and higher weight groups. Our findings revealed candidate genes associated with the weight of Hanwoo cattle and perhaps can provide comprehensive insights into the association of weight with various tissues beyond adipose tissue and muscle, indicating the potential for expanding the focus of livestock trait research.

KEYWORDS

**Hanwoo, body weight, eQTL analysis, differentially expressed genes, transcriptome**

## Introduction

Hanwoo cattle (Korean native cattle) is a breed of cattle indigenous to Korea that was previously used for agricultural, transportation and religious purposes but later evolved into beef cattle and remains one of the country's most important food sources to this day (Lee et al., 2014). Hanwoo cattle are recognized for their high fertility, but their slow growth rate hinders their meat production capability (Choi et al., 2019). For efficient meat production of beef cattle, it is important to maximize their weight, which is an economic trait (Fink et al., 2017). Livestock weight is economically important because it indicates livestock ability, a standard for determining livestock rations and selling prices, and is also used as a trait to evaluate livestock breeding value (Wangchuk et al., 2018). To this end, a significant amount of genetic research has focused on

elucidating the genetic determinants of body weight and related traits in cattle and other livestock species (Littlejohn et al., 2012).

Genetic studies have progressively shed light on the complex underpinnings of body weight traits in cattle. For instance, Duan et al. (2021) pinpointed several single nucleotide polymorphisms (SNPs) in Chinese Simmental cattle that correlate with body weight, revealing genes that govern growth and development. Similarly, Naserkheil et al. (2020) conducted a GWAS that identified genes in Hanwoo cattle associated with metabolic processes and growth, highlighting the genetic complexity of these traits. Furthermore, Liu et al. (2018) provided insight into the role of lipid metabolism by analyzing differentially expressed genes (DEGs) in the subcutaneous adipose tissue of Lilu cattle, underscoring the multifaceted nature of weight regulation at the genetic level.

While these studies offer valuable information, they primarily rely on GWAS for gene identification, and little attention has been given to integrating transcriptomic data with expression quantitative trait loci (eQTL) mapping to gain a more holistic understanding of weight traits. Recognizing this gap, our study harnesses RNA-seq technology to profile gene expression across Hanwoo cattle tissues, integrating eQTL analysis to elucidate the genetic mechanisms influencing weight. The eQTL analysis is crucial in elucidating the significant association between gene expression levels and genetic polymorphisms, providing a profile that highlights the unique biological significance (Peng et al., 2018; Cai et al., 2023). This comprehensive approach aims to build on the existing genetic framework, adding depth to our understanding of how genetic variations contribute to phenotypic expressions related to body weight in Hanwoo cattle.

In recent studies on weight and body composition traits of Chinese Lilu cattle and common beef breeds in the United States (such as angus, beefmaster, brahman, etc.), identified significant candidate genes have been reported to be involved in lipid metabolism pathways (Liu et al., 2018; Lindholm-Perry et al., 2020). Lipid metabolism, which contributes to the characteristics of body weight, encompasses a range of biochemical pathways including fat synthesis, lipolysis, lipid transport, and oxidation (Muradian et al., 2015). These processes occur not only in adipose tissue but also in various other tissues such as the brain, liver, and muscles (Adibhatla and Hatcher, 2008; Yang et al., 2013; Zhang et al., 2022). Therefore, the aim of this study was to perform expression quantitative trait loci (eQTL) analysis and differential gene expression analysis in different tissues to detect candidate genes influencing weight in Hanwoo cattle.

# Materials and methods

## Experimental overview and sample collection

To identify expression quantitative trait loci (eQTL) in Hanwoo cattle, 32 animals from the same farm were provided by the Hanwoo Cattle Research Institute, National Institute of Animal Science, South Korea. The age (Mean ± Sd, 15.6 ± 5.5) and body weight (Mean ± Sd, 388.9 ± 115.6) of the 32 samples were measured at the time of slaughter (Supplementary Table S1). The 22 tissues collected for RNA-sequencing are as follows: abdominal fat (ABF), abomasum (ABO), back fat (BFT), blood (BLO), cecum (CEC), colon (COL), duodenum (DUO), heart (HEA), ileum (ILE), jejunum (JEJ), kidney (KID), kidney fat (KIF), liver (LIV), sirloin (LOM), lung (LUN), omasum (OMA), rectum (REC), reticulum (RET), round (RMP), rumen (RUM), spleen (SPL) and tenderloin (TEN). Three of 32 individuals (Sample IDs: 192018, 192032, 202012) included missing tissue samples. Information about the tissues collected per individual is provided in Supplementary Table S2. Ethics approval was obtained from the National Institute of Animal Science (approval no: NIAS20201979).

## RNA isolation and sequencing

Tissue samples harvested from 32 castrated Hanwoo cattle were processed for RNA preparation using two distinct methods. The first method involved RNA extraction following the Trizol Beating RLT Dnase column protocol, utilizing the QIAamp 96 Viral RNA Kit in conjunction with QIAzol Lysis Reagent. The second method extracted RNA based on the Trizol beating isopropanol column DNase tissue RNA protocol, employing the QIAamp DNA Mini Kit and QIAzol Lysis Reagent. Additionally, for the blood samples, RNA was extracted by referencing the 900 μL Trizol Isopropanol column protocol, using a combination of QIAzol_3X and QIAzol Lysis Reagent. RNA concentration was checked using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, United States). RNA extracted from tissues and blood was all subjected to RNA QC using the TapeStation RNA Screen Tape, and the criteria were Concentrations (total amount) > 0.5 (ug), RINs value > 6, and rRNA ratio > 1. The quality and the integrity of the RNA was assessed using a bioanalyzer (Agilent, Santa Clara, United States) and only samples with a RIN value greater than 8.0 was used for cDNA library construction. Individual libraries were generated using Illumina TruSeq™ RNA Sample Preparation Kit (Illumina, San Diego, CA, United States). All samples were sequenced on the Illumina NovaSeq 6000 sequencer, generating 100bp paired-end reads at a sequencing depth of 6 Gb. Sequencing for all samples was conducted across separate lanes as per the workflow schedule, rather than being performed on a single lane. The goal was to produce data of at least 60 million reads for each sample. The raw reads were freely deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under accession number E-MTAB-13398.

## RNA-seq data production and RNA SNP calling

We sought to obtain quantified expression values transcripts per million (TPM) to compare gene expression levels and for use in eQTL analysis. RNA-seq data of the 22 tissues sourced from 32 samples were quality-checked with FastQC (version 0.11.9) (Brown et al., 2017), and low-quality reads were filtered through the Trimmomatic (version 0.39) process (Bolger et al., 2014). Expression levels were quantified using the rsem-calculate-expression function of the RNA-seq by expectation maximization (RSEM) software (version 1.3.1) (Li and Dewey, 2011), generating TPM values for each of the 22 tissues across the 32 samples. The ARS-UCD1.2 of cattle was used as the reference genome.

To acquire single nucleotide polymorphisms (SNP) information requisite for principal component analysis (PCA) and eQTL analysis, SNP data was derived from RNA-seq datasets. RNA-seq data from 32 samples of BLO tissue were subjected to SNP calling using genome analysis toolkit (GATK) (version 4.1.4.0) (DePristo et al., 2011)

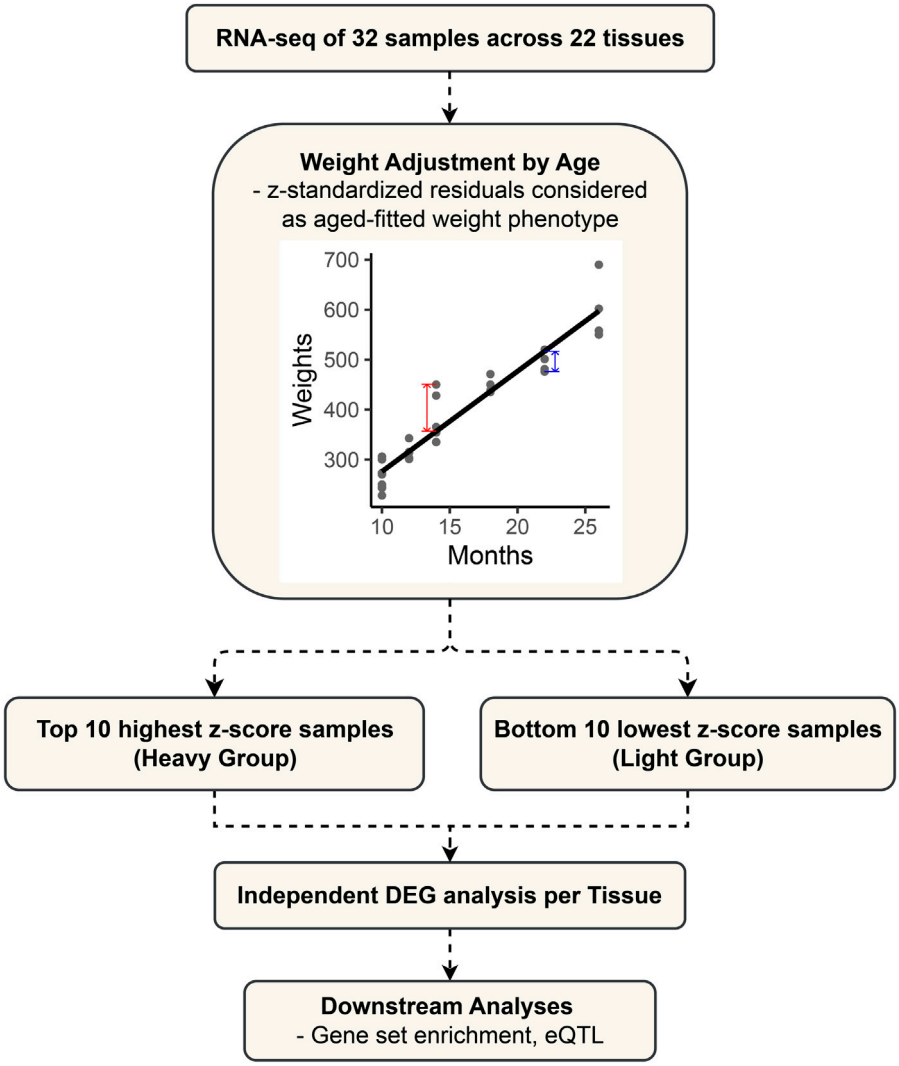



FIGURE 1
Flowchart outlining the study design and analytical approach for differential expression gene (DEG) analysis in cattle.

adhering to the best practice guidelines (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels). First, after the previously mentioned Trimmomatic process, we mapped the reference using the spliced transcripts alignment to a reference (STAR) (version 0.11.9) tool (Dobin et al., 2013) and removed duplicates using the MarkDuplicatesSpark tool in GATK. After going through the SplitNCigarReads process of the GATK and base quality recalibration, variant calling was performed using the HaplotypeCaller tool. To enhance the accuracy of RNA SNP variants, we treated genotypes with a genotype quality (GQ) less than 20 and a read depth (DP) less than 5 as missing. SNPs exceeding a 10% missing rate (-geno 0.1) were filtered. We then excluded the sex chromosomes and indels, focusing our analysis solely on autosomal SNPs.

## Study design

The 32 samples varied in age, as detailed in Supplementary Table S1. For an accurate analysis, the weight phenotype was adjusted for the effect of age (Figure 1). We adjusted the weight phenotypes for age by simple linear regression and standardized the residuals to z-scores by using the "lm" function in the R (version 4.2.2) software (Ihaka and Gentleman, 1996). The heavy group consists of the top 10 samples with the highest z-scores after age fitting, while the light group comprises the bottom 10 samples. They represent groups at the ends of a continuum in the age-adjusted weight phenotype. For each tissue, the transcriptomic comparison between the two defined groups was performed to identify differentially expressed genes.

## Principal component analysis for genetic similarity and pattern identification

To ensure the independence of the samples and to detect any potential correlations resulting from genetic or environmental influences, we performed principal component analysis (PCA). The PCA was executed using the genome-wide complex trait analysis tool (GCTA64) (Yang et al., 2011), utilizing variant call

format (VCF) files obtained from RNA sequencing data of BLO (Granato et al., 2018). The VCF files were first converted to Plink format with Plink software for compatibility with the GCTA64 tool. Subsequent analyses were restricted to autosomal chromosomes to avoid confounding factors associated with sex chromosomes. The PCA was computed using the "--pca 21" option to extract the first 21 principal components. The principal components (PC1 and PC2) were adjusted as covariates in the differential gene expression (DEG) analysis to mitigate potential confounding effects.

## Differential gene expression analysis and inter-tissue correlation analysis

In the RNA-seq process, the number of reads of each gene was calculated using the FeatureCount (version 2.0.1) program (Liao et al., 2014) for the bam file generated after the STAR process (Dobin et al., 2013). Read counts were converted to counts per million (CPM), and genes with maximum CPM values less than 1 in the samples were removed. When specifying the model to be fitted, the values of PC1 and PC2 were applied as covariates to control for external factors. Differentially expressed genes (DEGs) were identified using the limma (version 3.52.4) -voom package (Law et al., 2014) of R software (Ihaka and Gentleman, 1996), and 10 samples with high z-scores were grouped as cases (= heavy group), and 10 samples with low z-scores were grouped as controls (= light group) for comparative analysis. DEGs were identified based on a $p$-value < 0.05 and $|\text{Log}_2\text{FC}| > 1$. In the reference genome ARS-UCD1.2 obtained with Ensembl's BioMart tool (https://asia.ensembl.org/index.html), overlapping analysis was performed using the Ensembl Gene ID to identify candidate DEGs.

The comprehensive assessment of gene expression correlations across all tissues was conducted using t-values derived from the limma-voom package, which was employed in the DEG analysis. To test the correlation between tissues, we utilized the Pearson coefficient test. Using the "corrplot (version 0.92)" function in R package, we distinguished and visualized three primary clusters through the hierarchical clustering algorithm (hclust).

## *cis*-eQTL analysis

Following the correlation analysis, which identified three adipose tissues [ABF ($n$ = 342), BFT ($n$ = 195), KIF ($n$ = 160)] as significant clusters, we proceeded with an eQTL analysis aimed at uncovering variants that influence the regulatory mechanisms of gene expression for DEGs within these specified tissues. First, the previously generated vcf file, which only included SNP calling, was converted into PLINK format using VCFtools (version 0.1.13) (Danecek et al., 2011) and PLINK (version 1.90b6.24) (Purcell et al., 2007). The phenotypic dataset was represented by the TPM expression levels for each sample across all identified DEGs within the three examined tissues. Genome-wide efficient mixed model association (GEMMA) (version 0.98.5) (Zhou and Stephens, 2012) was utilized to perform association tests between the expression levels of DEGs and genotypes. Our selection criteria for candidate *cis*-eQTL variants entailed choosing SNPs that were not only on the same chromosome as their corresponding gene but also within a proximity of 500 Kb to the gene's transcription start or polyadenylation site, with an emphasis on those exhibiting a $p$-value below 5e-05. It was run on each gene separately to identify variants with a minimum minor allele frequency (MAF) of 5%.

Next, we aimed to examine the linkage disequilibrium (LD) relationships of surrounding SNPs with the candidate variant that satisfied both the $p$-value and MAF criteria. To observe the LD relationships with a larger number of SNPs, we utilized variant data generated up to the GATK variant calling stage using RNA-seq data. The vcf file processed through the GATK pipeline was then converted to PLINK format using VCFtools and PLINK. We assessed LD relationships within a 250 Kb window on both sides of the significant SNP of interest and retrieved all reported pairs. The results were visualized using bar plots to illustrate the LD relationships between the candidate variant and other SNPs.

## Gene set enrichment analysis and visualization

DAVID v6.8 (https://david.ncifcrf.gov/) tool Field (Huang et al., 2007) was used for functional annotation and enrichment analysis of the DEG list. A $p$-value of 0.05 was used as the criterion for statistical significance. The gene set of genes reported in the GWAS catalog was identified using the GENE2FUNC process of the FUMA GWAS (https://fuma.ctglab.nl/) Field (Watanabe et al., 2017). For both tools, a $p$-value of 0.05 was used as the criterion for statistical significance.

# Results and discussion

## Experimental design and data quantification

We obtained the RNA-seq data for 22 tissue samples from 32 castrated Hanwoo cattle of different ages and weights (Materials and Methods) from the same farm managed by the Hanwoo Research Institute. Of the 32 samples, 3 samples (Sample ID: 192018, 192031, 202012) did not have sequencing data for some tissues, so the data were generated only with the tissues that had complete sequencing data (Supplementary Table S2). Statistical data regarding sequence quality and alignment information have been compiled and summarized for three adipose tissues in Supplementary Table S3. The 32 cattle were of different ages and had different body weights for each age (Supplementary Table S1). Our data show a high correlation between age and weight, and we confirmed that age explains a significant portion of the variation in body weight through the scatter plot ($R^2 = 0.9075$, $p$-value < 2.2E-16) (Supplementary Figure S1). Our goal was to detect differentially expressed genes (DEGs) that affect body weight and strictly control for other environmental factors. Body weight as a phenotype was adjusted considering age through simple regression analysis, and residuals were standardized as the z-score to control for the age factor (Supplementary Table S1).

In addition, we sought to control the genetic architecture and external influences on body weight among Hanwoo cattle. To achieve this, we evaluated the sample independence and potential

**TABLE 1 Differentially expressed genes (DEGs) Top 1 by tissue.**

| Tissue | Gene | Gene name | Log2FC | *p*-value | Function & association |
|---|---|---|---|---|---|
| ABF | *NFATC1* | nuclear factor of activated T cells 1 | 1.107661292 | 2.05E-04 | Estimated glomerular filtration rate in diabetes, Estimated glomerular filtration rate in diabetes, Estimated glomerular filtration rate in non-diabetics |
| ABO | *C1H3orf52* | chromosome 1 C3orf52 homolog | −1.22000226 | 2.26E-04 | - |
| BFT | *ZNF385A* | Zinc Finger Protein 385A | 1.030168135 | 2.37E-04 | Adipogenesis through 3′-UTR binding and translational regulation of CEBPA mRNA, Body height, BMI-adjusted waist circumference |
| BLO | *BTBD16* | BTB Domain Containing 16 | −3.192387309 | 1.42E-04 | Bipolar disorder disease |
| CEC | *SCGB1D* | secretoglobin, family 1D | 2.761929305 | 2.69E-04 | - |
| COL | *REG3G* | Regenerating Family Member 3 Gamma | 5.174855604 | 3.47E-04 | Regenerating islet-derived protein 3-gamma levels |
| DUO | *BTNL9* | Butyrophilin Like 9 | 1.219409464 | 3.32E-03 | Blood protein levels |
| HEA | *EFEMP1* | EGF Containing Fibulin Extracellular Matrix Protein 1 | −1.19589485 | 1.41E-04 | Body height, Body fat distribution, Body weight, Body mass index |
| ILE | *CDCA7* | Cell Division Cycle Associated 7 | −1.17600901 | 9.35E-06 | MYC-mediated cell transformation and apoptosis |
| JEJ | *IL21* | Interleukin 21 | −2.137673447 | 2.05E-04 | Cytokines with immunomodulatory activity |
| KID | *TOPAZ1* | Testis And Ovary Specific TOPAZ 1 | 1.771979303 | 3.66E-03 | Body mass index, Sperm development and sperm cell division |
| KIF | *CD300H* | CD300H Molecule (Gene/Pseudogene) | 3.114638566 | 2.35E-04 | Involved in innate immunity and autoimmune response |
| LIV | *CYP1A1* | Cytochrome P450 Family 1 Subfamily A Member 1 | −1.068010846 | 1.34E-04 | Involved in the metabolism of various endogenous substrates including fatty acids, steroid hormones and vitamins |
| LOM | *BARX1* | BARX Homeobox 1 | 1.52242117 | 2.13E-03 | Body mass index, Inhibits endoderm Wnt activity |
| LUN | *FOSB* | FosB Proto-Oncogene, AP-1 Transcription Factor Subunit | −1.71298643 | 2.74E-03 | Coexistence of osteoporosis, colon cancer and obesity |
| OMA | *ADAMTSL3* | ADAMTS Like 3 | 1.038905769 | 4.64E-05 | Body fat distribution, Body fat percentage, Abdominal adipose tissue volumes, Type 2 diabetes, Weight |
| REC | *UBD* | Ubiquitin D | −1.929973168 | 6.09E-4 | Inflammation, apoptosis and tumorigenesis, Adipogenesis and proliferation |
| RET | *SLC6A14* | Solute Carrier Family 6 Member 14 | −1.243818073 | 9.84E-04 | Mutations in this gene are associated with X-linked obesity |
| RMP | *CASQ2* | Calsequestrin 2 | −1.278916229 | 1.55E-03 | Serves as an internal calcium store in muscle |
| RUM | *SH2D1A* | SH2 Domain Containing 1A | 1.240464379 | 1.13E-04 | Inhibitors of transmembrane proteins |
| SPL | *OR5E1* | Olfactory receptor family 5 subfamily E member 1 | −2.866277868 | 5.28E-03 | - |
| TEN | *ALB* | Albumin | −3.212922518 | 1.05E-03 | Control of colloidal osmotic pressure in the blood |

correlations due to genetic or extrinsic factors. The analysis revealed that the first principal component (PC1) accounted for 7.0% of the variance in the data, while the second principal component (PC2) explained 5.6% (Supplementary Figure S2A). Although it seemed to form a cluster, it was confirmed that samples of similar weight or similar age were not grouped. We used a scatter plot to examine the causal or correlational relationship between PC1 and PC2 and the sample's age, weight, and z-scores (Supplementary Figure S2B) (Kahng et al., 1998). All scatter plots showed apparent linear relationships, but PC1 × months, PC1 × body weights, PC1 × z-score, PC2 × months, and PC2 × body weights demonstrated non-significant correlations with *p*-values greater than 0.05. The scatter plot of the PC2 × z-score was considered statistically significant with a *p*-value of 0.03, but it exhibited a weak correlation with an $R^2$ value of 0.16. These results suggest that the formed clusters are not strongly associated with age or weight, and there are no genetically related individuals. However, it also demonstrates the necessity to consider PC1 and PC2 to obtain robust evidence of genetic mechanisms underlying body weight regulation.

## Analysis of differentially expressed candidate genes involved in body weight in each tissue

To carefully detect differential gene expression in relation to weight, we defined the heavy (case) and light (control) groups

which correspond to the top and bottom 10 individuals according to the age-fitted body weight z-scores, respectively. The transcriptomic comparison between groups was performed for each tissue independently while adjusting for the covariates PC1 and PC2 (Supplementary Table S4). By comparing the control and case groups in each of the 22 tissues, we found a maximum of 602 DEGs in the ILE and a minimum of 17 DEGs in the Tenderloin (TEN) (Supplementary Table S5) (Zhou et al., 2019). The top 1 genes that were significantly identified in each tissue, *Zinc Finger Protein 385A (ZNF385A), EGF Containing Fibulin Extracellular Matrix Protein (EFEMP1), Testis And Ovary Specific TOPAZ 1 (TOPAZ1), Cytochrome P450 Family 1 Subfamily A Member 1 (CYP1A1), BARX Homeobox 1 (BARX1), FosB Proto-Oncogene (FOSB), ADAMTS like 3 (ADAMTSL3), Ubiquitin D (UBD), Solute carrier family 6 member (SLC6A14)* are known to be associated with adipogenesis, adipocyte proliferation, height, body mass index (BMI), lipid metabolism, weight and obesity-related functions (Table 1) (Suviolahti et al., 2003; Safran et al., 2010; DuBois et al., 2012; Welter et al., 2014; Skrypnik et al., 2017; Zhao et al., 2018; Rask-Andersen et al., 2019; Wang et al., 2020; Xiao et al., 2020). The genes identified that genes contributing to weight-related functions are not exclusively expressed in adipose tissue but also in other tissues such as the heart (HEA), kidney (KID), liver (LIV), sirloin (LOM), lung (LUN), omasum (OMA), rectum (REC), and reticulum (RET). This suggests a broader biological involvement of these genes across various tissue types in the regulation of body weight.

In addition, 381 genes were differentially expressed in two or more tissues (Supplementary Table S6). Notably, the *RGS2* and the *UBD* exhibited differential expression in multiple tissues. The *Regulator Of G Protein Signaling 2 (RGS2)* showed significant differential expression in 9 tissues (Abomasum (ABO), BLO, Colon (COL), ILE, Sirloin (LOM), Omasum (OMA), Rectum (REC), Spleen (SPL) and TEN), and the *UBD* was differentially expressed in 7 tissues (COL, Jejunum (JEJ), Liver (LIV), LOM, OMA, REC, Rumen (RUM)). It has been reported that the loss of *RGS2* is advantageous for glucose production but disadvantageous for glycogen and lipid production, contributing to a lean phenotype with a lower body weight (Nunn et al., 2011). This has previously been demonstrated to emphasize the importance of *RGS2* in obesity control and insulin sensitivity because it is involved in adipocyte differentiation, and impaired adipocyte differentiation can contribute to a lower body weight phenotype. The downregulation of the *UBD* has been reported in previous studies to partially inhibit adipogenesis of subcutaneous adipocyte precursor cells within pig muscles and to suppress cell proliferation, indicating its essential role in the differentiation of adipocyte precursor cells (Zhao et al., 2018). This suggests that in order to understand the mechanisms by which differentially expressed genes (DEGs) in non-adipose tissues influence body weight, additional analyses are also necessary.

## Correlation of gene expression patterns between tissues

To examine the overall relationship of gene expression patterns across the entire tissue, we generated a correlation plot using the T-values from the Limma-Voom analysis, representing the expression differences between the case and control groups (Figure 2). We used a hierarchical clustering algorithm (hclust) to generate correlation plots by dividing into three major clusters to view the clustering of tissues with similar gene expression patterns. The highest positive correlation coefficient observed (correlation coefficient = 0.58) was between the Abdominal Fat (ABF) and Back Fat (BFT), and the highest negative correlation coefficient (=−0.42) was between the OMA and Cecum (CEC). One of the three clusters was identified to consist exclusively of adipose tissue, including ABF, BFT and kidney fat (KIF). This result indicates a similarity in gene expression patterns among adipose tissues, and we specifically focused on adipose tissue as the primary tissue for further investigation.

## Functional annotation of adipose tissue DEGs

Figures 3A–C depict the differential expression patterns of all genes in three adipose tissues (ABF, BFT and KIF) and showcase the DEGs that were significantly selected in two or more adipose tissues. Figure 3D depicts the differential expression profiles and related enrichment analysis of significant DEGs in the three adipose tissues using the Gene Ontology (GO) (Consortium, 2004), conducted through the FUMA GWAS (Watanabe et al., 2017). The gene set enrichment analysis mapped 40 genes to the obesity-related traits ($p$-value < 0.05) among the GO terms (Figure 3D).

In the GO analysis of adipose tissue using DAVID (Huang et al., 2007), the genes were examined separately as upregulated genes and downregulated genes (Table 2). The downregulated genes, including those involved in signal transduction, organ or limb morphogenesis, bone cell differentiation or development, various metabolic biosynthetic processes, and other related biological processes (BP), were enriched with 41, 24, and 7 significant GO terms in the ABF, BFT, KIF tissues, respectively. The upregulated genes, primarily involved in metabolism, fatty acid metabolism, lipid metabolism, glutathione, and other metabolic processes, were enriched with 18, 3, and 2 significant GO terms in the ABF, BFT and KIF tissues, respectively. This indicates that most of these are associated with metabolic processes such as fatty acid and lipid metabolism and signal transduction.

## Expression comparison of fatty acid metabolism

We discovered through gene set enrichment analysis that many DEGs are associated with metabolic processes such as fatty acid and lipid metabolism in adipose tissues. We examined the expression levels of *Fatty acid synthase (FASN)* and *Stearoyl-CoA desaturase (SCD)*, DEGs involved in well-known fatty acid metabolism according to the literature, in all tissues (Figure 4) (Zhang et al., 2018). We confirmed that the *FASN* and *SCD* are expressed significantly more in the adipose tissues than in the other tissues. *FASN* is a complex homodimeric enzyme that regulates the *de novo* synthesis of long-chain fatty acids in mammals (Chakravarty et al., 2004). It catalyzes the formation of fatty acids with a 16-carbon atom length from acetyl-CoA and malonyl-CoA (Chakravarty et al., 2004). The expression product of the *FASN* is involved in lipid metabolism, and it is known to participate in fat accumulation and

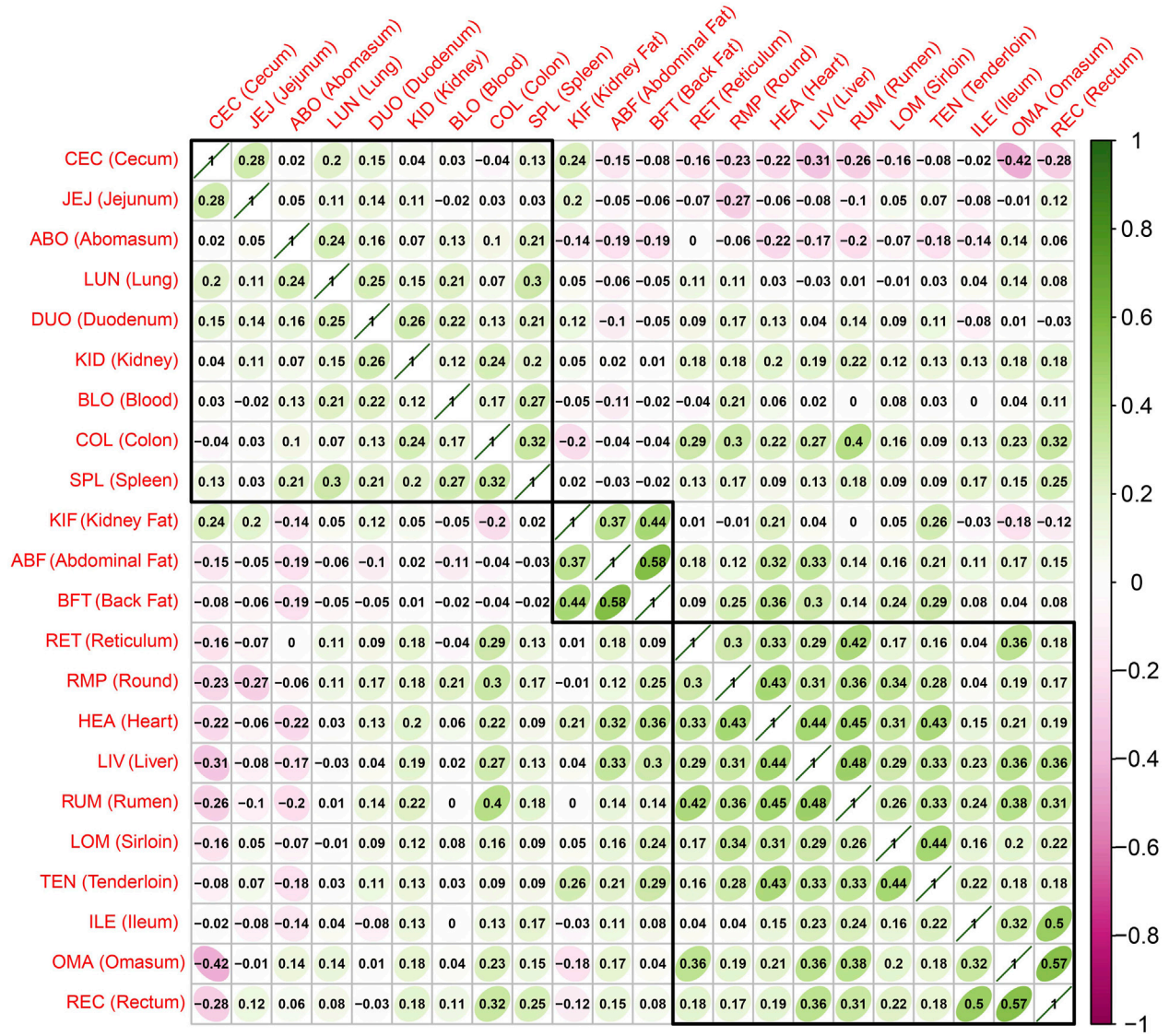



FIGURE 2
Correlations and hierarchical clustering between tissues based on the t value representing the difference in expression between the case group (sample with a high z-score) and the control group (sample with a low z-score). The closer the correlation value is to 1, the higher the correlation.

fatty acid composition in pigs and cattle (Grzes et al., 2016; Raza et al., 2018). The *SCD* functions as an enzyme in mammalian adipocytes, converting saturated fatty acids into monounsaturated fatty acids (MUFAs) (Taniguchi et al., 2004). The conversion of saturated fatty acids to MUFAs by the *SCD* enzyme has a role in signal transduction, cell differentiation, and cell apoptosis. It can influence the development of certain tumor mutations (Dobrzyn and Ntambi, 2004). Considering the various roles of these MUFAs, changes in *SCD* activity in mammals can potentially impact key physiological variables such as differentiation, insulin sensitivity, metabolic rate, obesity, atherosclerosis, and cancer (Dobrzyn and Ntambi, 2004). The *SCD* is an important metabolic control point in weight regulation. One study identified it as one of the genes exerting the greatest influence on intramuscular fat content and fatty acid composition in Angus cattle through a Genome-wide association study (GWAS) (Dobrzyn and Ntambi, 2004; Ros-Freixedes et al., 2016).

## The candidate *cis*-eQTL variant regulating the expression level of *TRIM31*

In the adipose tissues of ABF, BFT, and KIF, expression quantitative trait loci (eQTL) analysis was conducted on 697 genes identified as differentially expressed, which yielded three SNPs meeting the significance threshold of $p$-value < 5e-05 and possessing a minor allele frequency (MAF) greater than 5% (Supplementary Table S7). The three SNPs that met both the $p$-value and MAF criteria were exclusively discovered in the ABF tissue; no SNPs in the BFT and KIF tissues satisfied these conditions. One SNP was located within the *UBD* and emerged as a significant variant influencing the expression levels of *Tripartite Motif Containing 31* (*TRIM31*) (chr23:29119138, $p$-value = 8.86e-06, MAF = 19.4%). The other two SNPs were positioned near the *Adenosine Deaminase RNA Specific* (*ADAR*) and identified as significant variants affecting

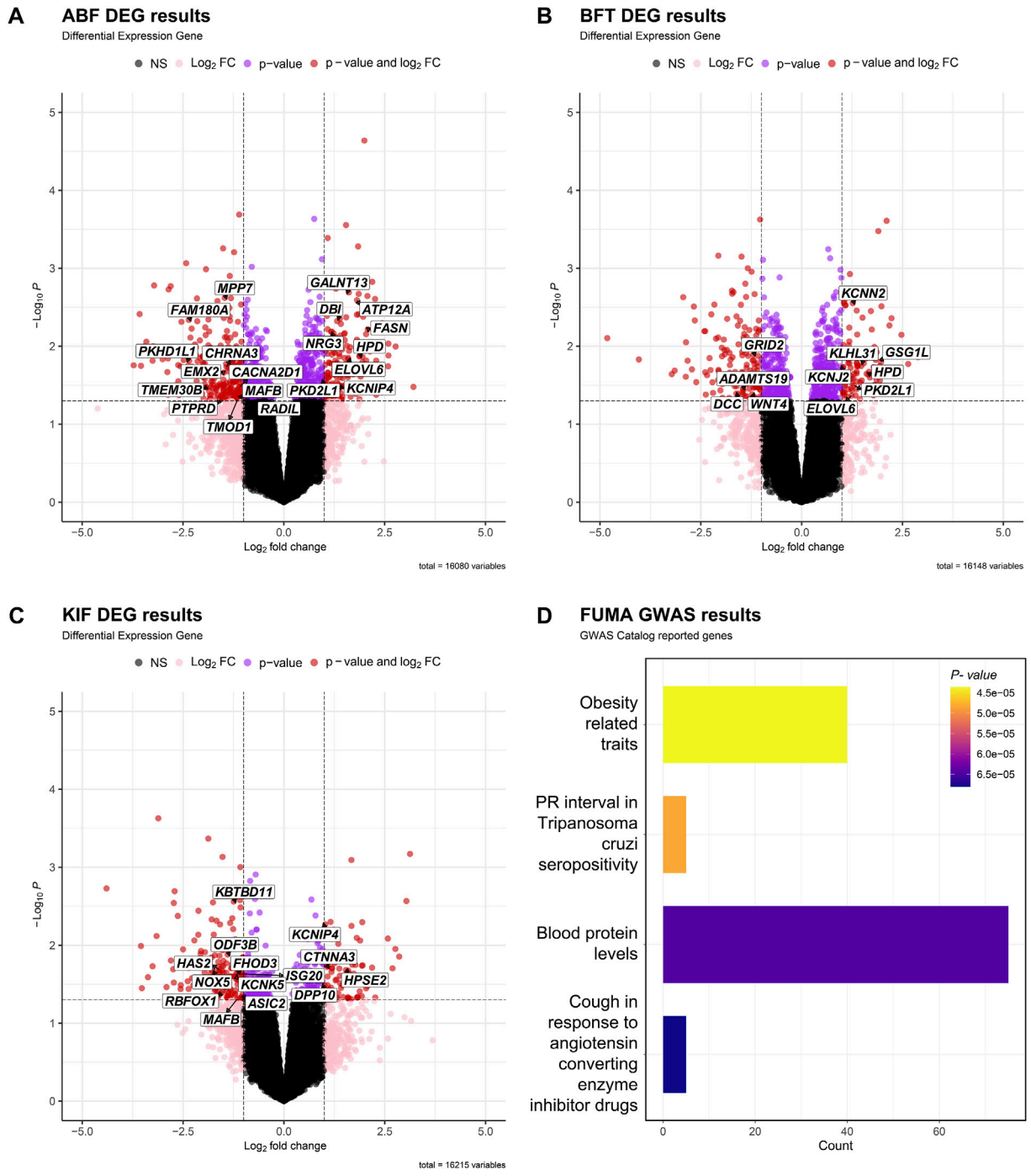



FIGURE 3
Volcano plot of DEGs in adipose tissue [**(A)** ABF, **(B)** BFT, **(C)** KIF]. The X-axis is $Log_2FC$, the Y-axis is $-Log_{10}p$- value, and the cutoff criterion is $p$-value < 0.05, $|Log_2FC| > 1$. Black dots are points that do not satisfy both the $p$-value and the $Log_2FC$ FC criterion, pink dots are points that satisfy only the $Log_2FC$ criterion, purple dots are points that satisfy only the $p$-value criterion, and red dots are differentially expressed genes (DEGs) that satisfy both $p$-value and $Log_2FC$ criteria. We displayed the genes associated with obesity-related traits based on the results of **(D)**. **(D)** Table showing the gene set reported in the GWAS Catalog based on the results of FUMA GWAS using candidate genes from the three tissues.

the expression of the *Tudor Domain Containing 10* (*TDRD10*) (chr3: 15995172, $p$-value = 1.31e-05, MAF = 5%; chr3:15995183, $p$-value = 1.31e-05, MAF = 5%). *TRIM31* has been identified as a "Janus-faced" regulator of innate immune responses, facilitating signal transduction through target substrate degradation or ubiquitin modification (Xu et al., 2022). Furthermore, in some studies, functional impairment (knockout) of *TRIM31* has been shown to significantly increase body weight, fasting blood glucose levels, and fasting insulin levels induced by a high-fat diet (HFD), suggesting that reduced expression of *TRIM31* can contribute to weight gain (Xu et al., 2022). Due to the lack of literature supporting the involvement of *TDRD10* in weight regulation, the SNP (chr23:29119138) within

10.3389/fgene.2023.1304638

**TABLE 2 Significant gene set enrichment analysis in 3 adipose tissue (ABF, BFT, KIF) DEGs.**

| Tissue | Regulation | Pathways | Term | *p*-value |
|---|---|---|---|---|
| ABF | Downregulated | GO-Biological Pathways | Positive regulation of mesenchymal cell proliferation | 3.07E-05 |
| | | | Mesenchyme migration | 1.19E-03 |
| | | | Animal organ morphogenesis | 1.62E-03 |
| | | | Proximal/distal pattern formation | 2.00E-03 |
| | | | Positive regulation of gene expression | 2.29E-03 |
| | | | Embryonic limb morphogenesis | 2.86E-03 |
| | | | Positive regulation of canonical Wnt signaling pathway | 3.54E-03 |
| | | | Embryonic forelimb morphogenesis | 3.93E-03 |
| | | | Positive regulation of smoothened signaling pathway | 4.33E-03 |
| | | | Extracellular matrix organization | 5.15E-03 |
| | | | Osteoblast differentiation | 5.18E-03 |
| | | | Lymphangiogenesis | 6.26E-03 |
| | | | Regulation of heart contraction | 6.26E-03 |
| | | | Cartilage development | 7.25E-03 |
| | | | Chondrocyte differentiation | 8.43E-03 |
| | | | Embryonic digestive tract development | 8.75E-03 |
| | | | Positive regulation of endothelial cell migration | 1.04E-02 |
| | | | Inner ear morphogenesis | 1.11E-02 |
| | | | Negative regulation of osteoblast differentiation | 1.19E-02 |
| | | | Collagen fibril organization | 1.26E-02 |
| | | | Positive regulation of protein kinase B signaling | 1.29E-02 |
| | | | Intermediate filament cytoskeleton organization | 1.48E-02 |
| | | | Intermediate filament organization | 1.84E-02 |
| | | | Oogenesis | 1.84E-02 |
| | | | Positive regulation of cell proliferation | 2.52E-02 |
| | | | Wnt signaling pathway | 2.57E-02 |
| | | | Positive regulation of MAPK cascade | 2.74E-02 |
| | | | Lung development | 2.87E-02 |
| | | | Positive regulation of vascular endothelial growth factor production | 3.08E-02 |
| | | | Ureter maturation | 3.28E-02 |
| | | | Vascular smooth muscle contraction | 3.28E-02 |
| | | | Signal transduction | 3.29E-02 |
| | | | Positive regulation of axonogenesis | 3.32E-02 |
| | | | Eye development | 4.32E-02 |
| | | | Zymogen activation | 4.32E-02 |
| | | | Regulation of Ras protein signal transduction | 4.35E-02 |
| | | | Negative regulation of inflammatory response | 4.55E-02 |
| | | | Inner ear development | 4.58E-02 |

(Continued on following page)

**TABLE 2 (*Continued*) Significant gene set enrichment analysis in 3 adipose tissue (ABF, BFT, KIF) DEGs.**

| Tissue | Regulation | Pathways | Term | *p*-value |
|---|---|---|---|---|
| | | | Negative regulation of Wnt signaling pathway | 4.58E-02 |
| | | | Brain development | 4.76E-02 |
| | | | Regulation of heart rate by cardiac conduction | 4.85E-02 |
| | Upregulated | GO-Biological Pathways | Glutathione metabolic process | 1.80E-05 |
| | | | ATP synthesis coupled electron transport | 6.75E-04 |
| | | | Mitochondrial electron transport, NADH to ubiquinone | 8.85E-04 |
| | | | Fatty acid biosynthetic process | 2.01E-03 |
| | | | Phospholipid biosynthetic process | 3.40E-03 |
| | | | Unsaturated fatty acid biosynthetic process | 4.54E-03 |
| | | | Mitochondrial respiratory chain complex I assembly | 8.27E-03 |
| | | | Response to stilbenoid | 1.35E-02 |
| | | | Regulation of phospholipid biosynthetic process | 1.35E-02 |
| | | | Response to glucose | 1.87E-02 |
| | | | Tryptophan transport | 2.03E-02 |
| | | | Proteolysis | 3.04E-02 |
| | | | Regulation of cytokine production | 3.16E-02 |
| | | | Protein homotetramerization | 3.90E-02 |
| | | | Negative regulation by host of viral process | 4.01E-02 |
| | | | Glycerol-3-phosphate metabolic process | 4.01E-02 |
| | | | Fatty acid elongation, polyunsaturated fatty acid | 4.66E-02 |
| | | | Lung lobe morphogenesis | 4.66E-02 |
| BFT | Downregulated | GO-Biological Pathways | Protein urmylation | 2.25E-02 |
| | | | tRNA wobble position uridine thiolation | 2.80E-02 |
| | | | Lung vasculature development | 2.80E-02 |
| | | | Axon guidance | 4.23E-02 |
| | | | Extracellular matrix organization | 4.78E-02 |
| | | | Tyrosine phosphorylation of STAT protein | 4.99E-02 |
| | | | Digestive tract morphogenesis | 4.99E-02 |
| | Upregulated | | Glutathione metabolic process | 3.00E-05 |
| | | | CDP-diacylglycerol biosynthetic process | 2.94E-02 |
| | | | Axonemal dynein complex assembly | 4.73E-02 |
| KIF | Downregulated | GO-Biological Pathways | Superoxide anion generation | 1.93E-06 |
| | | | Extracellular matrix assembly | 1.59E-03 |
| | | | Thyroid hormone generation | 2.24E-03 |
| | | | Stabilization of membrane potential | 3.85E-03 |
| | | | Negative regulation of osteoclast differentiation | 4.81E-03 |
| | | | Positive regulation of apoptotic cell clearance | 1.10E-02 |
| | | | Regulation of thyroid hormone generation | 1.10E-02 |

(Continued on following page)

**TABLE 2 (*Continued*) Significant gene set enrichment analysis in 3 adipose tissue (ABF, BFT, KIF) DEGs.**

| Tissue | Regulation | Pathways | Term | *p*-value |
|---|---|---|---|---|
| | | | Defense response to Gram-positive bacterium | 1.38E-02 |
| | | | Proteolysis | 1.52E-02 |
| | | | Melatonin biosynthetic process | 1.64E-02 |
| | | | Extracellular polysaccharide biosynthetic process | 1.64E-02 |
| | | | Positive regulation of hydrogen peroxide biosynthetic process | 1.64E-02 |
| | | | Negative regulation of BMP signaling pathway | 2.13E-02 |
| | | | Cell adhesion | 2.21E-02 |
| | | | Hyaluronan biosynthetic process | 2.72E-02 |
| | | | Hydrogen peroxide metabolic process | 2.72E-02 |
| | | | Ossification | 2.86E-02 |
| | | | Hydrogen peroxide biosynthetic process | 3.25E-02 |
| | | | Positive regulation of platelet aggregation | 3.25E-02 |
| | | | Response to light stimulus | 3.78E-02 |
| | | | Bone trabecula formation | 3.78E-02 |
| | | | Respiratory burst | 4.31E-02 |
| | | | Positive regulation of phosphatidylinositol 3-kinase signaling | 4.70E-02 |
| | | | Embryonic eye morphogenesis | 4.84E-02 |
| | Upregulated | GO-Biological Pathways | Regulation of potassium ion transmembrane transport | 1.40E-04 |
| | | | Regulation of membrane potential | 1.29E-02 |

*UBD* that regulates the expression of *TRIM31* was selected as the most prominent candidate *cis*-eQTL variant associated with weight control.

We sought to determine whether the genotype of the candidate cis-eQTL variant (chr23:29119138) is associated with the regulation of *TRIM31* expression and body weight control. Samples with the GG genotype at chr23:29119138 showed lower expression levels of *TRIM31* compared to samples carrying the alternative allele A ($p$-value = 8.86e-06) (Figure 5A). Upon comparing *TRIM31* expression levels between the heavy and light groups, the light group exhibited a significantly higher level of *TRIM31* expression than the heavy group (Figure 5B). Although the $p$-value did not reach statistical significance, visualization of the raw body weight differences based on the genotypes at chr23:29119138 showed a trend where individuals carrying allele A exhibited a lower weight distribution (Supplementary Figure S3). These research findings, while not statistically significant, support the directionality of weight regulation associated with *TRIM31* expression previously reported in the literature (Luo et al., 2022; Xu et al., 2022). This suggests that a reduction in *TRIM31* expression may contribute to an increase in body weight, and the variant at chr23:29119138 could potentially regulate the expression of *TRIM31*.

However, the candidate *cis*-eQTL variant is located within the exon of the *UBD* gene, which has been reported in mice to be a gene upregulated by an HFD, and the deficiency of *UBD* has been associated with a reduction in body fat due to increased energy expenditure (Choi et al., 2015). Given that the *TRIM31* gene (but not the *UBD*) showed the differential expression between weight groups, the significant association of the SNP within the *UBD* gene and the expression level of *TRIM31* may have been attributable to strong linkage disequilibrium (LD) within this region. To examine this further, an LD pattern analysis was conducted within a 250 Kb range on either side of the candidate variant to assess the LD relationship with neighboring SNPs. The LD analysis revealed a high level of linkage around the candidate *cis*-eQTL variant and also confirmed a strong LD ($R^2$ = 1) with SNPs within the *TRIM31* (Figure 5C; Supplementary Figure S4). These findings suggest that the neighboring SNPs may be co-inherited with the candidate *cis*-eQTL variant and could be implicated in gene expression regulation, even if they are not the direct causative variants. Moreover, the candidate variant could be correlated with the causal variant due to close genetic linkage. Consequently, the candidate *cis*-eQTL variant may be closely linked to the actual causative variant or contribute to the modulation of *TRIM31* expression. This underscores the need for further investigation to elucidate the fundamental genetic mechanisms.

## Limitation

The current study's outcomes are subject to several limitations. Weight gain over time serves as an important indicator of feed efficiency and would likely yield better performance in research outcomes. Additionally, the absence of information on environmental factors presents a limitation in completely controlling for external influences. Nevertheless, it is acknowledged that body weight is significantly influenced by both environmental

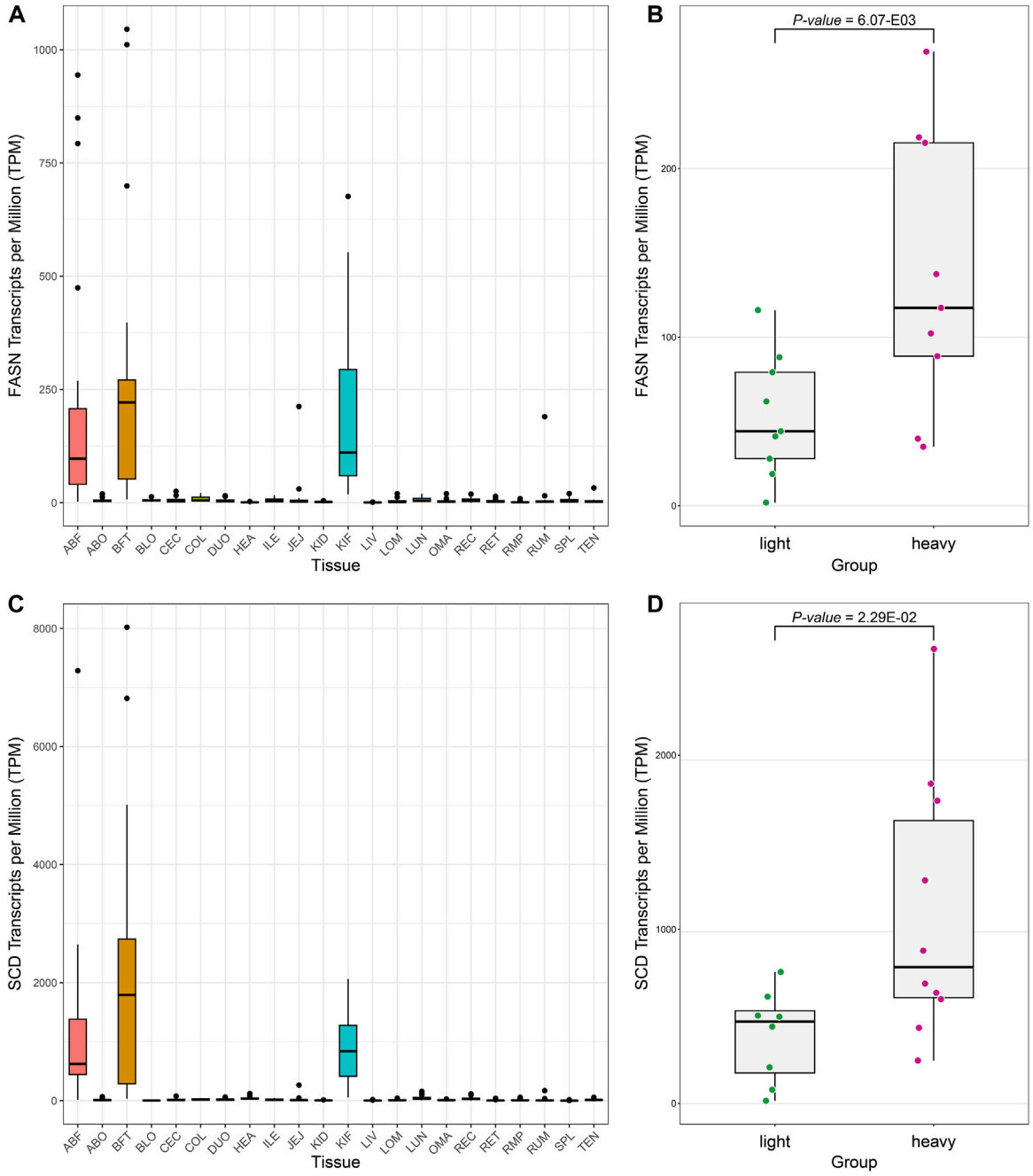



FIGURE 4
Boxplot to confirm the expression level of representative genes (*FASN, SCD*) related to fat metabolism. **(A)** The expression pattern of the *FASN* across all tissues. **(B)** Boxplot showing the differential expression pattern of *FASN* in ABF tissue based on groups (z-scores of top and bottom 10 samples). **(C)** Expression pattern of the *SCD* across all tissues. **(D)** Boxplot showing the differential expression pattern of the *SCD* gene in ABF tissue based on groups.

and genetic factors. Particularly in humans, the heritability estimate for body weight has been reported to be between 0.7 and 0.81, while in cattle, heritability estimates vary but have been reported to range from approximately 0.3 to 0.6 (Toshniwal et al., 2008; Russo et al., 2010; Polizel et al., 2018; Rezende et al., 2022). Recognizing the weight of genetic contributions, this study meticulously incorporated principal components (PC1 and PC2) as covariates in the DEG analysis to mitigate environmental biases, thus sharpening the focus on genetic correlations with body weight.

The sample size, comprising 32 individuals, is relatively small, which restricts the ability to detect trans-eQTLs that typically exhibit smaller effect sizes compared to cis-eQTLs. In both DEG and eQTL analyses, the *p*-value significance threshold did not meet the stringent standards set by FDR adjustment, posing challenges in identifying influential genes and SNPs. This raises concerns regarding the incidence of Type I errors. The imperative for subsequent analyses with augmented datasets is clear, to yield more precise and dependable outcomes.

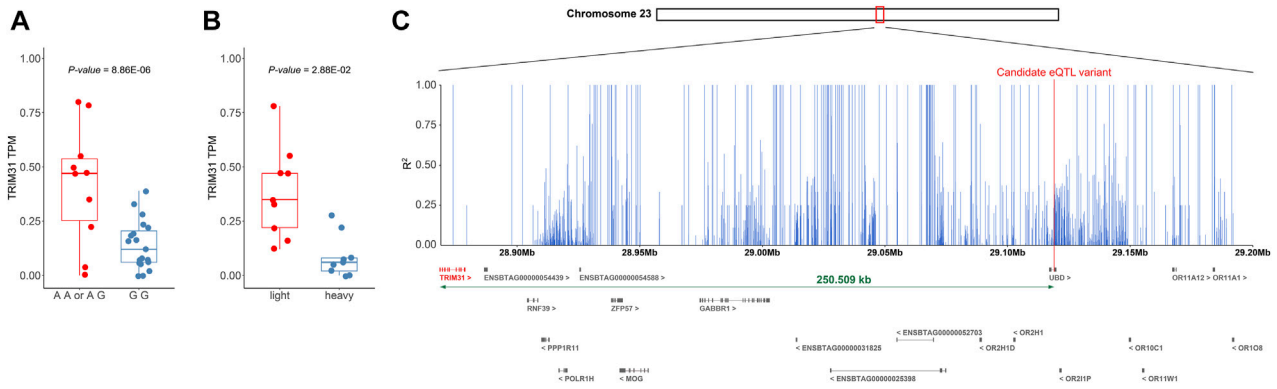



FIGURE 5
**(A)** Distance between *TRIM31* and candidate variants regulating the expression level. **(B)** Boxplot showing the change in *TRIM31* expression level according to the genotype of the candidate *cis*-eQTL variant (chr23:29119138) that affects *TRIM31* expression. **(A, B)** Additionally, outliers were excluded during the visualization process in creating the boxplot. **(C)** Bar plot depicting the LD values between the candidate *cis*-eQTL variant and surrounding SNPs within 250 Kb. The figure displays the range chr 23:28,868,629-29,200,000. A plot of the entire 500 Kb region is provided in Supplementary Figure S4. The x-axis represents genetic coordinates while the y-axis indicates the LD values between the candidate variant and SNPs. Differentially expressed gene *TRIM31* and the candidate *cis*-eQTL variant are highlighted in red.

Despite these issues, the study leverages the convergence of DEG and eQTL analyses to enhance the reliability of the genetic associations identified. The utilization of RNA-seq data for eQTL analysis is a novel approach for Hanwoo cattle research, marking a significant contribution that paves the way for future inquiry. This research underscores the importance of continuous investigation, bolstered by broader datasets, to reinforce the preliminary findings presented.

## Conclusion

We analyzed the gene expression data from multiple tissues to identify genes and biological mechanisms at the transcriptome level that influence the weight of Hanwoo cattle. Our study has uncovered transcriptional changes associated with weight in previously overlooked tissues. We have confirmed that the candidate genes we discovered are associated with biological pathways involving various metabolic processes, such as lipid metabolism, adipogenesis, and adipocyte proliferation. Using RNA-seq data in expression quantitative trait loci (eQTL) studies enabled us to identify allele-specific gene expression easily. By integrating eQTL and differentially expressed genes (DEGs) analysis results, we have identified genomic regions that may regulate the expression of candidate genes, such as *TRIM31* and provided insights into their association with the expression levels. Of particular interest, we found that the variant regulating the expression of *TRIM31* is located within the *UBD*, which is known to regulate adipogenesis. The findings suggest that further analysis is necessary to fine-map causal *cis*-eQTL variants regulating *TRIM31*. Moreover, it emphasizes the necessity to broaden the focus and understanding of research on various tissues that can influence the weight of Hanwoo cattle and other livestock. Our study may represent a comprehensive genomic and transcriptomic portrait of livestock body weight by utilizing the RNA-seq data of many tissues and progress toward understanding the role of eQTLs in determining livestock phenotypic diversity.

## 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 approved by National Institute of Animal Science. The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

SbJ: Conceptualization, Writing–original draft, Writing–review and editing, Investigation. SnJ: Writing–review and editing, Funding acquisition, Resources. JK: Writing–review and editing, Conceptualization, Supervision, Writing–original draft. WP: Conceptualization, Funding acquisition, Investigation, Resources, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was carried out with the support of 2020 the RDA Fellowship Program of the NIAS and Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ01492004, “Determination of growing performance, tissue physiology and meat quality change in Hanwoo steers”), Rural Development Administration, Republic of Korea.

## 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.2023.1304638/full#supplementary-material

## References


Adibhatla, R. M., and Hatcher, J. (2008). Altered lipid metabolism in brain injury and disorders. *Lipids health Dis.* 49, 241–268. doi:10.1007/978-1-4020-8831-5_9

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

Brown, J., Pirrung, M., and Mccue, L. A. (2017). FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* 33, 3137–3139. doi:10.1093/bioinformatics/btx373

Cai, W., Zhang, Y., Chang, T., Wang, Z., Zhu, B., Chen, Y., et al. (2023). The eQTL colocalization and transcriptome-wide association study identify potentially causal genes responsible for economic traits in Simmental beef cattle. *J. Animal Sci. Biotechnol.* 14, 78. doi:10.1186/s40104-023-00876-7

Chakravarty, B., Gu, Z., Chirala, S. S., Wakil, S. J., and Quiocho, F. A. (2004). Human fatty acid synthase: structure and substrate selectivity of the thioesterase domain. *Proc. Natl. Acad. Sci.* 101, 15567–15572. doi:10.1073/pnas.0406901101

Choi, J.-Y., Shin, D., Lee, H.-J., and Oh, J.-D. (2019). Comparison of long noncoding RNA between muscles and adipose tissues in Hanwoo beef cattle. *Animal Cells Syst.* 23, 50–58. doi:10.1080/19768354.2018.1512522

Choi, M.-S., Kim, Y.-J., Kwon, E.-Y., Ryoo, J. Y., Kim, S. R., and Jung, U. J. (2015). High-fat diet decreases energy expenditure and expression of genes controlling lipid metabolism, mitochondrial function and skeletal system development in the adipose tissue, along with increased expression of extracellular matrix remodelling-and inflammation-related genes. *Br. J. Nutr.* 113, 867–877. doi:10.1017/S0007114515000100

Consortium, G. O., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids Res.* 32, D258–D261. doi:10.1093/nar/gkh036

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

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi:10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Dobrzyn, A., and Ntambi, J. M. (2004). The role of stearoyl-CoA desaturase in body weight regulation. *Trends Cardiovasc. Med.* 14, 77–81. doi:10.1016/j.tcm.2003.12.005

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

Dubois, B. N., O'tierney-Ginn, P., Pearson, J., Friedman, J. E., Thornburg, K., and Cherala, G. (2012). Maternal obesity alters feto-placental cytochrome P4501A1 activity. *Placenta* 33, 1045–1051. doi:10.1016/j.placenta.2012.09.008

Fink, T., Tiplady, K., Lopdell, T., Johnson, T., Snell, R. G., Spelman, R. J., et al. (2017). Functional confirmation of PLAG1 as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. *Sci. Rep.* 7, 44793. doi:10.1038/srep44793

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., and Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: a critical perspective. *Trends Food Sci. Technol.* 72, 83–90. doi:10.1016/j.tifs.2017.12.006

Grzes, M., Sadkowski, S., Rzewuska, K., Szydlowski, M., and Switonski, M. (2016). Pig fatness in relation to FASN and INSIG2 genes polymorphism and their transcript level. *Mol. Biol. Rep.* 43, 381–389. doi:10.1007/s11033-016-3969-z

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, W169–W175. doi:10.1093/nar/gkm415

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. statistics* 5, 299–314. doi:10.1080/10618600.1996.10474713

Kahng, S., Iwata, B. A., Fischer, S. M., Page, T. J., Treadwell, K. R. H., Williams, D. E., et al. (1998). Temporal Distributions Of Problem Behavior Based On Scatter Plot Analysis. *J. Appl. Behav. Analysis* 31, 593–604. doi:10.1901/jaba.1998.31-593

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29–R17. doi:10.1186/gb-2014-15-2-r29

Lee, S.-H., Park, B.-H., Sharma, A., Dang, C.-G., Lee, S.-S., Choi, T.-J., et al. (2014). Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. *J. Animal Sci. Technol.* 56, 2. doi:10.1186/2055-0391-56-2

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* 12, 323. doi:10.1186/1471-2105-12-323

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Lindholm-Perry, A. K., Freetly, H. C., Oliver, W. T., Rempel, L. A., and Keel, B. N. (2020). Genes associated with body weight gain and feed intake identified by meta-analysis of the mesenteric fat from crossbred beef steers. *PLoS One* 15, e0227154. doi:10.1371/journal.pone.0227154

Littlejohn, M., Grala, T., Sanders, K., Walker, C., Waghorn, G., Macdonald, K., et al. (2012). Genetic variation in PLAG1 associates with early life body weight and peripubertal weight and growth in *Bos taurus*. *Anim. Genet.* 43, 591–594. doi:10.1111/j.1365-2052.2011.02293.x

Liu, G., Liu, X., Shahzad, K., You, W., Loor, J. J., and Wan, F. (2018). Bioinformatics analyses of bovine adipose tissue transcriptome from lilu beef cattle at different stages of growth. *Pak. J. Zool.* 50, 1847–1855. doi:10.17582/journal.pjz/2018.50.5.1847.1855

Luo, J., Tan, J., Zhao, J., Wang, L., Liu, J., Dai, X., et al. (2022). Cynapanoside A exerts protective effects against obesity-induced diabetic nephropathy through ameliorating TRIM31-mediated inflammation, lipid synthesis and fibrosis. *Int. Immunopharmacol.* 113, 109395. doi:10.1016/j.intimp.2022.109395

Muradian, K., Vaiserman, A., Min, K. J., and Fraifeld, V. E. (2015). Fucoxanthin and lipid metabolism: a minireview. *Nutr. Metabolism Cardiovasc. Dis.* 25, 891–897. doi:10.1016/j.numecd.2015.05.010

Naserkheil, M., Bahrami, A., Lee, D., and Mehrban, H. (2020). Integrating single-step GWAS and bipartite networks reconstruction provides novel insights into yearling weight and carcass traits in Hanwoo beef cattle. *Animals* 10, 1836. doi:10.3390/ani10101836

Nunn, C., Zhao, P., Zou, M.-X., Summers, K., Guglielmo, C. G., and Chidiac, P. (2011). Resistance to age-related, normal body weight gain in RGS2 deficient mice. *Cell. Signal.* 23, 1375–1386. doi:10.1016/j.cellsig.2011.03.020

Peng, S., Deyssenroth, M. A., Di Narzo, A. F., Cheng, H., Zhang, Z., Lambertini, L., et al. (2018). Genetic regulation of the placental transcriptome underlies birth weight and risk of childhood obesity. *PLoS Genet.* 14, e1007799. doi:10.1371/journal.pgen.1007799

Polizel, G. H. G., Grigoletto, L., Carvalho, M. E., Rossi Junior, P., Ferraz, J. B. S., and Santana, M. H. D. A. (2018). Genetic correlations and heritability estimates for dry matter intake, weight gain and feed efficiency of Nellore cattle in feedlot. *Livest. Sci.* 214, 209–210. doi:10.1016/j.livsci.2018.06.013

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., 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

Rask-Andersen, M., Karlsson, T., Ek, W. E., and Johansson, Ç. (2019). Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nat. Commun.* 10, 339. doi:10.1038/s41467-018-08000-4

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

Rezende, M. P. G., Malhado, C. H. M., Biffani, S., Carrillo-Tabakman, J. A., Fabbri, M. C., Crovetti, A., et al. (2022). Heritability and genetic correlation of body weight and Kleiber ratio in Limousin and Charolais beef cattle breeds. *animal* 16, 100528. doi:10.1016/j.animal.2022.100528

Ros-Freixedes, R., Gol, S., Pena, R. N., Tor, M., Ibáñez-Escriche, N., Dekkers, J. C., et al. (2016). Genome-wide association study singles out SCD and LEPR as the two main loci influencing intramuscular fat content and fatty acid composition in Duroc pigs. *PLoS One* 11, e0152496. doi:10.1371/journal.pone.0152496

Russo, P., Lauria, F., and Siani, A. (2010). Heritability of body weight: moving beyond genetics. *Nutr. Metabolism Cardiovasc. Dis.* 20, 691–697. doi:10.1016/j.numecd.2010.09.007

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database* 2010, baq020. doi:10.1093/database/baq020

Skrypnik, K., Suliburska, J., Skrypnik, D., Pilarski, Ł., Reguła, J., and Bogdański, P. (2017). The genetic basis of obesity complications. *Acta Sci. pol. Technol. aliment.* 16, 83–91. doi:10.17306/J.AFS.2017.0442

Suviolahti, E., Oksanen, L. J., Øhman, M., Cantor, R. M., Ridderstråle, M., Tuomi, T., et al. (2003). The SLC6A14 gene shows evidence of association with obesity. *J. Clin. Investigation* 112, 1762–1772. doi:10.1172/JCI17491

Taniguchi, M., Utsugi, T., Oyama, K., Mannen, H., Kobayashi, M., Tanabe, Y., et al. (2004). Genotype of stearoyl-CoA desaturase is associated with fatty acid composition in Japanese Black cattle. *Mamm. Genome* 15, 142–148. doi:10.1007/s00335-003-2286-8

Toshniwal, J. K., Dechow, C. D., Cassell, B. G., Appuhamy, J. a.D. R. N., and Varga, G. A. (2008). Heritability of electronically recorded daily body weight and correlations with yield, dry matter intake, and body condition score. *J. Dairy Sci.* 91, 3201–3210. doi:10.3168/jds.2007-0627

Wang, J., Gaughan, S., Lamer, J. T., Deng, C., Hu, W., Wachholtz, M., et al. (2020). Resolving the genetic paradox of invasions: preadapted genomes and postintroduction hybridization of bigheaded carps in the Mississippi River Basin. *Evol. Appl.* 13, 263–277. doi:10.1111/eva.12863

Wangchuk, K., Wangdi, J., and Mindu, M. (2018). Comparison and reliability of techniques to estimate live cattle body weight. *J. Appl. Animal Res.* 46, 349–352. doi:10.1080/09712119.2017.1302876

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1826. doi:10.1038/s41467-017-01261-5

Welter, D., Macarthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids Res.* 42, D1001–D1006. doi:10.1093/nar/gkt1229

Xiao, Y., Liu, D., Cline, M. A., and Gilbert, E. R. (2020). Chronic stress, epigenetics, and adipose tissue metabolism in the obese state. *Nutr. Metabolism* 17, 88. doi:10.1186/s12986-020-00513-4

Xu, M., Tan, J., Dong, W., Zou, B., Teng, X., Zhu, L., et al. (2022). The E3 ubiquitin-protein ligase Trim31 alleviates non-alcoholic fatty liver disease by targeting Rhbdf2 in mouse hepatocytes. *Nat. Commun.* 13, 1052. doi:10.1038/s41467-022-28641-w

Yang, H., Li, F., Xiong, X., Kong, X., Zhang, B., Yuan, X., et al. (2013). Soy isoflavones modulate adipokines and myokines to regulate lipid metabolism in adipose tissue, skeletal muscle and liver of male Huanjiang mini-pigs. *Mol. Cell. Endocrinol.* 365, 44–51. doi:10.1016/j.mce.2012.09.002

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

Zhang, X., Hu, Y., Ansari, A. R., Akhtar, M., Chen, Y., Cheng, R., et al. (2022). Caecal microbiota could effectively increase chicken growth performance by regulating fat metabolism. *Microb. Biotechnol.* 15, 844–861. doi:10.1111/1751-7915.13841

Zhang, Y., Liu, Z., Liu, R., Wang, J., Zheng, M., Li, Q., et al. (2018). Alteration of hepatic gene expression along with the inherited phenotype of acquired fatty liver in chicken. *Genes* 9, 199. doi:10.3390/genes9040199

Zhao, C., Yao, X., Chen, X., Wu, W., Xi, F., Yang, G., et al. (2018). Knockdown of ubiquitin D inhibits adipogenesis during the differentiation of porcine intramuscular and subcutaneous preadipocytes. *Cell Prolif.* 51, e12401. doi:10.1111/cpr.12401

Zhou, J., Zhang, C., Wu, X., Xie, Q., Li, L., Chen, Y., et al. (2019). Identification of genes and pathways related to atherosclerosis comorbidity and depressive behavior via RNA-seq and bioinformation analysis in ApoE(-/-) mice. *Ann. Transl. Med.* 7, 733. doi:10.21037/atm.2019.11.118

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

frontiers | Frontiers in Veterinary Science

TYPE Review
PUBLISHED 11 January 2024
DOI 10.3389/fvets.2023.1334434



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Yongzhen Huang,
Northwest A&F University, China
Sharmila Ghosh,
University of California, Davis, United States

*CORRESPONDENCE
Muhammad Zahoor Khan
✉ zahoorkhattak91@yahoo.com
Changfa Wang
✉ wangchangfa@lcu.edu.cn

RECEIVED 07 November 2023
ACCEPTED 27 December 2023
PUBLISHED 11 January 2024

CITATION
Liu X, Chen W, Huang B, Wang X, Peng Y, Zhang X, Chai W, Khan MZ and Wang C (2024) Advancements in copy number variation screening in herbivorous livestock genomes and their association with phenotypic traits. *Front. Vet. Sci.* 10:1334434. doi: 10.3389/fvets.2023.1334434



# Advancements in copy number variation screening in herbivorous livestock genomes and their association with phenotypic traits

Xiaotong Liu, Wenting Chen, Bingjian Huang, Xinrui Wang, Yongdong Peng, Xinhao Zhang, Wenqiong Chai, Muhammad Zahoor Khan* and Changfa Wang*

Liaocheng Research Institute of Donkey High-Efficiency Breeding, Liaocheng University, Liaocheng, China

Copy number variations (CNVs) have garnered increasing attention within the realm of genetics due to their prevalence in human, animal, and plant genomes. These structural genetic variations have demonstrated associations with a broad spectrum of phenotypic diversity, economic traits, environmental adaptations, epidemics, and other essential aspects of both plants and animals. Furthermore, CNVs exhibit extensive sequence variability and encompass a wide array of genomes. The advancement and maturity of microarray and sequencing technologies have catalyzed a surge in research endeavors pertaining to CNVs. This is particularly prominent in the context of livestock breeding, where molecular markers have gained prominence as a valuable tool in comparison to traditional breeding methods. In light of these developments, a contemporary and comprehensive review of existing studies on CNVs becomes imperative. This review serves the purpose of providing a brief elucidation of the fundamental concepts underlying CNVs, their mutational mechanisms, and the diverse array of detection methods employed to identify these structural variations within genomes. Furthermore, it seeks to systematically analyze the recent advancements and findings within the field of CNV research, specifically within the genomes of herbivorous livestock species, including cattle, sheep, horses, and donkeys. The review also highlighted the role of CNVs in shaping various phenotypic traits including growth traits, reproductive traits, pigmentation and disease resistance etc., in herbivorous livestock. The main goal of this review is to furnish readers with an up-to-date compilation of knowledge regarding CNVs in herbivorous livestock genomes. By integrating the latest research findings and insights, it is anticipated that this review will not only offer pertinent information but also stimulate future investigations into the realm of CNVs in livestock. In doing so, it endeavors to contribute to the enhancement of breeding strategies, genomic selection, and the overall improvement of herbivorous livestock production and resistance to diseases.

KEYWORDS

**copy number variation, herbivorous livestock, phenotypes, genome, molecular markers**

# 1 Introduction

China, renowned as one of the earliest nations to engage in livestock domestication (1), has a rich history of nurturing herbivorous livestock, including cattle, sheep, horses, and donkeys. This ancient practice has played a pivotal role in fulfilling diverse human needs, ranging from the procurement of essential animal-derived products such as meat, eggs, milk, and leather to harnessing domesticated animals for laborious tasks (2). Over time, the scope of domestication has expanded to encompass a multitude of applications. Throughout this evolutionary process, natural and artificial selection mechanisms have yielded an array of domestic animal breeds characterized by varying traits (3), including phenotypic attributes, economic characteristics, environmental adaptability, and resistance to diseases. Nonetheless, the intricate genetic underpinnings responsible for these disparities remain incompletely elucidated.

In recent years, the exploration of genomic variation has emerged as a central focus of scientific inquiry in the fields of animal production and health regulation, as evidenced by numerous studies (4–10). This emphasis on genetic variation holds significant significance in our quest to comprehend the intricate interplay between genetic diversity and a wide array of phenotypic and economic traits exhibited by animals (11–13). Furthermore, it serves as a robust theoretical foundation for elucidating genetic mechanisms and advancing the field of molecular breeding. Since the introduction of genomic selection, a range of livestock species, including sheep, goats, cattle, and horses, have undergone genotyping to assess their suitability for important economic traits, as demonstrated by previous studies (14–16). Up to this point, single nucleotide polymorphisms (SNPs) have been the primary focus of genomic research within the animal breeding community (17, 18). Significant strides have been made in establishing a solid genetic foundation for enhancing production and disease resistance in animals (18, 19). However, it is worth noting that despite these advancements, ~25% of the identified copy number variants (CNVs) exhibit no significant linkage disequilibrium with any SNP, leading to the conclusion that CNVs harbor genetic information that cannot be solely elucidated through SNP analysis (20).

CNVs are heritable chromosomal structural variations, characterized by deletions or insertions exceeding 50 base pairs (21). Notably, CNVs encompass a larger proportion of the genome compared to SNPs (22, 23). Consequently, CNVs are being proposed as an additional reservoir of information to elucidate the genetic variance underlying complex traits that may not be fully accounted for by SNPs alone (20). To date, several methodologies have been commonly employed for CNV detection, including comparative genome hybridization, extracting CNV data from SNP arrays, and whole-genome sequencing (WGS) approaches (24–26). Notably, recent research endeavors have delved into the investigation of the association between CNV in specific genes and a variety of phenotypic traits in animals, including growth characteristics in cattle (27, 28), goats (29–31), sheep (32) and horses (33–38). These studies have also extended their focus to examine the link between CNVs and other vital phenotypes, such as reproduction traits (39, 40) and disease resistance (7, 41). These studies have unveiled CNV as a key player linked to diverse facets of phenotypic diversity and economic traits in animal realms.

The confluence of two pivotal trends, the rising prevalence of molecular markers in livestock breeding and the maturation of microarray and sequencing technologies, necessitates a contemporary and comprehensive review of the burgeoning body of research on CNVs. This review paper seeks to illuminate the intrinsic value and biological ramifications of CNV in the landscape of genetic variation, with a particular focus on its potential as a potent molecular marker in the realm of livestock breeding. Through this approach, our review aims to introduce fresh viewpoints regarding genetic diversity and molecular breeding. Thus, in current review, we have focused on the progress of CNV screening methods in genomes of various herbivorous livestock including cattle, horses, donkeys, sheep and goat. In addition, we have briefly evaluated the association of CNVs in genes with different phenotypic traits including growth traits, reproductive traits, pigmentation, disease resistance, and environmental adaptability, etc., in cattle, horses, sheep and goat. The overall progress in screening CNVs within livestock genomes is summarized in Figure 1 (32–35, 37, 42–70).

# 2 Overview of CNV biology

## 2.1 CNV definition

A CNV, typically resulting from genome rearrangement, refers to the amplification or reduction in the copy number of a large genome segment of 1 kb or greater in length, primarily demonstrated via sub-microscopic deletions and duplications (50, 71). The common variant forms of CNV are illustrated in Figure 2 (72).

Several nearby CNVs and partially overlapping CNVs in the same genomic region can be merged into a single CNV segment (73). CNVs play a vital role in genomic structural variation (SV) (74). The number of base pairs regulated by CNVs is over five times greater than the number regulated by single nucleotide variants (SNVs) in each individual. Each CNV is correlated more than three times with one genome-wide association signal and fifty times with expression quantitative trait loci (eQTL) compared to SNVs (75). Furthermore, CNVs have been recognized for its significant influence on the evolution of phenotypic diversity, disease resistance, and evolutionary processes in organisms (76, 77).

It is widely recognized that CNVs are common in the genomes of plants, animals, and humans (78), but there is no final conclusion on the mechanism of CNV formation. According to existing studies, three primary formation mechanisms may be involved: (1) Non-allelic homologous Recombination, NAHR: Rearrangements occurring between homologous chromosomes of an individual's genomic DNA during meiosis can lead to duplications, deletions, inversions, and translocations (79, 80). (2) Non-homologous end-joining, NHEJ: A mechanism of genomic rearrangement that occurs during the period of DNA double-strand break repair and that can result in numerous simple CNVs (79, 80). (3) Replication fork smiling and template Switching, FoSTeS: Refers to the stalling of DNA replication forks which causes the lagging strand to break away from the DNA template and switch to other replication forks to continue DNA replication

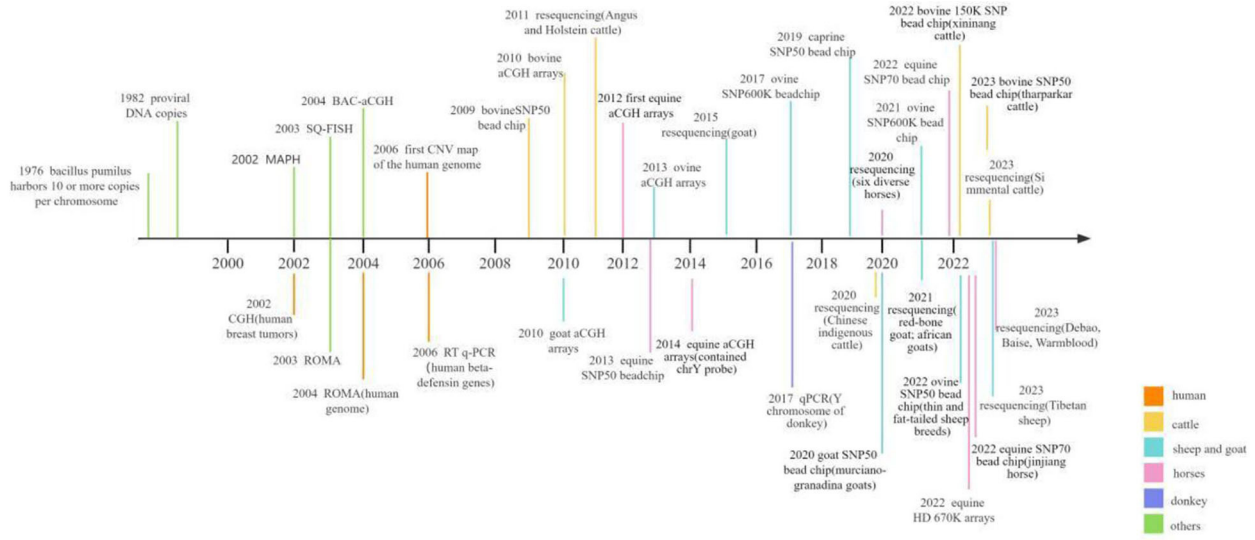



FIGURE 1
Graphical presentation of the research progress on CNV screening in herbivorous livestock and humans.

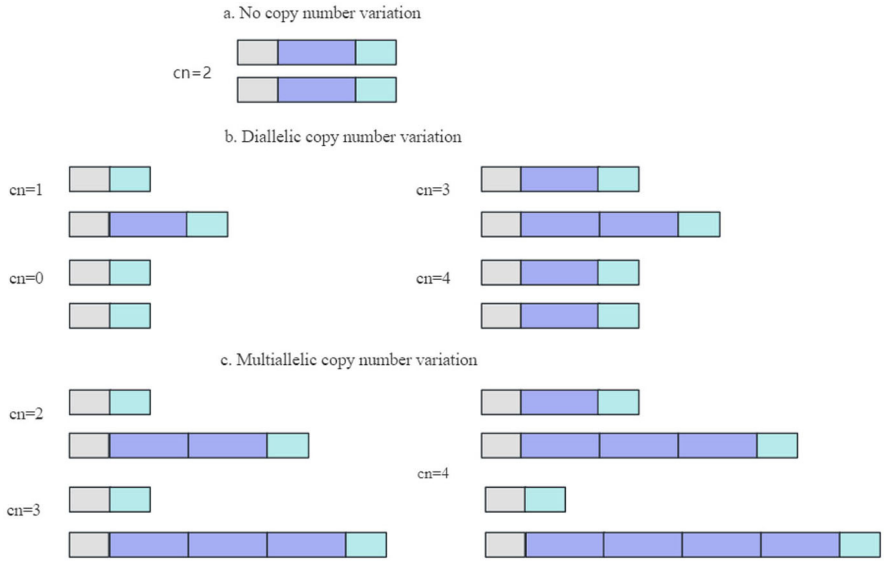



FIGURE 2
Common types of CNV.

synthesis. This can lead to DNA duplications or deletions and the emergence of a significant number of CNVs. Consequently, the formation of complex structural CNVs may be attributed to this phenomenon (80, 81).

## 2.2 CNV detection methods

### 2.2.1 Chip technology

Array comparative genomic hybridization (aCGH) has proven to be an efficient technique for the identification of CNVs at a genome-wide level (82–84). The fundamental principle of CGH involves labeling the DNA under analysis and control DNA with distinct fluorescent dyes. Subsequently, these treated DNAs are hybridized to standard chromosomes, and a digital fluorescence imaging system is employed to scan them, with a priority on fluorescence intensity ratios (85, 86). The aCGH is a microarray technology that utilizes two types of microarrays depending on the probes used during fabrication: Bacterial Artificial Chromosome CGH microarrays and Oligonucleotide Probe CGH microarrays. These aCGH microarray probes cover the entire genome and exhibit exceptional sensitivity, accuracy, and resolution, facilitating high-throughput screening. Previous research has provided robust support for the credibility of the experimental data generated by CGH (87, 88).

TABLE 1 Summary of CNVs and identification methods performed in cattle.

| References | Methods | Species | Samples | CNV number | CNVR number |
|---|---|---|---|---|---|
| Matukumalli et al. (51) | BovineSNP50 BeadChip | 6 | 576 | 79 | - |
| Liu et al. (52) | Bovine aCGH arrays | 3 | 90 | 1041 | 229 |
| Bae et al. (98) | BovineSNP50 BeadChip | 1 | 265 | 885 | 368 |
| Fadista et al. (22) | Bovine aCGH arrays | 4 | 20 | - | 304 |
| Stothard et al. (23) | Resequencing | 2 | 2 | 790 | - |
| Bickhart et al. (99) | Resequencing | 3 | 6 | 1265 | - |
| Zhang et al. (100) | Bovine aCGH arrays | 15 | 29 | - | 605 |
| Da Silva et al. (101) | Bovine HD Genotyping Bead Chip, Resequencing | 1 | 1717 | 68007 | 7319 |
| Zhou et al. (102) | BovineHD Genotyping BeadChip | 1 | 528 | 191 | - |
| Liu et al. (60) | Resequencing | 1 | 14 | - | 1344 |
| Mei et al. (63) | Resequencing | 8 | 75 | - | 11486 |
| Butty et al. (26) | BovineHD Beadchip, Genome Profiler Bovine 150 K, Genome Profiler Bovine HD, BovineSNP50 Beadchip | 1 | 96 | - | 52/36 |
| Butty et al. (24) | BovineHD Beadchip, Genome Profiler Bovine 150K, Genome Profiler Bovine HD, BovineSNP50 Beadchip, Genome Profiler Bovine 50K | 1 | 5845 | 23256 | 1645 |
| Zhou et al. (67) | Bovine 150K SNP BeadChip | 1 | 403 | - | 38/33 |
| Kooverjee et al. (103) | Resequencing | 3 | 5 | - | 355 |
| Kumar et al. (69) | BovineSNP50 BeadChip | 1 | 72 | 693 | 447 |
| Sun et al. (104) | Resequencing | 1 | 30 | - | 2944 |

Single nucleotide polymorphism (SNP) microarray technology represents another effective method for detecting CNVs. SNP microarray technology requires only a single hybridization, thus obviating the need for the simultaneous double hybridization of two DNA samples with probes. It determines the genomic copy equivalents of each locus by comparing variations in signal intensities between the samples being tested (89). A SNP microarray demonstrates remarkable stability and high resolution, enabling the detection of diverse forms of CNVs, including sub-microscopic deletions, duplications, and more (90). In comparison to Comparative Genomic Hybridization, SNP microarrays offer the advantage of simultaneously detecting CNVs while determining their genotypes and revealing heterozygous deletions. This approach is not only more cost-effective but also facilitates large-scale CNV testing (91). Several software and programs are currently available for CNV detection using chip technology, including CNVPartition, PennCNV, and QuantiSNP. A study revealed through genome sequencing of horses that the optimal order of performance for the three assays was CNVPartition, PennCNV, and QuantiSNP (92). Furthermore, the combination of PennCNV and QuantiSNP exhibited improved accuracy in CNV detection.

### 2.2.2 Sequencing technology

As gene sequencing technology has matured, numerous tools and software have been developed to enhance the efficiency and precision of copy number variant detection. Next-generation sequencing (NGS) has emerged as the most commonly used method for detecting CNVs in recent years (93), with Illumina's Solexa/HiSeq technology being a prominent representative. Second-generation sequencing technology aims to synthesize and sequence DNA simultaneously. Fluorescent signals are excited by lasers and recorded using optical equipment. These recorded signals are then converted into bases using computer technology. Second-generation sequencing methods are not limited to target sequences hybridized with primer probes and can identify genome-wide CNVs, structural variants, and other variations (94). In the current landscape, four primary strategies and methods are employed for detecting variants in NGS data: paired-end mapping (PEM), split-read (SR), read depth (RD), and *de novo* assembly (AS), in addition to a combined approach based on the above four methods. Compared to microarray-based methods, next-generation sequencing technology offers advantages in terms of speed, resolution, cost-effectiveness, and reproducibility (95–97). Furthermore, widely used sequencing technologies for identifying CNVs, software including CNVnator, CNVpytor, CNVcaller, and ERDS have been documented.

## 3 Overview of screening methods for CNVs in herbivorous livestock genomes

### 3.1 Cattle genome

Cattle ranching represent a promising industry with substantial economic, ecological, and social implications. The diversity of

cattle breeds, including Holstein for dairy production, Charolais for beef, and regionally adapted yaks and buffaloes, caters to a wide range of human needs. CNVs assume significant importance in the exploration of phenotypic traits and adaptation to various environments, shedding light on the domestication origins of these animals. As dietary preferences evolve toward beef with higher protein and lower fat content, copy number variation regions (CNVRs) have been found to be associated with cattle carcass traits, offering theoretical support for future breeding endeavors. In recent years, numerous investigations have identified CNVs within the bovine genome. Table 1 provides a comprehensive summary of some of these studies. Consistently, a study identified a total of 79 CNV loci across six distinct cattle breeds, employing the BovineSNP50 BeadChip (51). Notably, 10 CNVs were found to overlap with those previously identified through aCGH data. Similarly another study conducted an extensive genome-wide analysis of CNVs in 90 modern domesticated cattle, uncovering over 200 potential CNVRs (52). In addition, this high-quality bovine CNV map fills critical gaps in current genome-wide association and selection studies based on SNP genotyping. Previous studies generated genome-wide CNV maps for cattle using the BovineSNP50 BeadChip and aCGH arrays, respectively (22, 98). Similarly, Kumar et al. (69) pioneered CNV detection in Indian Tharparkar cattle, identifying a total of 8881 CNVs, which were subsequently filtered down to 693 CNVs and merged into 447 CNVRs, representing ∼2.17% of the cattle genome. Utilizing a purebred Angus cow as a reference, a study documented 605 CNVRs through a genome-wide analysis of CNVs in comparative genomic crossbreeding arrays of 29 Chinese domestic bulls (100). Detailed distribution maps of these CNVRs were constructed on their respective genomes. In a consistent manner, Kooverjee et al. (103) conducted a comprehensive investigation, wherein they successfully identified a total of 355 CNVRs in a cohort of five crossbred cows. This identification was accomplished through the utilization of the Panelcn.MOPS software. Notably, these CNVRs exhibited an average length of 9318 base pairs, collectively representing ∼0.15% coverage of the bovine genome. Previous studies employed different platforms for genotyping Holstein cattle, followed by CNV analysis, to investigate the impact of genotype array density on CNV detection, thereby contributing to our understanding of genetic variation in Holstein cattle (24, 26). Accordingly, Sun et al. (104) harnessed sequencing technology to sequence the entire genome of Simmental bulls, detecting 2944 CNVRs, which were subsequently subject to genetic analysis, revealing associations with reproduction, immunity, and fertility. These findings constitute a valuable molecular breeding resource for cattle. A study analyzed common CNV regions in Xinjiang brown cattle and compared differences between the ARS and UMD reference genomes, suggesting the ARS reference genome's superior effectiveness in CNV detection (102). Likewise, previously studies conducted on CNVs in the bovine genome utilizing the BovineHD Genotyping BeadChip (38, 102). Conversely, studies undertook an analysis of CNVs in the bovine genome through sequencing technologies (23, 63, 99, 105).

## 3.2 Sheep and goat genomes

Sheep products, encompassing mutton meat, milk, wool, and cashmere, are of considerable significance within the sheep industry. They not only enhance the quality of life for individuals but also contribute substantially to industrial development, augment the incomes of farmers and herdsmen, and provide high-quality fertilizers for farmland, among other advantages. Genomic selection technology has emerged as a pivotal approach in sheep breeding. Current research has unveiled associations between CNVs and various traits in sheep, including growth characteristics, wool color, cashmere quality, disease resistance, and reproduction.

Numerous studies focused on mapping CNVs within the sheep genome have significantly enriched our understanding of genomic variations in sheep. Table 2 provides a comprehensive summary of research pertaining to CNVs in the sheep genome. A study identified a total of 238 CNVRs and establishing a CNV map in the genomes of three distinct sheep breeds using the sheep SNP50K microarray (54). Consequently, Ma et al. (106) investigated CNVs in the genomes of eight sheep breeds using the sheep SNP50K microarray, identifying 111 CNV regions from 160 sheep and mapping the distribution of CNVRs across autosomal chromosomes. Consistently, total of 13,347 CNVs based on sequencing data from six domesticated goats and two wild goats were detected (56). While another experimental trial employed high-density sheep SNP microarrays to identify 371, 301, and 66 autosomal CNVRs within the genomes of big-tailed frigid sheep, Altai sheep, and Tibetan sheep, respectively (107). This endeavor resulted in the creation of the first high-resolution sheep CNV map, offering a valuable resource for comprehending genomic variation in sheep. In continuity, Yang et al. (108) identified 24,558 CNVs from 2,254 sheep across various geographic regions worldwide, culminating in 619 CNV regions with a combined length of 197 megabases (Mb). This encompasses 6.9% of the sheep genome and establishes a comprehensive CNV map that can assist in genome annotation for sheep. In addition, Igoshin et al. (111) detected 4,527 CNVs among 354 sheep representing 16 Russian indigenous breeds. Gene function enrichment analysis revealed significant impacts of CNVs on olfactory perception and immunity within 12 of the breeds. In line, Taghizadeh et al. (68) identified 328 and 187 CNVRs in fat-tailed and thin-tailed sheep breeds, respectively. These CNVRs were found to be located within or overlapping with 790 known sheep genes, covering ∼73.85 Mb of the sheep genome. Previous report conducted a genome-wide analysis on 48 beach sheep, documented 1,296 CNV regions and constructed a CNVs map of the Tan sheep genome, thereby complemented the data on CNVs in the Chinese sheep genome (59). While another study identified 6,286 CNVs in a total of 1,023 sheep representing 50 different breeds worldwide, employing the pennCNV tool. The results unveiled differences in CNVs among populations across different geographic regions (60). Furthermore, a study reported 42 CNVs from 120 samples representing five dairy goat breeds and established significant associations between two CNVs (CNV5 and CNV25) and two milk production traits (109). Consequently, Guan et al. (62) detected 1,461 regions of

CNV within the Spanish dairy goat genome, with an average length of 196.89 kilobases (kb). The total length of all CNV regions accounted for 3.9% of the autosomal genome, leading to the creation of a CNV map. Previous study has identified 127 CNVRs in four breeds of goats, covering ~11.39 Mb of the bovine genome, thereby establishing the first CNVR map for goats (70). In addition, a study reported 1,217 CNVRs in 67 sheep breeds worldwide (66). Furthermore, a study identified 4,769 high-quality CNVRs in 47 sheep breeds globally, subsequently generating CNV maps. Additionally, they investigated the influence of solar radiation on CNVs within sheep genomes (66). Consistently, a study employed resequencing to detect CNVs in two Tibetan sheep breeds, ultimately identifying 368 distinct CNVRs, which may contribute to determining population disparities (32). Yuan and co-authors were the first to utilize resequencing technology to establish a CNV map of Chinese fine-wool sheep and analyzed the overlap of CNVRs with several quantitative trait loci related to economic traits, providing vital insights for the future improvement of fine-wool sheep (110). Consistently previous studies employed resequencing to analyze the genomic CNVs of Mile red-boned goats and African goats, respectively (64, 65). These findings facilitated a deeper understanding of the genetic traits of these animals. Consequently, another study identified a total of 702 CNVs in 120 dairy goats, resulting in the creation of a CNVR map, which promises to be beneficial for further research on the association between CNVs and phenotypic variations (13).

## 3.3 Horses genome

China's horse industry has experienced remarkable growth owing to the country's rapid social and economic development, coupled with continuous improvements in living standards. Consequently, equestrian sports have gained significant popularity as leisure and recreational activities within the nation. The maturation of microarray and sequencing technologies has played a pivotal role in the identification of equine CNVs. These CNVs are of paramount importance for the study of equine trait variations, disease prevention and treatment, exploration of genetic diversity, tracing the origins of domestication, and the development of new equine breeds.

In recent years, the substantial progress in microarray and sequencing technologies has facilitated the detection of CNVs. Researchers from both domestic and international institutions have undertaken extensive investigations into equine genome CNV identification, with select research findings summarized in Table 3. Doan et al. (53) were the pioneers in reporting CNVs within equine genomes. They identified 775 CNV regions in 16 horses using an aCGH microarray, demonstrating the significant impact of CNVs on biological phenotypic diversity. Subsequently, in an effort to detect CNVs in both normal horses and Przewalski horses, Ghosh et al. (55) prepared an aCGH microarray that also included a Y chromosome probe. They successfully detected 258 CNVRs in autosomes, chrX, and chrUn, but none were found in chrY. Notably, the majority of these CNVRs were associated with genes related to sensory perception, the immune system, and reproduction (55). Previous study meticulously described the mapping of CNVs in Chinese horses using high-resolution array Comparative Genomic Hybridization (aCGH), a highly effective method for genome-wide CNV detection in animals (114). Consequently, Kader et al. (57) through whole-genome analysis of CNVs in 96 horses representing three Chinese breeds: Debo Shorthorn, Mongolian horse, and Yili horse. Their work identified a total of 287 CNVs, which were combined to form 122 CNV regions (CNVRs) with sizes ranging from 199 base pairs to 2,344 kilobases. Consistently, a study reported 15,041 CNVs and 5,350 CNVRs in 222 Friesian horses, creating a distribution map of CNVRs within the equine genome (115). Similarly, a recent study performed an analysis of CNV in 469 horses from four Korean breeds, uncovering 843 CNVRs that overlapped with 7.2% of the reference genome for horses (33). Furthermore, they constructed an autosomal map of CNVRs in horses. In addition, Laseca et al. (34) analyzed high-density SNP genotyping data from 654 horses, identifying a total of 19,902 CNV segments and 1,007 CNV regions, with CNVs covering 4.4% of the equine genome.

A study utilized gene chips to identify CNVs in 1,755 horses representing eight breeds. Their findings revealed that the size of CNV regions varied from 1 kilobase to 21.3 megabases (116). Consequently a study revealed an average total of 1,540 CNVs per horse through whole-genome resequencing of six horses representing six different breeds. Their results suggested that a reduction in the number of LATH copies might be linked to the development of endurance in horses (61). While a recent study by Choudhury et al. (37) leveraged resequencing data for Debao (DB), Baise (BS), and Warmblood (WB) horses to identify CNVs and create a CNVR map of the equine genome. Their research indicated that differential CNVRs may influence the phenotypic characteristics of different breeds.

## 3.4 Donkey genome

The donkey industry has emerged as a significant contributor to the growth of the livestock sector in recent years. However, the detection of CNVs within the genetic variations of the donkey genome, both in China and internationally, remains a relatively understudied area. In a study conducted by Han, a cohort comprising 263 native Chinese donkeys representing 13 breeds from eight provinces and regions was employed to identify CNVs in five Y chromosome genes of donkeys (CUL4BY, ETSTY1, ETSTY4, ETSTY5, and SRY) through quantitative Polymerase Chain Reaction (qPCR) analysis (58). While studies related to Single Nucleotide Polymorphisms (SNPs) have been reported, studies focusing on CNVs in donkeys are limited. Consistently, another study detected ~7 million SNPs in 126 domestic donkeys and made a noteworthy discovery that black or chestnut coat color was attributed to a 1 base pair deletion downstream of the TBX3 gene (117). This deletion led to reduced gene expression and its inhibitory effect on pigmentation. In addition to SNPs, CNVs represent crucial genetic resources. They possess distinct advantages, including their ubiquitous presence in the genome, extensive coverage, and relative ease of detection compared to SNPs (50, 118).

TABLE 2 Summary of CNVs and identification methods performed in sheep and goat.

| References | Methods | Species | Samples | CNV number | CNVR number |
|---|---|---|---|---|---|
| Fontanesi et al. (70) | aCGH | 4 | 9 | 161 | 127 |
| Liu et al. (54) | Ovine SNP50 BeadChip | 3 | 50 | 3624 | 238 |
| Ma et al. (106) | Ovine SNP50 BeadChip | 8 | 160 | 173 | 111 |
| Dong et al. (56) | Resequencing | 2 | 8 | 13347 | - |
| Zhu et al. (107) | Ovine HD 600K SNP arrays | 3 | 110 | - | 738 |
| Ma et al. (59) | Ovine SNP600K BeadChip | 1 | 48 | 5190 | 1296 |
| Yang et al. (108) | Ovine SNP50 BeadChip | 68 | 2111 | 24588 | 619 |
| Liu et al. (105) | CaprineSNP50 BeadChip | 50 | 1023 | 6286 | 978 |
| Kang et al. (109) | CaprineSNP50 BeadChip | 5 | 120 | 42 | - |
| Guan et al. (62) | Goat SNP50 BeadChip | 1 | 1036 | - | 1461 |
| Di Gerlando et al. (13) | GoatSNP50 BeadChip | 4 | 120 | 702 | 75 |
| Yuan et al. (110) | Resequencing | 3 | 32 | 1747604 | 7228 |
| Salehian-Dehkordi et al. (66) | Ovine SNP600K BeadChip | 67 | 2059 | 18152 | 1217 |
| He et al. (64) | Resequencing | 2 | 72 | 5862 | - |
| Nandolo et al. (65) | Resequencing | 34 | 82 | 253553 | 6231 |
| Igoshin et al. (111) | Ovine Infinium HD SNP BeadChip | 16 | 354 | 4527 | 1450 |
| Taghizadeh et al. (68) | OvineSNP50 Beadchip | 3 | 192 | 815 | 515 |
| Salehian-Dehkordi et al. (112) | Ovine SNP600K BeadChip | 47 | 695 | 39145 | 4769 |
| Shi et al. (32) | Resequencing | 2 | 20 | 60429 | 4927 |

The continued advancement of sequencing technology promises to greatly facilitate the identification of CNVs within the donkey genome.

The subsequent section offers insights into potential reasons why CNVs have been relatively overlooked in donkey research:

1. Donkeys predominantly serve as working animals in many parts of the world. Consequently, researchers may prioritize the study of genetic markers associated with phenotype traits in species other than donkeys.
2. Initially, copy number variation was primarily detected using gene chips, which were less readily available for donkeys compared to other livestock species such as cattle, sheep, and horses. In recent years, the development of sequencing technology and the completion of whole-genome sequencing for donkeys have created favorable conditions for the detection of CNVs within the donkey genome.
3. Funding for research pertaining to donkeys may be comparatively constrained when compared to other livestock species like cattle and sheep. Consequently, researchers may face challenges in securing adequate resources for the study of CNVs in donkeys.
4. In many parts of the world, the commercial value of donkeys is primarily linked to their performance as working animals, rather than their utility for products such as milk or meat. As a result, there may be limited commercial interest in investigating CNVs in donkeys.

## 4 Gene ontology analysis for genes overlapping CNVRs

To identify genes that may be influenced by CNVRs within the genome, an annotation analysis of genes associated with CNVRs revealed that olfactory-related functions, specifically olfactory transduction and olfactory receptor activity, were frequently affected. This observation aligns with previous findings in various species, including cattle (119), sheep (62, 111), horses (34, 115, 120) and humans (121–123), where genes affected by copy number variation have shown enrichment for olfactory-related functions. In the case of humans, the sense of smell is considered a minor aspect of overall health and may not be closely linked with adaptation. Consequently, human olfactory receptor (OR) genes tend to evolve neutrally (121). However, in the animal kingdom, the sense of smell holds paramount importance as it plays a vital role in locating food, identifying harmful substances, avoiding predators, selecting mates, and ensuring long and healthy survival and reproduction (124). Consistently, a study proposed that olfactory receptors also play a role in appetite regulation and feeding efficiency in mammals. Hence, alterations in these receptors may lead to individual differences in feed intake, body weight, and body composition (125). Additionally, CNVRs have been found to intersect with Quantitative Trait Loci (QTL) associated with various factors such as morphology, disease resistance, and more. This intersection serves as a fundamental foundation for the examination of phenotypic diversity. Variations in CNVR frequency among different breeds have been identified in CNVR

TABLE 3 Summary of CNVs and identification methods performed in horses.

| References | Methods | Species | Samples | CNVs | CNVRs |
|---|---|---|---|---|---|
| Doan et al. (53) | aCGH arrays | 15 | 16 | 2368 | 775 |
| Dupuis et al. (113) | Equine SNP50 bead-chip array | 4 | 477 | 2797 | 478 |
| Metzger et al. (92) | Equine SNP50 bead-chip array | 17 | 717 | 50 | - |
| Ghosh et al. (55) | aCGH arrays | 16 | 38 | - | 258 |
| Wang et al. (114) | aCGH arrays | 6 | 6 | 700 | 353 |
| Kader et al. (57) | Equine SNP70 bead-chip array | 3 | 96 | 287 | 122 |
| Schurink et al. (115) | Equine genotyping array | 1 | 222 | 15041 | 5350 |
| Solé et al. (116) | Equine genotyping array | 8 | 1755 | 18800 | 939 |
| Al Abri et al. (61) | Resequencing | 6 | 6 | 1540 | - |
| Kim et al. (33) | Equine SNP70 bead-chip array | 4 | 469 | - | 843 |
| Laseca et al. (34) | Equine high-density 670K | 1 | 654 | 19902 | 1007 |
| Wang et al. (35) | Equine SNP70 bead-chip array | 10 | 70 | 577 | 239 |
| Choudhury et al. (37) | Resequencing | 3 | 26 | 18974 | 4279 |

tests involving animals from various regions (52, 111, 126). These variations can be attributed to breed domestication and environmental adaptation.

## 5 CNVs and their association with phenotypic traits in herbivorous livestock

It has been well-established that genomic CNVs exert an impact on an organism's phenotype through various mechanisms, including changes in gene dosage, modulation of gene expression, modulation of gene transcriptional regulators, and positional effects (127). The association of CNVs in genes and their association with various phenotypic traits (growth traits, reproductive traits, pigmentation, and diseases resistance) in herbivorous livestock (cattle, sheep, goat, and horses) have been summarized in Table 4.

### 5.1 CNVs associated with growth and reproductive traits in herbivorous livestock

CNVs have been extensively studied in various livestock species, including cattle, sheep, donkey and horses, and its association with important growth, reproduction, and fertility traits has been documented. This comprehensive review discusses key findings and contributions from various research studies in these livestock species. In cattle, Yang et al. (131) identified CNV in the cytochrome P-450 4A11 (CYP4A11) gene, which was associated with increased growth. Multiple copies of CYP4A11 were found to promote the differentiation of 3T3-L1 cells into adipocytes, potentially leading to increased fat deposition. Previous studies investigated SERPINA3-1 and GAL3ST1 gene CNVs in different Chinese cattle breeds, revealed associations with growth traits such as body height, body weight, and rump width (146, 180). These genes hold promise as candidate genes for Chinese cattle breeding. Correspondingly, Hu et al. (139) analyzed CCDC39 gene CNVs and their impact on body length and hip width, noting significant effects, particularly in the Pinan (PN) breed. Additionally, Shi et al. (128) found a correlation between leptin gene CNV and various phenotypic traits, including body weight, body height, body length, and brisket circumference, in multiple cattle breeds.

In sheep, Zhu et al. (107) identified adiposity-related genes, including PPARA, RXRA, KLF11, ADD1, FASN, PPP1CA, and PDGFA, in CNV regions of fat-tailed sheep. These genes were associated with fat deposition, with individuals carrying copy number deletions exhibiting higher body weight. Similarly, a recent study by Wang et al. (155) documented a significant correlation between CNV in the KAT6A gene and body height and hip width in Hu sheep (HU). They also found that CNV3 duplicates were associated with higher body height and weight. Yang et al. (153) highlighted the BAG4 gene's role in regulating body height in sheep and its potential as a molecular marker for molecular breeding. Feng et al. (156) discovered that CNVs of the PIGY gene significantly impacted body weight, chest circumference, and tube circumference in sheep. Additionally, Xu et al. (29) found that CNV types of the CCSER1 gene were correlated with body weight and heart girth traits in Guizhou white goats (GZW).

Horses were also subject to CNV studies. Consistently, Metzger et al. (92) conducted a genome-wide analysis of CNVs and their association with equine body height traits. They identified deleted regions in ECA1, ECA8, and ECA9, which were significantly linked to equine body height. While Kim et al. (33) reported CNVRs with overlapping quantitative trait loci (QTLs) associated with equine body height in Jeju riding horses and Hanra horses. These findings provided valuable insights into the genetic factors influencing equine body height.

In the context of fertility, CNVs have also been explored. Consistently, a deletion in the intronic region of the SPAG16 gene has been identified in bulls with poor sperm motility (PSM), suggesting its potential role in bull fertility (104). A comprehensive

TABLE 4 Summary of CNVs in genes and their association with phenotypic traits in herbivorous livestock.

| Genes (CNVs) | Phenotypes | Biological effect | Species | References |
|---|---|---|---|---|
| Changes in the CNV region within the LEPR intron 3 | Growth traits | Body weight, body height, body length, and brisket circumference | Cattle | (128) |
| Myosin heavy chain 3 (MYH3)-CNV | | Skeletal muscle development | | (129) |
| Mitogen-activated protein kinase 10 (MAPK10)-CNV | | Body weight ($P < 0.05$), body height and chest girth | | (130) |
| CYP4A11-CNV | | Lipid deposition | | (131) |
| Guanylate binding protein 2 (GBP2)-CNV | | Body height, body length, heart girth, hip width, rump length | | (132) |
| Insulin-like growth factor 1 receptor (IGF1R)-CNV | | Associated with body weight and body height of Jinnan cattle and was significantly linked with body height and hucklebone width of Qinchuan cattle | | (133) |
| Kupple like factor 3 (*KLF3*)-CNV | | Body mass and heart girth | | (134) |
| Potassium inwardly-rectifying channel, subfamily J 12 (KCNJ12)-CNV (1&2) | | Significant association with the body length, chest circumference, body weight, rump length | | (135) |
| MLLT10-CNV | | Hip width, rump length, hucklebone width, and cannon bone circumference | | (136) |
| Uanylate-binding protein 6 (GBP6)-CNV | | Associated with body weight, cannon circumference and chest circumference | | (137) |
| (CNV1: 3600 bp, including exon 2–11; CNV2: 4800 bp, including exon 21–22) of the CLCN2 gene | | Cannon circumference, body slanting length, chest girth, and body weight | | (138) |
| CCDC39- CNV (Normal, deletion, duplication) | | body length, hip width, heart girth and cannon bone, and circumference | | (139) |
| PLA2G2A-CNV *(Normal, Deletion)* | | Height at sacrum, heart girth and body height, chest depth | | (140) |
| SYT11-CNV | | Significantly correlated with body length, cannon circumference, chest depth, rump length, and forehead size of Yunling cattle, and was significantly correlated with the bodyweight of Xianan cattle | | (141) |
| Mitochondrial fusion protein (MFN1)-CNV | | Significant correlation with hucklebone width, hip width, height at sacrum, chest width and rump length | | (142) |
| DYNC1I2-CNV (Duplication and deletion) | | Associated with height at hip cross, body length, chest width and hucklebone width, chest depth | | (143) |
| WW domain binding protein 1-like (WBP1L)-CNV | | Associated with heart girth, rump length and body weight (Pinan cattle), withers height, rump length, body length, chest depth and BW of (Jiaxian cattle) | | (144) |
| MUC19-CNV | | Correlated with hip width, height at hip cross and withers height, body length, and huckle bone width | | (145) |
| SERPINA3-1-CNV | | Body height | | (12) |
| GAL3ST1- CNV (deletion) | | Body weight | | (146) |
| VAMP7-Duplication | | Body growth trait (height at the hip cross) | | (147) |
| ZNF679-CNV | | Body size and length | | (148) |
| CNVRs harbored genes (PPARA, RXRA, KLF11, ADD1, FASN, PPP1CA, PDGFA, and PEX6) | | Fat deposition | Sheep | (107) |
| Src homology 2 domain containing E (SHE)-CNV | | Correlated to body length, circumference of cannon bone, heart girth, chest width and high at the cross | | (149) |
| ORMDL sphingolipid biosynthesis regulator 1 (ORMDL1)-CNV | | Body weight, body height, body length, chest depth, and height of hip cross | | (150) |
| KMT2D-CNV | | Body length, withers height, hip width | | (151) |
| TOP2B-CNV | | Body length, chest circumference, canon circumference and height of hip cross | | (152) |

*(Continued)*

TABLE 4 (Continued)

| Genes (CNVs) | Phenotypes | Biological effect | Species | References |
|---|---|---|---|---|
| BAG4 -CNV | | Body height, body slanting length, body height and hip cross height | | (153) |
| TOP2B-CNV | | Body length, chest circumference, canon circumference and height of hip cross | | (152) |
| LRRFIP1-CNV | | Chest width, rump breadth and circumference of cannon, larger heart girth | | (154) |
| KAT6A-(CNV1, CNV2, CNV3) | | Body height and body length | | (155) |
| PIGY- CNV | | Body weight, chest circumference, and tube circumference | | (156) |
| Myosin light chain kinase-4(MYLK)-CNV | | Body weight, body length and body height | Goat | (157) |
| Opn4-CNV | | Body weight in Guizhou white goat<br>Body length in Guizhou black goat | | (158) |
| SNX29 gene-CNV<br>ADCY1-CNV | | Meat production traits | | (159) |
| CADM2-CNV (Deletion) | | Withers height and body length | | (160) |
| Sorting nexin 29 (SNX29)-CNV | | Body length, body height, heart girth, chest width, canon circumference | | (161) |
| Myogenic differentiation 1 (MyoD1)-CNV | | Body weight, height at hip cross, heart girth and hip width | | (30) |
| A-kinase-anchoring protein 13 (AKAP13)-CNV | | Body height and body length, chest depth, chest circumference, and cannon circumference | | (162) |
| Pleomorphic adenoma gene 1 (PLAG1)-CNV | | Body weight, heart girth, height at hip cross, and hip width | | (163) |
| CCSER1-CNV (deletion) | | Body weight and heart girth traits | | (29) |
| Deleted regions on ECA1, ECA8 and ECA9. | | Body height | Horses | (92) |
| 23 CNVRs with overlapping QTLs associated with equine body height | | Body height | | (33) |
| A deletion in the intronic region of the SPAG16 gene | Reproductive traits | Bull-fertility traits (sperm motility) | Cattle | (104) |
| CNV of ZNF280BY | | Negative correlation with the percentage of normal sperm and sperm concentration | | (164) |
| CNVs of the bovine HSFY and ZNF280BY | | Correlated negatively with testis size, while positively with sire conception rate. | | (165) |
| CNV of *ZNF280BY* | | Negatively correlated with testis size | Hu sheep | (166) |
| SMAD2 | | Litter size and semen quality | Goat | (40) |
| Sorting nexin 29 (SNX29)-CNV (Indel) | | Litter size and fertility | | (167) |
| Protein phosphatase 3 catalytic subunit alpha (PPP3CA)-CNV | | Litter size and semen quality | | (168) |
| PRP1 and PRP6 have CNV mutations in the HF group | | Litter size | | (169) |
| CNVs in regions of the Y chromosome | | Male development and equine fertility | Horses | (170) |
| An 1155 bp deletion in the ASIP gene | | Coat color darkening | | (171) |
| 2,809 bp LINE-1 insertion in ASIP gene | | White coat color | | (172) |
| ASIP-CNV (Duplication) | | Light coat color | Goat | (56) |
| ASIP- CNV | | White coat colounr | | (173) |
| 1 Mb CNV affect EDNRA gene | | White coat color | | (174) |
| 13.42kb duplication upstream of ASIP | | Non-classic swiss markings | | (175) |

*(Continued)*

TABLE 4 (Continued)

| Genes (CNVs) | Phenotypes | Biological effect | Species | References |
|---|---|---|---|---|
| A 4.6 kb duplication in intron 6 of STX17 | | Gray phenotype | Horses | (176) |
| A deletion including exon 3 of the ED1 gene | Disease resistance | Anhidrotic Ectodermal Dysplasia | Cattle | (177) |
| Low relative expression levels of KIF2A and PHKG2 | | Disease resistance | Sheep | (106) |
| CNVS in CCL1, CCL2, CCL8, CCL11, NOS2, TNF, CSF3, and STAT3 genes | | Resistance to natural *Haemonchus contortus* infections | | (41) |
| CNVR33, CNVR65, and CNVR7 overlap with immune system-related genes | | Strong resistance to infectious diseases | Goat | (13) |
| CNVRs located in the MHC region of ECA20 | | Insect bite hypersensitivity | Horses | (115) |
| A pure deletion of the AKR1C gene | | Disorders of sexual development | | (55) |
| CNV in GRIK4, IFNLR1, and LOC102275985 | Environmental adaptability | High-altitude adaptation | Cattle | (178) |
| CNV in LDHB and ME1 | | Cold and low oxygen environments | | (179) |
| CNV changes affect in ALKBH5, NARFL genes | | Plateau acclimatization | Sheep | (107) |
| CNVR is significantly correlated with solar radiation | | Solar radiation | | (112) |
| Genes associated with hemoglobin binding located on CNVRs | | Harsh plateau environment | Horses | (114) |
| Changes in NFKBIA, SOCS4, HSPA1A, and IL6 genes located in the CNVR | | High temperatures and humidity | | (35) |

analysis documented CNVs in Laoshan dairy goat populations with differing fertility levels and showed that CNV mutations in PRP1 and PRP6 genes, which affect mammalian fertility (169). A study investigated CNVs in Y chromosome-specific regions in male horses, identifying potential genes linked to stallion fertility and contributing to our understanding of equine male development and fertility (170). These studies collectively underscore the significance of CNVs in shaping various traits in cattle, sheep, and horses, from growth and fat deposition to fertility. The identification of specific genes and regions associated with these traits holds promise for selective breeding programs and further genetic research in these livestock species.

## 5.2 Pigmentation

The role of CNVs in determining coat color and disease resistance in cattle, horses, and sheep has been the subject of extensive scientific investigation. In this discussion, we will delve into various studies that have shed light on the influence of CNVs on these traits in these livestock species. A study reported two sequence translocations between chromosomes 6 and 29 in Belgian Blue and Swiss Brown cattle, affecting the KIT gene, leading to color-sidedness (181). Similarly another study reported that a deletion of 1155 bp within the ASIP gene in Nellore cattle results in dark hair color in specific regions by elevating melanin production (171). Furthermore, a 2,809 bp LINE-1 insertion in the ASIP gene icausing a white coat color phenotype by impeding melanin production has been identified in buffalo (172). Consequently a study postulated that a 4.6 kb duplication within intron 6 of the STX17 gene leads to a pronounced upregulation of both STX17 and NR4A3 gene expression (176). This heightened gene expression subsequently instigates the proliferation of melanocytes, ultimately culminating in the manifestation of a gray coat phenotype in affected horses. Furthermore, it is noteworthy that horses harboring this mutant phenotype exhibit a gradual transition in hair color from gray to white as they advance in age. In a similar vein, previous investigations encompassed genomic analyses of feral and domestic goat populations, which unveiled the intriguing revelation that 13 genes situated within CNV regions overlap with the comprehensive roster of cloned color genes provided by the European Society for Pigment Cell Research (ESPCR) (56). Moreover, these investigations substantiated the substantial impact of CNVs within the ASIP gene on the lightening of coat color in domestic goat breeds, employing rigorous resequencing analyses. A subsequent inquiry brought to light the potential influence of CNVs on the ASIP gene, which may, in turn, lead to the emergence of a white coat color phenotype in the Girgentana and Saanen goat breeds, as corroborated by Fontanesi et al. (173). A recent report presented compelling evidence demonstrating the significant association between a 13.42 kb repeat sequence located upstream of the ASIP gene and non-classic Swiss markings in goats, utilizing CNV assays and quantitative Polymerase Chain Reaction (qPCR) techniques (175). This finding further underscores the pivotal role of CNVs in shaping coat color patterns in goats. Furthermore, Menzi et al. (174) unraveled the involvement of the EDNRA gene, situated within a 1 Mb CNV region on chromosome 17, in potentially attenuating melanism among Boer goats. Notably, an elevation in the copy number within this CNV region could potentially lead to a reduction

in skin pigmentation, thereby culminating in the manifestation of a white-spotted phenotype. In summary, these comprehensive investigations collectively shed light on the intricate interplay between CNVs and the genetic determinants of coat color diversity in horses and goats, providing valuable insights into the underlying genetic mechanisms governing these phenotypic variations.

## 5.3 Disease resistance

The CNVs hold significant potential to influence disease resistance in livestock species, including cattle, sheep, and horses. Notably, Liu et al. (52) conducted an exhaustive investigation in cattle, uncovering multiple CNVs that play pivotal roles in crucial biological processes such as drug detoxification, innate and adaptive immunity, as well as receptor and signal recognition. In the context of cattle, it is worth highlighting the association between CNVs and disease susceptibility. For instance, Drögemüller et al. (177) reported that a deletion encompassing exon 3 of the ED1 gene has been linked to anhidrotic ectodermal dysplasia, underscoring the critical role of CNVs in disease vulnerability. Furthermore, a study reported CNVs in 18 candidate genes (TERT, NOTCH1, SLC6A3, CLPTM1L, PPARα, BCL-2, ABO, VAV2, CACNA1S, TRAF2, RELA, ELF3, DBH, CDK5, NF2, FASN, EWSR1 and MAP3K11) which were associated with milk somatic cells count and mastitis resistance in dairy cattle (182). Consistently, another study reported that CNVs in ZNF496 and NLRP3 were significantly associated with resistance to gastrointestinal nematodes in Angus cattle (129).

Shifting our focus to small ruminants, Di Gerlando et al. (13) embarked on an exploration of Sicilian goat breeds. Their study unveiled intriguing findings, as CNVR33, CNVR65, and CNVR7 were found to overlap with genes closely associated with the immune system. This discovery offers a potential explanation for the remarkable resistance of these goat breeds against infectious diseases. In the case of Florida Native sheep, Estrada-Reyes et al. (41) delved into the genetic underpinnings of resistance against gastrointestinal nematodes. Their meticulous investigation revealed that 14 CNVs exhibited overlaps with QTLs associated with gastrointestinal nematode resistance. Moreover, these CNVs demonstrated significant correlations with fecal egg count (FEC), suggesting a potential influence of CNVs on parasite resistance in these sheep. Ma et al. (106) conducted an extensive study encompassing 160 Chinese sheep breeds, leading to the identification of 111 CNV regions. Their functional analysis highlighted an enrichment of CNV regions with genes closely linked to environmental responses. Notably, 17 candidates genes emerged from this analysis, primarily associated with specific diseases, metabolic processes, and development. Particularly intriguing was the observation of lower relative expression levels of KIF2A and PHKG2 in domestic sheep breeds compared to introduce sheep breeds, implying enhanced disease resistance within modern Gansu sheep breeding populations.

In the equine domain, Schurink et al. (115) undertook a thorough investigation into CNVs within Friesian horses. Their research unveiled a staggering 15,041 CNVs identified across 222 individuals. Importantly, this study integrated genome-wide association study (GWAS) leveraging both SNPs and CNVs data. A significant finding was the association between CNV regions situated in the major histocompatibility complex (MHC) region of ECA20 and insect bite hypersensitivity (IBH) in Friesian horses. Notably, approximately half of the horses included in the study were afflicted by this condition. Consistently, Ghosh et al. (55) conducted a comprehensive genomic analysis encompassing healthy horses representing 16 distinct breeds. Their investigation identified 258 CNV regions. Notably, the study extended its inquiry to horses exhibiting sexual developmental impairments, wherein they identified a pure deletion of the AKR1C gene in two male pseudohermaphrodites. This discovery suggests a potential association between this gene deletion and the observed abnormalities.

## 5.4 Environmental adaptability

A comparative examination of CNVs in herbivorous livestock originating from diverse regional breeds has illuminated the potential influence of CNVs on their environmental adaptability. A performed a comprehensive investigation into the differential distribution of CNVs within yak populations hailing from the Tibet and Gansu regions (178). Their study identified seven candidate CNVs, specifically annotating three genes (GRIK4, IFNLR1, and LOC102275985) enriched in five well-established signaling pathways that play pivotal roles in the animals' acclimatization to their environments and various physiological functions. Of particular note is the regulation of physiological processes in hypoxic environments. This research significantly contributes to our understanding of the molecular mechanisms underlying the high-altitude environmental adaptability of yaks. Qaidam cattle, known for their adaptation to cold and low-oxygen environments, underwent genomic analysis by Guo et al. (179). This study identified LDHB and ME1 as potential key genes influencing the cattle's remarkable adaptability to such harsh conditions.

Tibetan sheep, native to high-altitude plateaus, were investigated by Zhu et al. (107), who identified 66 CNVRs associated with their plateau acclimatization. Notably, α-ketoglutarate-dependent dioxygenase alkB homolog 5 (ALKBH5) and nuclear prelamin A recognition factor-like (NARFL) were found to be associated with plateau adaptation within the identified CNVRs. A study performed by Salehian-Dehkordi et al. (112) documented CNVs in 47 sheep breeds. Their research revealed 155 CNVs highly significantly correlated with various environmental parameters, with 35 CNVRs showing significant correlations with solar radiation. Moreover, genes overlapping with CNVs, such as B3GNTL1, UBE2L3, TRAF2, GTF2F1, and IGFALS, were significantly correlated with climatic variables, further emphasizing the role of CNVs in environmental adaptation.

In the equine domain, Wang et al. (114) utilized aCGH to identify CNVs in six horse breeds, uncovering a total of 700 CNVs. These CNVs were classified into 353 CNVRs, and their genetic examination revealed specific genes associated with hemoglobin binding, suggesting a potential influence on horses' adaptability to the challenging plateau environment. The Jinjiang horse, an indigenous breed exclusive to the southeastern coast of China and

adapted to high temperatures and humidity, was the focus of a study by Wang et al. (35). They identified 229 genes that overlapped with CNVRs, with four specific candidate genes (NFKBIA, SOCS4, HSPA1A, and IL6) highlighted due to their strong correlation with cellular thermal acclimatization.

## 6 Conclusions

Common herbivorous livestock, like cattle and sheep, are vital to human daily life due to their significant contributions to meat and dairy production, playing crucial roles in animal husbandry. Recently, China has seen remarkable growth in its equine and donkey industries, driven by advancements in science, technology, and societal progress, establishing themselves as emerging specialties within the livestock sector. Therefore, investigating CNVs within the genomes of these animals holds profound importance. Advancements in microarray and sequencing technologies, along with decreasing sequencing costs, have provided a robust foundation for identifying and studying CNVs. These CNVs, characterized by extended mutant fragment lengths and their substantial impact on genes, represent a formidable genetic resource for exploring genetic variations in livestock and poultry. Extensive research focusing on CNVs in herbivore genomes has unequivocally demonstrated the pivotal role CNVs play in shaping phenotypic diversity, influencing economic traits, enhancing disease resistance, and facilitating environmental adaptation. Researchers worldwide have dedicated their efforts to elucidating the connection between CNVs and phenotypic differences as well as diseases in livestock and poultry. These findings offer compelling support for exploring the potential applications of CNVs as genetic markers in regulation of various productive and disease resistance traits. Consequently, CNVs emerge as a promising avenue for augmenting genetic diversity and expediting molecular breeding strategies in common herbivorous livestock, including cattle, sheep, horses, and donkeys. In conclusion, CNVs represent a valuable and dynamic field of study poised to make a lasting impact on the genetic improvement of herbivorous livestock species, ultimately benefiting both human society and the global livestock industry.

## Author contributions

XL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Writing—original draft, Writing—review & editing. WChe: Methodology, Writing—review & editing, Data curation, Formal analysis. BH: Writing—review & editing, Investigation, Software. XW: Investigation, Software, Writing—review & editing, Data curation, Formal analysis. YP: Data curation, Software, Writing—review & editing, Conceptualization. XZ: Data curation, Software, Writing—review & editing, Investigation. WCha: Software, Writing—review & editing, Formal analysis, Methodology. MK: Methodology, Writing—review & editing, Conceptualization, Project administration, Supervision, Writing—original draft. CW: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Key R&D Program of China (grant number 2022YFD1600103), The Shandong Province Modern Agricultural Technology System Donkey Industrial Innovation Team (grant no. SDAIT-27), Livestock and Poultry Breeding Industry Project of the Ministry of Agriculture and Rural Affairs (grant number 19211162), The National Natural Science Foundation of China (grant no. 31671287), The Open Project of Liaocheng University Animal Husbandry Discipline (grant no. 319312101-14), The Open Project of Shandong Collaborative Innovation Center for Donkey Industry Technology (grant no. 3193308), Research on Donkey Pregnancy Improvement (grant no. K20LC0901), and Liaocheng University scientific research fund (grant no. 318052025).

## 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



## References

1. Hancock JF. *Origins of World Crops and Livestock, World Agriculture Before and After 1492: Legacy of the Columbian Exchange*. Cham: Springer (2022), 5–18.

2. Caldwell CD. Domestication in agricultural systems In CD Caldwell, S Wang, editors. *Introduction to Agroecology*. Singapore: Springer Singapore, 157–67.

3. Andersson L, Purugganan M. Molecular genetic variation of animals and plants under domestication. *Proc Nat Acad Sci.* (2022) 119:e2122150119. doi: 10.1073/pnas.2122150119

4. Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T, et al. Mutation and selection processes regulating short tandem repeats give

rise to genetic and phenotypic diversity across species. *J Evol Biol.* (2023) 36:321–36. doi: 10.1111/jeb.14106

5. Lee YL, Bosse M, Takeda H, Moreira GC, Karim L, Druet T, et al. High-resolution structural variants catalogue in a large-scale whole genome sequenced bovine family cohort data. *BMC Genomics.* (2023) 24:1–7. doi: 10.1186/s12864-023-09259-8

6. Khan MZ, Wang J, Ma Y, Chen T, Ma M, Ullah Q, et al. Genetic polymorphisms in immune-and inflammation-associated genes and their association with bovine mastitis resistance/susceptibility. *Front Immunol.* (2023) 14:1082144. doi: 10.3389/fimmu.2023.1082144

7. Khan IM, Khan A, Liu H, Khan MZ. Genetic markers identification for animal production and disease resistance. *Front Genet.* (2023) 14:1243793. doi: 10.3389/fgene.2023.1243793

8. Wang T, Wang X, Liu Z, Shi X, Ren W, Huang B, et al. Genotypes and haplotype combination of DCAF7 gene sequence variants are associated with number of thoracolumbar vertebrae and carcass traits in Dezhou donkey. *J Appl Anim Res.* (2023) 51:31–9. doi: 10.1080/09712119.2022.2149538

9. Khan MZ, Ma Y, Ma J, Xiao J, Liu Y, Liu S, et al. Association of DGAT1 with cattle, buffalo, goat, and sheep milk and meat production traits. *Frontiers in Veterinary Science.* (2021) 8:712470. doi: 10.3389/fvets.2021.712470

10. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet.* (2013) 14:125–38. doi: 10.1038/nrg3373

11. Begna T. Role and economic importance of crop genetic diversity in food security. *Int J Agric Sci Food Technol.* (2021) 7:164–9. doi: 10.17352/2455-815X.000104

12. Huang B, Khan MZ, Chai W, Ullah Q, Wang C. Exploring genetic markers: mitochondrial dna and genomic screening for biodiversity and production traits in donkeys. *Animals.* (2023) 13:2725. doi: 10.3390/ani13172725

13. Di Gerlando R, Mastrangelo S, Moscarelli A, Tolone M, Sutera AM, Portolano B, et al. Genomic structural diversity in local goats: analysis of copy-number variations. *Animals.* (2020) 10:1040. doi: 10.3390/ani10061040

14. Haque MA, Alam MZ, Iqbal A, Lee YM, Dang CG, Kim JJ, et al. Genome-wide association studies for body conformation traits in korean holstein population. *Animals.* (2023) 13:2964. doi: 10.3390/ani13182964

15. Wang P, Li X, Zhu Y, Wei J, Zhang C, Kong Q, et al. Genome-wide association analysis of milk production, somatic cell score, and body conformation traits in Holstein cows. *Front Vet Sci.* (2022) 9:932034. doi: 10.3389/fvets.2022.932034

16. Yang P, Wang G, Jiang S, Chen M, Zeng J, Pang Q, et al. Comparative analysis of genome-wide copy number variations between Tibetan sheep and White Suffolk sheep. *Anim Biotechnol.* (2023) 34:986–93. doi: 10.1080/10495398.2021.2007937

17. Sölzer N, May K, Yin T, König S. Genomic analyses of claw disorders in Holstein cows: genetic parameters, trait associations, and genome-wide associations considering interactions of SNP and heat stress. *J Dairy Sci.* (2022) 105:8218–36. doi: 10.3168/jds.2022-22087

18. Junior LP, Pinto LF, Cruz VA, Junior GA, Oliveira HR, Chud TS, et al. Genome-wide association and functional genomic analyses for various hoof health traits in North American Holstein cattle. *J Dairy Sci.* (2023).

19. Schneider H, Segelke D, Tetens J, Thaller G, Bennewitz J. A genomic assessment of the correlation between milk production traits and claw and udder health traits in Holstein dairy cattle. *J Dairy Sci.* (2023) 106:1190–205. doi: 10.3168/jds.2022-22312

20. Hay EH, Utsunomiya YT, Xu L, Zhou Y, Neves HH, Carvalheiro R, et al. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. *BMC Genomics.* (2018) 19:1–8. doi: 10.1186/s12864-018-4787-6

21. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* (2015) 526:75–81. doi: 10.1038/nature15394

22. Fadista J, Thomsen B, Holm L, Bendixen C. Copy number variation in the bovine genome. *BMC Genomics.* (2010) 11:284. doi: 10.1186/1471-2164-11-284

23. Stothard P, Choi J, Basu U, Sumner-Thomson JM, Meng Y, Liao X, et al. Whole genome resequencing of black angus and holstein cattle for snp and cnv discovery. *BMC Genomics.* (2011) 12:559. doi: 10.1186/1471-2164-12-559

24. Butty AM, Chud TCS, Cardoso DF, Lopes LSF, Miglior F, Schenkel FS, et al. Genome-wide association study between copy number variants and hoof health traits in holstein dairy cattle. *J Dairy Sci.* (2021) 104:8050–61. doi: 10.3168/jds.2020-19879

25. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet.* (2011) 12:363–76. doi: 10.1038/nrg2958

26. Butty AM, Chud TCS, Miglior F, Schenkel FS, Kommadath A, Krivushin K, et al. High confidence copy number variants identified in holstein dairy cattle from whole genome sequence and genotype array data. *Sci Rep.* (2020) 10:8044. doi: 10.1038/s41598-020-64680-3

27. Braga LG, Chud TC, Watanabe RN, Savegnago RP, Sena TM. Identification of copy number variations in the genome of Dairy Gir cattle. *PLoS ONE.* (2023) 18:e0284085. doi: 10.1371/journal.pone.0284085

28. Liu Y, Mu Y, Wang W, Ahmed Z, Wei X, Lei C, et al. Analysis of genomic copy number variations through whole-genome scan in Chinese Qaidam cattle. *Front Vet Sci.* (2023) 10:1148070. doi: 10.3389/fvets.2023.1148070

29. Xu Z, Wang X, Song X, An Q, Wang D, Zhang Z, et al. Association between the copy number variation of ccser1 gene and growth traits in chinese capra hircus (goat) populations. *Anim Biotechnol.* (2023) 34:1377–83. doi: 10.1080/10495398.2022.2025818

30. Ren H, Wei Z, Li X, Wang Q, Chen H, Lan X, et al. Goat MyoD1: mRNA expression, InDel and CNV detection and their associations with growth traits. *Gene.* (2023) 866:147348. doi: 10.1016/j.gene.2023.147348

31. Yao Z, Zhang S, Wang X, Guo Y, Xin X, Zhang Z, et al. Genetic diversity and signatures of selection in BoHuai goat revealed by whole-genome sequencing. *BMC Genomics.* (2023) 24:116. doi: 10.1186/s12864-023-09204-9

32. Shi H, Li T, Su M, Wang H, Li Q, Lang X, et al. Identification of copy number variation in tibetan sheep using whole genome resequencing reveals evidence of genomic selection. *BMC Genomics.* (2023) 24:555. doi: 10.1186/s12864-023-09672-z

33. Kim Y, Ha S, Seong H, Choi J, Baek H, Yang B, et al. Identification of copy number variations in four horse breed populations in south korea. *Animals.* (2022) 12. doi: 10.3390/ani12243501

34. Laseca N, Molina A, Valera M, Antonini A, Demyda-Peyrás S. Copy number variation (cnv): a new genomic insight in horses. *Animals.* (2022) 12:1435. doi: 10.3390/ani12111435

35. Wang M, Liu Y, Bi X, Ma H, Zeng G, Guo J, et al. Genome-wide detection of copy number variants in chinese indigenous horse breeds and verification of cnv-overlapped genes related to heat adaptation of the jinjiang horse. *Genes.* (2022) 13:603. doi: 10.3390/genes13040603

36. Zandi MB, Salek Ardestani S, Vahedi SM, Mahboudi H, Mahboudi F, Meskoob A, et al. Detection of common copy number of variants underlying selection pressure in middle eastern horse breeds using whole-genome sequence data. *J Heredity.* (2022) 113:421–30. doi: 10.1093/jhered/esac027

37. Choudhury MP, Wang Z, Zhu M, Teng S, Yan J, Cao S, et al. Genome-wide detection of copy number variations associated with miniature features in horses. *Genes.* (2023) 14:1934. doi: 10.3390/genes14101934

38. Castaneda C, Radović L, Felkel S, Juras R, Davis BW, Cothran EG, Wallner B, Raudsepp T. Copy number variation of horse Y chromosome genes in normal equine populations and in horses with abnormal sex development and subfertility: relationship of copy number variations with Y haplogroups. *G3 J.* (2022) 12:jkac278. doi: 10.1093/g3journal/jkac278

39. Sun X, Jiang J, Wang G, Zhou P, Li J, Chen C, et al. Genome-wide association analysis of nine reproduction and morphological traits in three goat breeds from Southern China. *Anim Biosci.* (2023) 36:191. doi: 10.5713/ab.21.0577

40. Wijayanti D, Luo Y, Bai Y, Pan C, Qu L, Guo Z, et al. New insight into copy number variations of goat SMAD2 gene and their associations with litter size and semen quality. *Theriogenology.* (2023) 206:114–22. doi: 10.1016/j.theriogenology.2023.05.012

41. Estrada-Reyes ZM, Ogunade IM, Pech-Cervantes AA, Terrill TH. Copy number variant-based genome wide association study reveals immune-related genes associated with parasite resistance in a heritage sheep breed from the united states. *Parasite Immunol.* (2022) 44:e12943. doi: 10.1111/pim.12943

42. Lovett PS, Duvall EJ, Keggins KM. Bacillus pumilus plasmid ppl10: properties and insertion into bacillus subtilis 168 by transformation. *J Bacteriol.* (1976) 127:817–28. doi: 10.1128/jb.127.2.817-828.1976

43. Buckler CE, Staal SP, Rowe WP, Martin MA. Variation in the number of copies and in the genomic organization of ecotropic murine leukemia virus proviral sequences in sublines of akr mice. *J Virol.* (1982) 43:629–40. doi: 10.1128/jvi.43.2.629-640.1982

44. Hollox EJ, Armour JAL, Barber JCK. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. *The Am J Hum Genetics.* (2003) 73:591–600. doi: 10.1086/378157

45. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc Nat Acad Sci.* (2002) 99:12963–8. doi: 10.1073/pnas.162471999

46. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* (2003) 13:2291–305. doi: 10.1101/gr.1349003

47. Guillaud-Bataille M, Valent A, Soularue P, Perot C, Inda M, Receveur A, et al. Detecting single dna copy number variations in complex genomes using one nanogram of starting dna and bac-array cgh. *Nucleic Acids Res.* (2004) 32:e112. doi: 10.1093/nar/gnh108

48. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science.* (2004) 305:525–8. doi: 10.1126/science.1098918

49. Chen Q, Book M, Fang X, Hoeft A, Stuber F. Screening of copy number polymorphisms in human β-defensin genes using modified real-time quantitative pcr. *J Immunol Methods.* (2006) 308:231–40. doi: 10.1016/j.jim.2005.11.001

50. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature.* (2006) 444:444–54. doi: 10.1038/nature05329

51. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density snp genotyping assay for cattle. *PLoS ONE.* (2009) 4:e5350. doi: 10.1371/journal.pone.0005350

52. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. *Genome Res.* (2010) 20:693–703. doi: 10.1101/gr.105403.110

53. Doan R, Cohen N, Harrington J, Veazy K, Juras R, Cothran G, et al. Identification of copy number variants in horses. *Genome Res.* (2012) 22:899–907. doi: 10.1101/gr.128991.111

54. Liu J, Zhang L, Xu L, Ren H, Lu J, Zhang X, et al. Analysis of copy number variations in the sheep genome using 50k snp beadchip array. *BMC Genomics.* (2013) 14:229. doi: 10.1186/1471-2164-14-229

55. Ghosh S, Qu Z, Das PJ, Fang E, Juras R, Cothran EG, et al. Copy number variation in the horse genome. *PLoS Genet.* (2014) 10:e1004712. doi: 10.1371/journal.pgen.1004712

56. Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W, et al. Reference genome of wild goat (*Capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genomics.* (2015) 16:431. doi: 10.1186/s12864-015-1606-1

57. Kader A, Liu X, Dong K, Song S, Pan J, Yang M, et al. Identification of copy number variations in three chinese horse breeds using 70k single nucleotide polymorphism beadchip array. *Anim Genet.* (2016) 47:560–9. doi: 10.1111/age.12451

58. Han H, Zhao X, Xia X, Chen H, Lei C, Dang R, et al. Copy number variations of five y chromosome genes in donkeys. *Arch Anim Breed.* (2017) 60:391–7. doi: 10.5194/aab-60-391-2017

59. Ma Q, Liu X, Pan J, Ma L, Ma Y, He X, et al. Genome-wide detection of copy number variation in chinese indigenous sheep using an ovine high-density 600 k SNP array. *Sci Rep.* (2017) 7:912. doi: 10.1038/s41598-017-00847-9

60. Liu M, Zhou Y, Rosen BD, Van Tassell CP, Stella A, Tosser-Klopp G, et al. Diversity of copy number variation in the worldwide goat population. *Heredity.* (2019) 122:636–46. doi: 10.1038/s41437-018-0150-6

61. Al Abri MA, Holl HM, Kalla SE, Sutter NB, Brooks SA. Whole genome detection of sequence and structural polymorphism in six diverse horses. *PLoS ONE.* (2020) 15:e230899. doi: 10.1371/journal.pone.0230899

62. Guan D, Martínez A, Castelló A, Landi V, Luigi-Sierra MG, Fernández-Álvarez J, et al. genome-wide analysis of copy number variation in Murciano-Granadina goats. *Genetics Select Evol.* (2020) 52:1–0. doi: 10.1186/s12711-020-00564-4

63. Mei C, Junjvlieke Z, Raza SHA, Wang H, Cheng G, Zhao C, et al. Copy number variation detection in chinese indigenous cattle by whole genome sequencing. *Genomics.* (2020) 112:831–6. doi: 10.1016/j.ygeno.2019.05.023

64. He Y, Hong Q, Zhou D, Wang S, Yang B, Yuan Y, et al. Genome-wide selective detection of mile red-bone goat using next-generation sequencing technology. *Ecol E11.* (2021) 14805–12. doi: 10.1002/ece3.8165

65. Nandolo W, Mészáros G, Wurzinger M, Banda LJ, Gondwe TN, Mulindwa HA, et al. Detection of copy number variants in african goats using whole genome sequence data. *BMC Genomics.* (2021) 22:398. doi: 10.1186/s12864-021-07703-1

66. Salehian-Dehkordi H, Xu Y, Xu S, Li X, Luo L, Liu Y, et al. Genome-wide detection of copy number variations and their association with distinct phenotypes in the world's sheep. *Front Genet.* (2021) 12:670582. doi: 10.3389/fgene.2021.670582

67. Zhou J, Liu L, Reynolds E, Huang X, Garrick D, Shi Y, et al. Discovering copy number variation in dual-purpose xinjiang brown cattle. *Front Genet.* (2022) 12. doi: 10.3389/fgene.2021.747431

68. Taghizadeh S, Gholizadeh M, Rahimi-Mianji G, Moradi MH, Costilla R, Moore S, et al. Genome-wide identification of copy number variation and association with fat deposition in thin and fat-tailed sheep breeds. *Sci Rep.* (2022) 12:8834. doi: 10.1038/s41598-022-12778-1

69. Kumar H, Panigrahi M, Saravanan KA, Rajawat D, Parida S, Bhushan B, et al. Genome-wide detection of copy number variations in tharparkar cattle. *Anim Biotechnol.* (2023) 34:448–55. doi: 10.1080/10495398.2021.1942027

70. Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, et al. An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC genomics.* (2010) 11:1–5. doi: 10.1186/1471-2164-11-639

71. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet.* (2006) 7:85–97. doi: 10.1038/nrg1767

72. Mei-lin J, Zeng-kui L, Qing L, Xiao-juan F, Ming-xing C, Cai-hongW. Research progress on copy number variation of livestock and poultry. *J Agric Biotechnol.* (2019) 10:1840–48.

73. Fernandes AC, da Silva VH, Goes CP, Moreira GC, Godoy TF, Ibelli AM, et al. Genome-wide detection of CNVs and their association with performance traits in broilers. *BMC Genomics.* (2021) 22:1–8. doi: 10.1186/s12864-021-07676-1

74. Pokrovac I, Pezer Ž. Recent advances and current challenges in population genomics of structural variation in animals and plants. *Front Genet.* (2022) 13:1060898. doi: 10.3389/fgene.2022.1060898

75. Hsieh P, Vollger MR, Dang V, Porubsky D, Baker C, Cantsilieris S, et al. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science.* (2019) 366: aax2083. doi: 10.1126/science.aax2083

76. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. *Genome Res.* (2006) 16:949–61. doi: 10.1101/gr.3677206

77. Pös O, Radvanszky J, Buglyó G, Pös Z, Rusnakova D, Nagy B, et al. copy number variation: main characteristics, evolutionary significance, and pathological aspects. *Biomed J.* (2021) 44:548–59.

78. Huang Y, Li Y, Wang X, Yu J, Cai Y, Zheng Z, et al. An atlas of cnv maps in cattle, goat and sheep. *Sci China Life Sci.* (2021) 64:1747–64. doi: 10.1007/s11427-020-1850-x

79. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* (2005) 1:e49. doi: 10.1371/journal.pgen.0010049

80. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics.* (2008) 1:4. doi: 10.1186/1755-8417-1-4

81. Lee JA, Carvalho CMB, Lupski JR. A dna replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell.* (2007) 131:1235–47. doi: 10.1016/j.cell.2007.11.037

82. Pinkel D, Albertson DG. Comparative genomic hybridization. *Annu Rev Genomics Hum Genet.* (2005) 6:331–54. doi: 10.1146/annurev.genom.6.080604.162140

83. Chehbani F, Tomaiuolo P, Picinelli C, Baccarin M, Castronovo P, Scattoni ML, et al. Yield of array-cgh analysis in tunisian children with autism spectrum disorder. *Mol Genet Genom Med.* (2022) 10:e1939. doi: 10.1002/mgg3.1939

84. Kowalczyk K, Smyk M, Bartnik-Głaska M, Plaskota I, Wiśniowiecka-Kowalnik B, Bernaciak J, et al. Application of array comparative genomic hybridization (ACGH) for identification of chromosomal aberrations in the recurrent pregnancy loss. *J Assist Reprod Genet.* (2022) 39:357–67. doi: 10.1007/s10815-022-02400-8

85. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet.* (2004) 66:488–95. doi: 10.1111/j.1399-0004.2004.00322.x

86. De Ravel D, Devriendt TJL, Fryns K, Vermeesch JJR. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). *Eur J Pediatr.* (2007) 166:637–43. doi: 10.1007/s00431-007-0463-6

87. Vissers LELM, de Vries BBA, Osoegawa K, Janssen IM, Feuth T, Choy CO, et al. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *The Am J Hum Genetics.* (2003) 73:1261–70. doi: 10.1086/379977

88. Weiss MM, Snijders AM, Kuipers EJ, Ylstra B, Pinkel D, Meuwissen SG, et al. Determination of amplicon boundaries at 20q13. 2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The J Pathol Soc Great Britain Ireland.* (2003) 200:320–6. doi: 10.1002/path.1359

89. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* (2009) 37:4181–93. doi: 10.1093/nar/gkp552

90. Nowak D, Hofmann W, Koeffler HP. Genome-wide mapping of copy number variations using snp arrays. *Transfus Med Hemother.* (2009) 36:246–51. doi: 10.1159/000225372

91. Walker BA, Morgan GJ. Use of single nucleotide polymorphism–based mapping arrays to detect copy number changes and loss of heterozygosity in multiple myeloma. *Clin Lymphoma Myeloma.* (2006) 7:186–92. doi: 10.3816/CLM.2006.n.057

92. Metzger J, Philipp U, Lopes MS, da Camara Machado A, Felicetti M, Silvestrelli M, et al. Analysis of copy number variants by three detection algorithms and their association with body size in horses. *BMC Genomics.* (2013) 14:1–5. doi: 10.1186/1471-2164-14-487

93. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC Bioinf.* (2013) 14:S1. doi: 10.1186/1471-2105-14-S11-S1

94. Liu GE, Bickhart DM. Copy number variation in the cattle genome. *Funct Integr Genomics.* (2012) 12:609–24. doi: 10.1007/s10142-012-0289-9

95. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* (2012) 2012:251364. doi: 10.1155/2012/251364

96. Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. *J Invest Dermatol.* (2013) 133:1–4. doi: 10.1038/jid.2013.248

97. Chen D, Xu Y, Ding C, Wang Y, Fu Y, Cai B, et al. The inconsistency between two major aneuploidy-screening platforms—single-nucleotide polymorphism array and next-generation sequencing—in the detection of embryo mosaicism. *BMC Genomics.* (2022) 23:62. doi: 10.1186/s12864-022-08294-1

98. Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY, et al. Identification of copy number variations and common deletion

polymorphisms in cattle. *BMC Genomics.* (2010) 11:1–10. doi: 10.1186/1471-2164-11-232

99. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* (2012) 22:778–90. doi: 10.1101/gr.133967.111

100. Zhang L, Jia S, Yang M, Xu Y, Li C, Sun J, et al. Detection of copy number variations and their effects in chinese bulls. *BMC Genomics.* (2014) 15:480. doi: 10.1186/1471-2164-15-480

101. Da Silva D, Giachetto JM, Da Silva PF, Cintra LO, Paiva LC, Yamagishi SR, et al. Genome-wide copy number variation (CNV) detection in nelore cattle reveals highly frequent variants in genome regions harboring QTLS affecting production traits. *BMC Genomics.* (2016) 17:454. doi: 10.1186/s12864-016-2752-9

102. Zhou Y, Connor EE, Wiggans GR, Lu Y, Tempelman RJ, Schroeder SG, et al. Genome-wide copy number variant analysis reveals variants associated with 10 diverse production traits in holstein cattle. *BMC Genomics.* (2018) 19:314. doi: 10.1186/s12864-018-4699-5

103. Kooverjee BB, Soma P, van der Nest MA, Scholtz MM, Neser FW. Copy number variation discovery in South African Nguni-Sired and Bonsmara-Sired crossbred cattle. *Animals.* (2023) 13:2513. doi: 10.3390/ani13152513

104. Sun T, Pei S, Liu Y, Hanif Q, Xu H, Chen N, et al. Whole genome sequencing of simmental cattle for snp and cnv discovery. *BMC Genomics.* (2023) 24:179. doi: 10.1186/s12864-023-09248-x

105. Liu S, Kang X, Catacchio CR, Liu M, Fang L, Schroeder SG, et al. Computational detection and experimental validation of segmental duplications and associated copy number variations in water buffalo (*Bubalus bubalis*). *Funct Integr Genomics.* (2019) 19:409–19. doi: 10.1007/s10142-019-00657-4

106. Ma Y, Zhang Q, Lu Z, Zhao X, Zhang Y. Analysis of copy number variations by snp50 beadchip array in chinese sheep. *Genomics.* (2015) 106:295–300. doi: 10.1016/j.ygeno.2015.08.001

107. Zhu C, Fan H, Yuan Z, Hu S, Ma X, Xuan J, et al. Genome-wide detection of cnvs in chinese indigenous sheep with different types of tails using ovine high-density 600k snp arrays. *Sci Rep.* (2016) 6:27822. doi: 10.1038/srep27822

108. Yang L, Xu L, Zhou Y, Liu M, Wang L, Kijas JW, et al. Diversity of copy number variation in a worldwide population of sheep. *Genomics.* (2018) 110:143–8. doi: 10.1016/j.ygeno.2017.09.005

109. Kang X, Li M, Liu M, Liu S, Pan MG, Wiggans GR, et al. Copy number variation analysis reveals variants associated with milk production traits in dairy goats. *Genomics.* (2020) 112:4934–7. doi: 10.1016/j.ygeno.2020.09.007

110. Yuan C, Lu Z, Guo T, Yue Y, Wang X, Wang T, et al. A global analysis of cnvs in chinese indigenous fine-wool sheep populations using whole-genome resequencing. *BMC Genomics.* (2021) 22:78. doi: 10.1186/s12864-021-07387-7

111. Igoshin AV, Deniskova TE, Yurchenko AA, Yudin NS, Dotsev AV, Selionova MI, et al. Copy number variants in genomes of local sheep breeds from russia. *Anim Genet.* (2022) 53:119–32. doi: 10.1111/age.13163

112. Salehian-Dehkordi H, Huang J, Pirany N, Mehrban H, Lv X, Sun W, et al. Genomic landscape of copy number variations and their associations with climatic variables in the worldandrsquo;s sheep. *Genes.* (2023) 14:1256. doi: 10.3390/genes14061256

113. Dupuis MC, Zhang Z, Durkin K, Charlier C, Lekeux P, Georges M, et al. Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy. *Anim Genet.* (2013) 44:206–8. doi: 10.1111/j.1365-2052.2012.02373.x

114. Wang W, Wang S, Hou C, Xing Y, Cao J, Wu K, et al. Genome-wide detection of copy number variations among diverse horse breeds by array cgh. *PLoS ONE.* (2014) 9:e86860. doi: 10.1371/journal.pone.0086860

115. Schurink A, Da Silva D, Velie VH, Dibbits BD, Crooijmans BW. Copy number variations in friesian horses and genetic risk factors for insect bite hypersensitivity. *BMC Genet.* (2018) 19:49. doi: 10.1186/s12863-018-0657-0

116. Solé M, Ablondi M, Binzer-Panchal A, Velie BD, Hollfelder N, Buys N, et al. Inter-and intra-breed genome-wide copy number diversity in a large cohort of European equine breeds. *BMC Genomics.* (2019) 20:1–2. doi: 10.1186/s12864-019-6141-z

117. Wang C, Li H, Guo Y, Huang J, Sun Y, Min J, et al. Donkey genomes provide new insights into domestication and selection for coat color. *Nat Commun.* (2020) 11:6014. doi: 10.1038/s41467-020-19813-7

118. Huang Y, Xu T, Li L, Li Z. Jun-yong W, Hong C, Zhao LC. Research progress of copy number variation of cattle and sheep china cattle. *Science.* (2018) 44:55–60. doi: 10.3969/j.issn.1001-9111.2018.04.015

119. Keel BN, Lindholm-Perry AK, Snelling WM. Evolutionary and functional features of copy number variation in the cattle genome. *Front Genet.* (2016) 7:207. doi: 10.3389/fgene.2016.00207

120. Corbi-Botto CM, Morales-Durand H, Zappa ME, Sadaba SA, Peral-García P, Giovambattista G, et al. Genomic structural diversity in criollo argentino horses: analysis of copy number variations. *Gene.* (2019) 695:26–31. doi: 10.1016/j.gene.2018.12.067

121. Nozawa M, Kawahara Y, Nei M. Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Nat Acad Sci.* (2007) 104:20421–6. doi: 10.1073/pnas.0709956104

122. Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ, et al. Extensive copy-number variation of the human olfactory receptor gene family. *The Am J Hum Genetics.* (2008) 83:228–42. doi: 10.1016/j.ajhg.2008.07.005

123. Veerappa AM, Vishweswaraiah S, Lingaiah K, Murthy M, Manjegowda DS, Nayaka R, et al. Unravelling the complexity of human olfactory receptor repertoire by copy number analysis across population using high resolution arrays. *PLoS ONE.* (2013) 8:e66843. doi: 10.1371/journal.pone.0066843

124. Spehr M, Munger SD. Olfactory receptors: g protein-coupled receptors and beyond. *J Neurochem.* (2009) 109:1570–83. doi: 10.1111/j.1471-4159.2009.06085.x

125. Palouzier-Paulignan B, Lacroix MC, Aime P, Baly C, Caillol M, Congar P, et al. Olfaction under metabolic influences. *Chem Senses.* (2012) 37:769–97. doi: 10.1093/chemse/bjs059

126. Ghosh S, Das PJ, McQueen CM, Gerber V, Swiderski CE, Lavoie JP, et al. Analysis of genomic copy number variation in equine recurrent airway obstruction (heaves). *Anim Genet.* (2016) 47:334–44. doi: 10.1111/age.12426

127. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet.* (2009) 10:451–81. doi: 10.1146/annurev.genom.9.081307.164217

128. Shi T, Xu Y, Yang M, Huang Y, Lan X, Lei C, et al. Copy number variations at lepr gene locus associated with gene expression and phenotypic traits in chinese cattle. *Anim Sci J.* (2016) 87:336–43. doi: 10.1111/asj.12531

129. Xu L, Hou Y, Bickhart DM, Song J, Van Tassell CP, Sonstegard TS, et al. genome-wide survey reveals a deletion polymorphism associated with resistance to gastrointestinal nematodes in Angus cattle. *Funct Integ Genomics.* (2014) 14:333–9. doi: 10.1007/s10142-014-0371-6

130. Liu M, Li B, Huang Y, Yang M, Lan X, Lei C, et al. Copy number variation of bovine MAPK10 modulates the transcriptional activity and affects growth traits. *Livest Sci.* (2016) 194:44–50. doi: 10.1016/j.livsci.2016.09.014

131. Yang M, Lv J, Zhang L, Li M, Zhou Y, Lan X, et al. Association study and expression analysis of cyp4a11 gene copy number variation in chinese cattle. *Sci Rep.* (2017) 7:46599. doi: 10.1038/srep46599

132. Zhang GM, Zheng L, He H, Song CC, Zhang ZJ, Cao XK, et al. Associations of GBP2 gene copy number variations with growth traits and transcriptional expression in Chinese cattle. *Gene.* (2018) 647:101–6. doi: 10.1016/j.gene.2018.01.004

133. Ma YL, Wen YF, Cao XK, Cheng J, Huang YZ, Ma Y, et al. Copy number variation (CNV) in the IGF1R gene across four cattle breeds and its association with economic traits. *Archives Anim Breeding.* (2019) 62:171–9. doi: 10.5194/aab-62-171-2019

134. Xu JW, Zheng L, Li LJ, Yao YF, Hua H, Yang SZ, et al. Novel copy number variation of the KLF3 gene is associated with growth traits in beef cattle. *Gene.* (2019) 680:99–104. doi: 10.1016/j.gene.2018.08.040

135. Zheng L, Xu JW Li JC, Wang DH, An QM, Xu LN, et al. Distribution and association study in copy number variation of KCNJ12 gene across four Chinese cattle populations. *Gene.* (2019) 689:90–6. doi: 10.1016/j.gene.2018.12.019

136. Yang P, Zhang Z, Xu J, Qu K, Lyv S, Wang X, et al. The association of the copy number variation of the MLLT10 gene with growth traits of Chinese cattle. *Animals.* (2020) 10:250. doi: 10.3390/ani10020250

137. Hao D, Wang X, Thomsen B, Kadarmideen HN, Wang X, Lan X, et al. Copy number variations and expression levels of guanylate-binding protein 6 gene associated with growth traits of Chinese cattle. *Animals.* (2020) 10:566. doi: 10.3390/ani10040566

138. Tang J, Shen X, Yang Y, Yang H, Qi A, Yang S, et al. Two different copy number variations of the clcn2 gene in chinese cattle and their association with growth traits. *Animals.* (2021) 12:41. doi: 10.3390/ani12010041

139. Hu L, Yu J, Huang R, Yang P, Zhang Z, Chai Y, et al. Copy number variation of the CCDC39 gene is associated with growth traits in Chinese cattle. *Vet Med Sci.* (2022) 8:917–24. doi: 10.1002/vms3.712

140. Yang P, Cai C, Niu M, Liu X, Wang H, Liang H, et al. of copy number variation of PLA2G2A gene to growth traits in Chinese cattle. *Gene.* (2022) 809:146014. doi: 10.1016/j.gene.2021.146014

141. Yang H, Yue B, Yang Y, Tang J, Yang S, Qi A, et al. Distribution of copy number variation in syt11 gene and its association with growth conformation traits in Chinese cattle. *Biology.* (2022) 11:223. doi: 10.3390/biology11020223

142. Yao Z, Li J, Zhang Z, Chai Y, Liu X, Li J, et al. The relationship between MFN1 copy number variation and growth traits of beef cattle. *Gene.* (2022) 811:146071. doi: 10.1016/j.gene.2021.146071

143. Li X, Ding X, Liu L, Yang P, Yao Z, Lei C, et al. Copy number variation of bovine DYNC1I2 gene is associated with body conformation traits in chinese beef cattle. *Gene.* (2022) 810:146060. doi: 10.1016/j.gene.2021.146060

144. Zhang J, Zhang Z, Liu X, Chai Y, Yang P, Li J, et al. number variation of WBP1L gene revealed its association with growth traits across chinese cattle populations. *J Agric Sci.* (2022) 160:528–34. doi: 10.1017/S0021859622000387

145. Chen Y, Peng W, Zhang Z, Liu X, Yang P, Fu C, et al. The relationship between MUC19 copy number variation and growth traits of Chinese cattle. *Gene.* (2023) 851:147010. doi: 10.1016/j.gene.2022.147010

146. Liang J, Liu X, Yang P, Yao Z, Qu K, Wang H, et al. Copy number variation of gal3st1 gene is associated with growth traits of chinese cattle. *Anim Biotechnol.* (2023) 34:672–8. doi: 10.1080/10495398.2021.1996385

147. Liu X, Yang P, Sun H, Zhang Z, Cai C, Xu J, et al. analysis of VAMP7 gene reveals variation associated with growth traits in Chinese cattle. *Anim Biotechnol.* (2023) 34:1095–101. doi: 10.1080/10495398.2021.2011741

148. Song X, Li X, Liu X, Zhang Z, Ding X, Chai Y, et al. Copy number variation of the ZNF679 gene in cattle and its association analysis with growth traits. *Anim Biotechnol.* (2023) 21:1–7. doi: 10.1080/10495398.2023.2185628

149. Jiang R, Cheng J, Cao XK, Ma YL, Chaogetu B, Huang YZ, et al. Copy number variation of the SHE gene in sheep and its association with economic traits. *Animals.* (2019) 9:531. doi: 10.3390/ani9080531

150. Wang X, Cao X, Wen Y, Ma Y, Elnour IE, Huang Y, et al. Associations of ORMDL1 gene copy number variations with growth traits in four Chinese sheep breeds. *Arch Anim Breeding.* (2019) 62:571–8. doi: 10.5194/aab-62-571-2019

151. Cheng J, Jiang R, Yang Y, Cao X, Huang Y, Lan X, et al. Association analysis of KMT2D copy number variation as a positional candidate for growth traits. *Gene.* (2020) 753:144799. doi: 10.1016/j.gene.2020.144799

152. Toremurat Z, Ibrahim EE, Huang YZ, Lan X, Pi L, Chaogetu B, et al. Copy number variations of TOP2B gene are associated with growth traits in Chinese sheep breeds. *Anim Biotechnol.* (2022) 33:85–9. doi: 10.1080/10495398.2020.1773490

153. Yang Z, Cao X, Ma Y, Cheng J, Song C, Jiang R, et al. Novel copy number variation of the bag4 gene is associated with growth traits in three chinese sheep populations. *Anim Biotechnol.* (2021) 32:461–9. doi: 10.1080/10495398.2020.1719124

154. Wen Y, Wang E, Wang X, Qing S, Chaogetu B, Wang C, et al. Copy number variations of LRRFIP1 gene and the relationship with growth traits in four Chinese sheep. *Anim Biotechnol.* (2023) 34:3008–15. doi: 10.1080/10495398.2022.2126981

155. Wang X, Wang Y, Cao X, Huang Y, Li P, Lan X, et al. Copy number variations of the kat6a gene are associated with body measurements of chinese sheep breeds. *Anim Biotechnol.* (2023) 34:947–54. doi: 10.1080/10495398.2021.2005616

156. Feng Z, Li X, Cheng J, Jiang R, Huang R, Wang D, et al. Copy number variation of the pigy gene in sheep and its association analysis with growth traits. *Animals.* (2020) 10:688. doi: 10.3390/ani10040688

157. Shi SY Li LJ, Zhang ZJ, Wang EY, Wang J, Xu JW, et al. Copy number variation of MYLK4 gene and its growth traits of *Capra hircus* (goat). *Anim Biotechnol.* (2020) 31:532–7. doi: 10.1080/10495398.2019.1635137

158. Li L, Yang P, Shi S, Zhang Z, Shi Q, Xu J, et al. Association analysis to copy number variation (CNV) of Opn4 gene with growth traits of goats. *Animals.* (2020) 10:441. doi: 10.3390/ani10030441

159. Liu M, Woodward-Greene J, Kang X, Pan MG, Rosen B, Van Tassell CP, et al. Genome-wide CNV analysis revealed variants associated with growth traits in African indigenous goats. *Genomics.* (2020) 112:1477–80. doi: 10.1016/j.ygeno.2019.08.018

160. Xu Z, Wang X, Zhang Z, An Q, Wen Y, Wang D, et al. Copy number variation of CADM2 gene revealed its association with growth traits across Chinese *Capra hircus* (goat) populations. *Gene.* (2020) 741:144519. doi: 10.1016/j.gene.2020.144519

161. Wang Q, Wei Z, Zhu H, Pan C, Akhatayeva Z, Song X, et al. Goat pleomorphic adenoma gene 1 (PLAG1): mRNA expression, CNV detection and associations with growth traits. *Animals.* (2023) 13:2023. doi: 10.3390/ani13122023

162. Song X, Bai Y, Yuan R, Zhu H, Lan X, Qu L, et al. InDel and CNV within the AKAP13 gene revealing strong associations with growth traits in goat. *Animals.* (2023) 13:2746. doi: 10.3390/ani13172746

163. Wang Q, Song X, Bi Y, Zhu H, Wu X, Guo Z, et al. Detection distribution of CNVs of SNX29 in three goat breeds and their associations with growth traits. *Front Vet Sci.* (2023) 10:1132833. doi: 10.3389/fvets.2023.1132833

164. Pei SW, Qin F, Li WH, Li FD, Yue XP. Copy number variation of ZNF280AY across 21 cattle breeds and its association with the reproductive traits of Holstein and Simmental bulls. *J Dairy Sci.* (2019) 102:7226–36. doi: 10.3168/jds.2018-16063

165. Yue XP, Dechow C, Chang TC, DeJarnette JM, Marshall CE, Lei CZ, et al. Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics.* (2014) 15:1–2. doi: 10.1186/1471-2164-15-113

166. Pei S, Xu H, Wang L, Li F, Li W, Yue X. Copy number variation of ZNF280BY across eight sheep breeds and its association with testicular size of Hu sheep. *J Anim Sci.* (2022) 100:skac232. doi: 10.1093/jas/skac232

167. Wang Q, Bi Y, Wang Z, Zhu H, Liu M, Wu X, et al. Goat SNX29: mRNA expression, InDel and CNV detection, and their associations with litter size. *Front Vet Sci.* (2022) 9:981315. doi: 10.3389/fvets.2022.981315

168. Bai Y, Zhang T, Liu N, Wang C, Guo Z, Pan C, et al. Investigation of copy number variations (CNVs) of the goat PPP3CA gene and their effect on litter size and semen quality. *Animals.* (2022) 12:445. doi: 10.3390/ani12040445

169. Zhang R, Wang J, Zhang T, Zhai H, Shen W. Copy-number variation in goat genome sequence: a comparative analysis of the different litter size trait groups. *Gene.* (2019) 696:40–6. doi: 10.1016/j.gene.2019.02.027

170. Janečka JE, Davis BW, Ghosh S, Paria N, Das PJ, Orlando L, et al. Horse y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nat Commun.* (2018) 9:2945. doi: 10.1038/s41467-018-05290-6

171. Trigo BB, Utsunomiya ATH, Fortunato AAAD, Milanesi M, Torrecilha RBP, Lamb H, et al. Variants at the asip locus contribute to coat color darkening in nellore cattle. *Genet Sel E.* (2021) 53:40. doi: 10.1186/s12711-021-00633-2

172. Liang D, Zhao P, Si J, Fang L, Pairo-Castineira E, Hu X, et al. Genomic analysis revealed a convergent evolution of line-1 in coat color: a case study in water buffaloes (bubalus bubalis). *Mol Biol E38.* (2021) 1122–36. doi: 10.1093/molbev/msaa279

173. Fontanesi L, Beretti F, Riggio V, Gómez González E, Dall Olio S, Davoli R, et al. Copy number variation and missense mutations of the agouti signaling protein (asip) gene in goat breeds with different coat colors. *Cytogenet Genome Res.* (2010) 126:333–47. doi: 10.1159/000268089

174. Menzi F, Keller I, Reber I, Beck J, Brenig B, Schütz E, et al. Genomic amplification of the caprine ednra locus might lead to a dose dependent loss of pigmentation. *Sci Rep.* (2016) 6:28438. doi: 10.1038/srep28438

175. Guo J, Sun X, Mao A, Liu H, Zhan S, Li L, et al. 13 42-kb tandem duplication at the ASIP locus is strongly associated with the depigmentation phenotype of non-classic Swiss markings in goats. *BMC Genomics.* (2022) 23:437. doi: 10.1186/s12864-022-08672-9

176. Rosengren Pielberg G, Golovko A, Sundström E, Curik I, Lennartsson J, Seltenhammer MH, et al. A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nature Genet.* (2008) 40:1004–9. doi: 10.1038/ng.185

177. Drögemüller C, Distl O, Leeb T. Partial deletion of the bovine ed1 gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Res.* (2001) 11:1699–705. doi: 10.1101/gr.182501

178. Guang-Xin E, Yang BG, Zhu YB, Duang XH, Basang WD, Luo XL, et al. Genome-wide selective sweep analysis of the high-altitude adaptability of yaks by using the copy number variant. *Biotech.* (2020) 10:1–6. doi: 10.1007/s13205-020-02254-w

179. Guo S, Wu X, Pei J, Wang X, Bao P, Xiong L, et al. Genome-wide cnv analysis reveals variants associated with high-altitude adaptation and meat traits in qaidam cattle. *Electron J Biotechnol.* (2021) 54:8–16. doi: 10.1016/j.ejbt.2021.07.006

180. Huang Y, Shi QT, Shi S, Yang P, Zhang Z, Lyu S, et al. Association between copy number variation of serpina3-1 gene and growth traits in chinese cattle. *Anim Biotechnol.* (2023) 34:1524–31. doi: 10.1080/10495398.2022.2038183

181. Durkin K, Coppieters W, Drögemüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature.* (2012) 482:81–4. doi: 10.1038/nature10757

182. Durán Aguilar M, Román Ponce SI, Ruiz López FJ, González Padilla E, Vásquez Peláez CG, Bagnato A, et al. Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. *J Anim Breeding Genetics.* (2017) 134:49–59. doi: 10.1111/jbg.12238

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 January 2024
DOI 10.3389/fgene.2023.1348329



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Mohammad Hossein Banabazi,
Swedish University of Agricultural Sciences, Sweden
Mohammed Ali Al Abri,
Sultan Qaboos University, Oman

*CORRESPONDENCE
Wujun Liu,
lwj_ws@163.com
Xiangmin Yan,
yanxiangmin1014@sohu.com

RECEIVED 02 December 2023
ACCEPTED 28 December 2023
PUBLISHED 12 January 2024

CITATION
Wang X, Ma Z, Gao L, Yuan L, Ye Z, Cui F, Guo X, Liu W and Yan X (2024), Genome-wide survey reveals the genetic background of Xinjiang Brown cattle in China.
*Front. Genet.* 14:1348329.
doi: 10.3389/fgene.2023.1348329



# Genome-wide survey reveals the genetic background of Xinjiang Brown cattle in China

Xiao Wang[1,2], Zhen Ma[3], Liang Gao[2], Lixin Yuan[3], Zhibing Ye[3], Fanrong Cui[3], Xiaoping Guo[4], Wujun Liu[1]* and Xiangmin Yan[3]*

[1]College of Animal Science, Xinjiang Agricultural University, Urumqi, China, [2]Yili Vocational and Technical College, Yili, China, [3]Institute of Animal Science, Xinjiang Academy of Animal Science, Urumqi, China, [4]Yili Kazakh Autonomous Prefecture General Animal Husbandry Station, Yili, China

**Introduction:** Xinjiang Brown cattle are a famous dual-purpose (dairy-beef) cultivated breed in China that occupy a pivotal position within the cattle breeding industry in Xinjiang, China. However, little information is available on the genetic background of this breed. To fill this research gap, we conducted a whole-genome screen using specific-locus amplified fragment sequencing to examine the genetic structure and diversity of 130 Xinjiang Brown cattle-grazing type (XBG, traditional type) cattle.

**Methods:** A subsequent joint analysis incorporating two ancestral breeds, specifically 19 Brown Swiss (BS) foreign and nine Kazakh (KZ) Chinese cattle, as well as 20 Xinjiang Brown cattle-housing type (XBH) cattle, was used to explore the genetic background of the Xinjiang Brown cattle.

**Results:** The results showed that, after nearly a century of crossbreeding, XBG cattle formed a single population with a stable genetic performance. The genetic structure, genetic diversity, and selection signature analysis of the two ancestral types showed highly different results compared to that of XBH cattle. Local ancestry inference showed that the average proportions of XGB cattle within the BS and KZ cattle lineages were 37.22% and 62.78%, respectively, whereas the average proportions of XBH cattle within the BS and KZ cattle lineages were 95.14% and 4.86%, respectively. Thus, XGB cattle are more representative of all Xinjiang Brown cattle, in line with their breeding history, which involves crossbreeding. Two complementary approaches, fixation index and mean nucleotide diversity, were used to detect selection signals in the four aforementioned cattle breeds. Finally, the analysis of 26 candidate genes in Xinjiang Brown cattle revealed significant enrichment in 19 Gene Ontology terms, and seven candidate genes were enriched in three pathways related to disease resistance (*CDH4*, *SIRPB1*, and *SIRPα*) and the endocrine system (*ADCY5*, *ABCC8*, *KCNJ11*, and *KCNMA1*). Finally, development of the core SNPs in XBG cattle yielded 8,379 loci.

**Conclusion:** The results of this study detail the evolutionary process of crossbreeding in Xinjiang Brown cattle and provide guidance for selecting and breeding new strains of this species.

KEYWORDS

**Xinjiang Brown cattle, specific-locus amplified fragment-sequencing, genetic structure, genetic diversity, candidate genes, ancestry proportion**

# 1 Introduction

Xinjiang Brown cattle is a dual-purpose (dairy-beef) cultivated breed bred independently in China. These cattle exhibit strong adaptability, superior grazing, and production performance within the extreme arid, cold, and barren environment of northern Xinjiang (Zhou et al., 2019) (Figures 1A, B). These qualities make them an important cattle breed for local breeders, farmers, and herders. By 2022, the stock of purebred Xinjiang Brown cattle amounted to approximately 1,169 thousand heads. Specifically, 27 thousand heads were in feedlots, whereas 1,142 thousand of them were in pastures. In addition, there were specifically 760 thousand and 14 thousand fertile cows and bulls, respectively. Overall, this stock accounted for approximately 1/5th of all types of cattle stock in Xinjiang. Xinjiang Brown cattle were selectively bred during the early 20th century, in which the Kazakh (KZ) cow was the female parent and underwent three-stage hybridization with the Brown Swiss (BS) bull or the Kostroma or Ala-Tau bulls, two breeds of Brown Swiss cattle origin (Yurchenko et al., 2017). As a result, the Ministry of Agriculture of China certified the Xinjiang Brown as a novel dual-purpose (dairy-beef) cattle breed in 1983 (Zhang et al., 2022). In the following 40 years, this breed has also transitioned into the expansion phase of selection and improvement (1987–2006), followed by a breeding phase of specialized strains (2007 to present). With the introduction of frozen semen from BS bulls primarily from Germany, the United States of America, and Canada, Xinjiang Brown cattle have been further improved and various new strains have been bred to meet the demands of the market and socio-economic development. Nearly 97% of Xinjiang Brown cattle are reared under semi-herding and grazing conditions and are referred to in this study as the Xinjiang Brown cattle-grazing type (XBG) cattle (Figure 1C). The XBG population is large and has been less affected by frozen sperm from BS cattle since 2007, which has helped the genetic preservation of the original makeup of Xinjiang Brown cattle. However, the genetic background of XBG cattle characteristics remains mostly unknown at the genomic level and the extent to which their breed contribution is influenced by Chinese KZ or foreign BS breeds is uncertain.

In recent years, developments in genomic technology have facilitated genomic analyses that have enabled access to individual DNA information via whole-genome sequencing (WGS) (Yin et al., 2019) and genotype-by-sequencing (GBS) (Elshire et al., 2011). Research on Xinjiang Brown cattle has also entered the omics era, as evidenced by genome-wide association studies for milk production and reproductive traits (Zhou et al., 2019), genome-wide identification and analysis of long non-coding RNAs in the longest dorsal muscle tissue (Yan et al., 2021), and genomic selection for milk production traits (Zhang et al., 2022). A WGS analysis was also performed to study the genetic evolution of Xinjiang Brown cattle, which more comprehensively revealed their genetic background, genetic diversity, and adaptive mechanisms (Chen et al., 2022a); however, the samples used in this study were not genuinely representative, as all 50 samples were collected from Xinjiang Brown cattle-housing type (XBH) cattle (Figure 1D) at the Urumqi breeding farm, XBH accounts for less than 3 percent of the

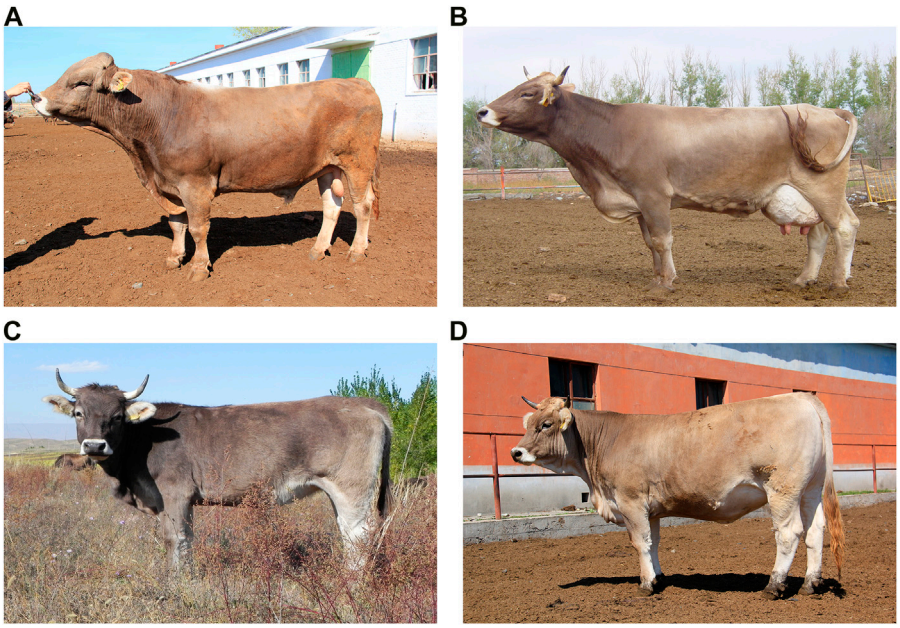



FIGURE 1
Xinjiang Brown cattle photos. **(A)** Xinjiang Brown cattle breeding bull in Yili-Tacheng region. **(B)** Xinjiang Brown cattle breeding cow in Yili-Nilka County. **(C)** Xinjiang Brown cattle-grazing type (XBG) cow in Yili-Tacheng region. **(D)** Xinjiang Brown cattle-housing type (XBH) cow in the Urumqi breeding farm.

total number of Xinjiang Brown cattle. Additionally, the XBH contains a disproportionately high percentage of BS genetic lineage and has low genetic diversity and closely related individuals as confirmed by trial results. Hence, gathering an extensive and more inclusive group of XBG cattle to obtain fresh test outcomes from is imperative. Additionally, the findings of the previous samples from XBH cattle should be blended with new results to reflect the genetic background of Xinjiang Brown cattle more precisely and impartially.

To this end, in this study, we used specific-locus amplified fragment (SLAF) sequencing (SLAF-seq) (Zhou and Pan, 2023) techniques to obtain individual DNA information from 130 XBG cattle. The published data for KZ, BS, and XBH cattle from 48 WGS analyses were downloaded from NCBI and combined with the SLAF-seq data of 129 XBG cattle for joint analysis; the proven feasibility of merging two types of data has previously confirmed the findings of an exploration of olive diversity in plants (Friel et al., 2021) and an assessment of the adaptation of Nigerian cattle in animals (Mauki et al., 2022). This study focused on analyzing the genetic structure, genetic diversity, and selection signatures of the XBG population and aimed to establish a molecular basis that could assist in the conservation, scientific introduction, and selection of breeding resources for Xinjiang Brown cattle.

# 2 Materials and methods

## 2.1 Sample collection

The study team visited the primary production area of XBG in the Yili Tacheng region in 2021. We collected blood samples from 130 XBG cattle on a large private ranch, despite difficulties in sampling resulting from grazing conditions. Whole-blood samples (10 mL) from 130 XBG cattle (all females) were divided into four groups depending on the color of the cow's coat (group A = 40, B = 76, C = 10, and D = 4). The coat color of group A was a normal brown, that of B was dark brown, with fawn for C, and light brown for D. Genomic DNA extractions were performed using the phenol–chloroform method (Sambrook and Russell, 2006) at the Xinjiang Academy of Animal Science. The DNA purity of the extracted samples was determined via quantification using a Thermo Scientific™ NanoDrop 2000 spectrophotometer (Thermo Scientific, United States). Furthermore, the quality of the DNA extracts was assessed by subjecting them to electrophoresis on a 2% agarose gel against a 2 kilobase (kb) DNA ladder marker. The 130 samples were then subjected to sequencing on the SLAF platform.

## 2.2 SLAF library construction and sequencing

The SLAF library was created as described previously with minor adjustments (Aerts et al., 2013). To ensure the anticipated SLAF output, we avoided repetitive SLAFs and selected a relatively uniform distribution of restriction fragments in the genome. Next, we conducted a simulated restriction enzyme digestion on the existing *B. taurus* genome (UMD 3.1) (Zimin et al., 2009). The genomic DNA from each sample was digested using a combination of RsaI and HaeIII restriction enzymes. This was followed by adding a single nucleotide (A) overhang to the 3′ end of the SLAF tags. To ensure ligation of dual-index sequencing adapters to A-tailed tags, we carried out restriction-ligation reactions using T4 DNA ligase (New England Biolabs). Subsequently, DNA amplification was performed using PCR and the resulting products were purified using the E. Z.N.A.H Cycle Pure Kit (Omega). The purified samples were combined and incubated with two specified restriction enzymes, RsaI and HaeIII. After being ligated with ATP and a Solexa adapter at the paired-end, the reaction was purified using a Quick Spin column (Qiagen, Venlo, Netherlands) and segregated on a 2% agarose gel. Fragments between 450 and 480 bp were extracted using a Gel Extraction Kit (Tiangen, China). These SLAFs were then subjected to PCR for barcode addition. The amplified DNA samples underwent re-purification before being prepared for 150-base paired-end sequencing using an Illumina NovaSeq6000 sequencing platform (Illumina, San Diego, CA, United States) at Biomarker Technologies Corporation (Beijing, China).

## 2.3 WGS library construction and sequencing

Forty-eight publicly available WGS genome datasets were acquired from previous studies (Chen et al., 2018; Chen et al., 2022a). Raw sequencing data of XBH ($n$ = 20), KZ ($n$ = 9), and BS ($n$ = 19) cattle are available at NCBI BioProject ID: PRJNA833533, PRJNA379859, and PRJEB28191, respectively (Supplementary Table S1).

## 2.4 Processing, mapping, filtering, and single-nucleotide polymorphism calling of SLAF reads

All indexed sequenced reads with clear information were clustered based on sequence similarity. Similarity clustering was used to group the sequenced reads from the same locus (Jones et al., 2013) and aligned to a reference genome (UMD 3.1) (Zimin et al., 2009) using BWA v0.7.17 software (Li and Durbin, 2009a). Single-nucleotide polymorphism (SNP) calling was performed using GATK v3.8 software (McKenna et al., 2010) and SAMtools v1.3.1 software (Li et al., 2009b). In addition, the filtered high-quality SNPs were used to annotate the SNP detection results using SnpEff software (Cingolani et al., 2012), which can provide the region of the genome where the variant locus occurs (intergenic region, gene region, or CDS region, etc.) and the effect of the variant (synonymous non-synonymous mutation, etc.).

## 2.5 Population genetic structure and genetic diversity analysis of SLAF datasets

The high-confidence SNPs produced via the above procedures were used to infer the genetic structure of the 130 XBG cattle. We constructed an unrooted phylogenetic tree using the neighbor-

joining method with the Kimura 2-parameter/p-distance model in MEGA-CC (MEGAX) software (Kumar et al., 2018), with 1,000 bootstrap replicates. Principal components analysis (PCA) was performed using the smartPCA module of EIGENSOFT v7.2.0 software, using the default parameters (Price et al., 2006). Estimation of the genetic relationships from SNPs using one of the five main functions of GCTA v1.91.7 software (Yang et al., 2011) to estimate the kinship between two individuals of a natural population is possible. The population structure within the 130 XBG cattle was inferred using ADMIXTURE v1.3.0 software (Alexander et al., 2009), with K values (the putative number of populations) ranging from 1 to 10. The optimal number of clusters K (best taxa) was determined as the one with the minimum cross-validation error rate. The Q matrix for each K value within stacked assignment bar plots was generated using the R package "Pophelpers" (Francis, 2017). Pi values were calculated using VCFtools v0.1.16 software based on the high-confidence filtered SNPs and a 100 kb window with a step size of 10 kb for each sub-population (Danecek et al., 2011).

## 2.6 Processing, mapping, filtering, SNP calling, and data merging of SLAF and published WGS reads

Raw paired-end reads of the 129 XBG SLAF-seq genome datasets and 48 publicly available WGS genome datasets were mapped to the *B. taurus* reference genome (ARS-UCD 1.2) (Rosen et al., 2020) using BWA v0.7.17 software (parameters: mem -t 4 -k 32 -M) (Li and Durbin, 2009a). SNP calling was performed using both GATK v3.8 software (McKenna et al., 2010) and SAMtools v1.3.1 software (Li et al., 2009b) (WGS parameter: rmdup; SLAF parameter: sort) analyses, and a locus was defined as a SNP if it was simultaneously called from these two packages. The "mpileup" command was used to identify SNPs with the parameters "-q 1 -C 50 -S -D -m 2 -F 0.002 -u". Then, to exclude SNP calling errors caused by incorrect mapping, only high-quality SNPs [coverage depth ≥ 4, RMS mapping quality ≥ 20, minor allele frequency (MAF) ≥ 0.01, miss ≤ 0.3] were retained for subsequent analysis. BCFtools v1.7 software was used to merge overlapping genomic regions between the SLAF and WGS datasets (Li et al., 2009b).

## 2.7 Population genetic structure and genetic diversity analysis of the merged SLAF and published WGS datasets

An individual-based neighbor-joining tree was constructed for the 177 evaluated cattle based on the p-distance, with one outgroup (i.e., *Bos mutus*) (Supplementary Table S1), using TreeBest v1.9.2 software (Vilella et al., 2009) with 1,000 bootstrap replicates. We further conducted PCA to evaluate genetic structures using GCTA v1.91.7 software (Yang et al., 2011). The population genetic structure was examined using ADMIXTURE v1.3.0 software (Alexander et al., 2009), and the number of assumed genetic clusters K ranged from 2 to 8, with 10,000 iterations for each run. In addition, RFMix v2.03 software (Maples et al., 2013) was used for Local-Ancestry inferencing. The indicators of observed heterozygosity (Ho), expected heterozygosity (He), polymorphism information content (Pic), and Nei's genetic diversity index (Nei) analysis were counted using the Stacks v1.45 populations program (Catchen et al., 2013).

## 2.8 Genome-wide selective sweep test, Gene Ontology annotation, and Kyoto Encyclopedia of genes and genomes functional enrichment

We used VCFtools v0.1.16 software (Danecek et al., 2011) to calculate Pi and the fixation index (FST) with the window size set to 50 k and sliding window 10 k. The genome-wide distribution of FST values and mean Pi ($\theta\pi$) ratios for the indicated group pairs were computed to discover genome-wide selection sweeps linked to cattle adaptability. The FST values were Z-transformed as Z (FST) = (FST—μFST)/σFST, where σFST represents the FST standard deviation and μFST is the FST mean. Log2-transformed $\theta\pi$ ratios were obtained. The empirical percentiles of Z (FST) and log2 ($\theta\pi$ ratio) in each window were then computed and ranked. Under strong selective sweeps, we looked at the windows that simultaneously had the top 5% Z (FST) and log2 ($\pi$ ratio) values as potential outliers. Every outlier window had a corresponding SNP and gene assigned to it (the selection method of the candidate window was modified according to the actual situation).

Protein-coding genes were functionally annotated through the utilization of BLASTp (with an E-value of <10–5) (Gish and States, 1993), with protein sequence databases sourced from SwissProt. Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis of differentially expressed genes was implemented via the GOseq R package (Young et al., 2010), in which gene length bias was corrected. GO terms with a corrected *p*-value less than 0.05 were considered significantly enriched by differentially expressed genes. We used KOBAS (Mao et al., 2005) to test the statistical enrichment of differentially expressed genes in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Pathways with a q-value < 0.05 were considered significantly enriched.

## 2.9 Core SNP development

The SNP markers of the 129 XBG cattle were screened for core markers. The first step of this screening process was depth filtering, during which SNP loci with a depth of at least ×4 were retained in this project and the low depth loci were filtered out. The second step was completeness filtering for markers with poor genotypic integrity coverage; markers with a genotypic coverage of at least 70% of all individuals in the population were retained (Tian et al., 2015; Gao et al., 2016). The third step was MAF filtering, which filtered out loci with MAF values below 0.01. The fourth step was to filter by Pic value; the Pic value was calculated as Pic = $1-\sum fi^r$, where fi is the gene frequency of locus I. Loci with Pic values less than 0.4 were filtered out. Finally, loci located in the intergenic region based on the functional annotation of the SNP loci were filtered out, retaining only loci located in the upstream and downstream areas of the gene and within the gene SNP loci (Zhang et al., 2013; Graebner et al., 2015).

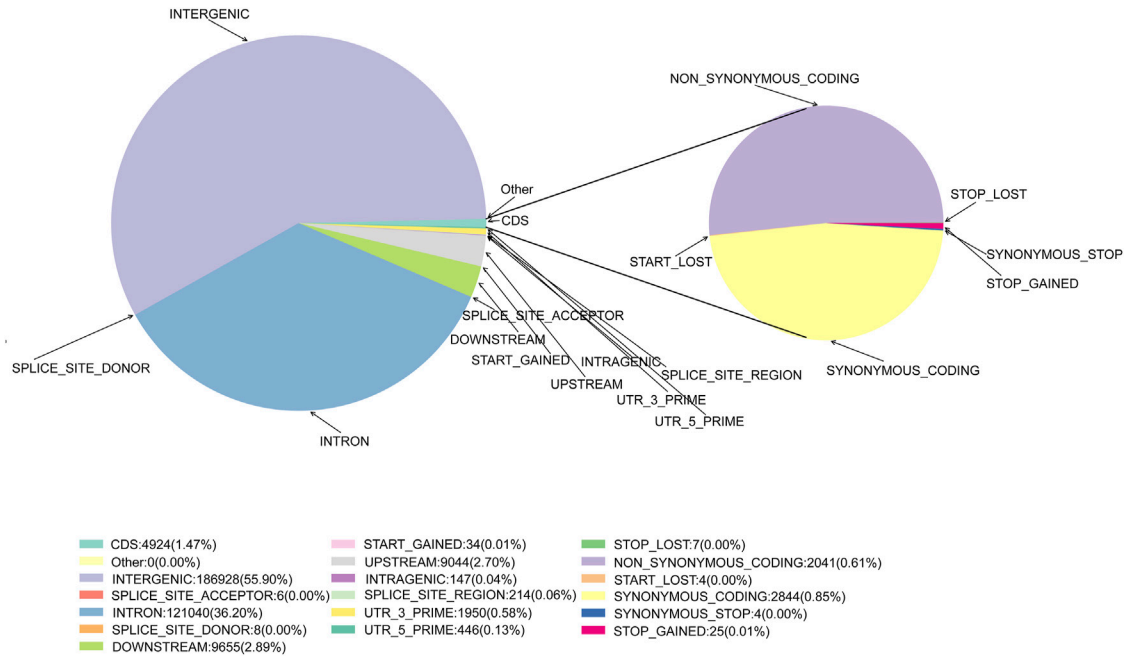



FIGURE 2
The position of the SNPs in gene structures and annotations of the SNPs in the exons.

# 3 Results

## 3.1 SLAF tag development

Following quality control and data filtering, a total of 130 libraries were constructed for the cattle genomic DNA samples. These libraries yielded 657.79 Mb clean reads from SLAF-seq. The total number of reads obtained from each sample ranged from 1,661,853 to 10,777,757. The GC content was 42.70%–49.16%, with an average of 46.02%, and the 3 M quality score of the sequenced bases was 87.97%–94.21%, with an average of 92.85% (Supplementary Table S2). A total of 984,712 SLAF tags were designed, with an average sequencing depth of 12.36X (Supplementary Figure S2; Supplementary Table S3). There were 28,208 polymorphic SLAF tags with a total of 4,839,549 SNP markers. SNP integrity was 22.52%–45.90%, and heterozygosity was 4.86%–11.04% (Supplementary Figures S2, S3; Supplementary Tables S4, S5). Further analysis of SNP distribution in the genome revealed that 55.90% were in intergenic regions, with 0.04% in intragenic regions, 36.20% in introns, and 2.70% and 2.89% in the 5 kb regions upstream and downstream, respectively (Figure 2, Supplementary Table S6). These potential functional SNPs provide valuable genetic resources for exploring the genetic background of the Xinjiang Brown cattle.

## 3.2 Population genetic structure and genetic diversity of XBG

According to phylogenetic analyses, the four-colored groups of XBG exhibit heterogeneity (i.e., no apparent clustering) (Figure 3A). Hree-dimensional PCA showed that the first, second, and third axis captured 1.19%, 1.14%, and 1.03% of the overall variance, respectively, and that the four-colored groups of XBG cattle are mostly clustered, a trend that is congruent with the PCA results (Figure 3B). The genetic structures of the four-colored groups of XBG were analyzed across different clusters (K from 1 to 10) using the cross-validation error rate (Figure 3C; Supplementary Figure S4). The cross-validation error rate was lowest when K = 1 (Supplementary Figure S5), confirming the lack of genetic differentiation among the four-colored groups; this finding is congruent with that obtained via the PCA and phylogenetic analyses. The heat map of kinship values is uniformly blue, indicative of the 130 XBG individuals being distantly related to each other (Figure 3D).

The observed heterozygosity (Ho), expected heterozygosity (He), polymorphism information content (Pic), and minor allele frequency (MAF) were calculated for the four XBG color groups based on SNP loci used to characterize the genetic diversity of the different cattle populations (Pan et al., 2016). The average Ho, He, Pic, and MAF of the four XBG groups were 0.2364, 0.2718, 0.2216, and 0.2241, respectively (Table 1).

## 3.3 Merged SLAF and published WGS datasets for population genetics analyses

The 48 publicly available genomes were downloaded from the NCBI database. Following quality control, the resultant high-quality (Q20 ≥ 90%, Q30 ≥ 85%) data (clean data) amounted to 165,561 gigabytes (Gb). Furthermore, the GC distribution of the resultant sequence data was normal, and none of the 48 samples were contaminated, allowing for their use in subsequent analyses (Supplementary Table S7). The reads were then aligned to the

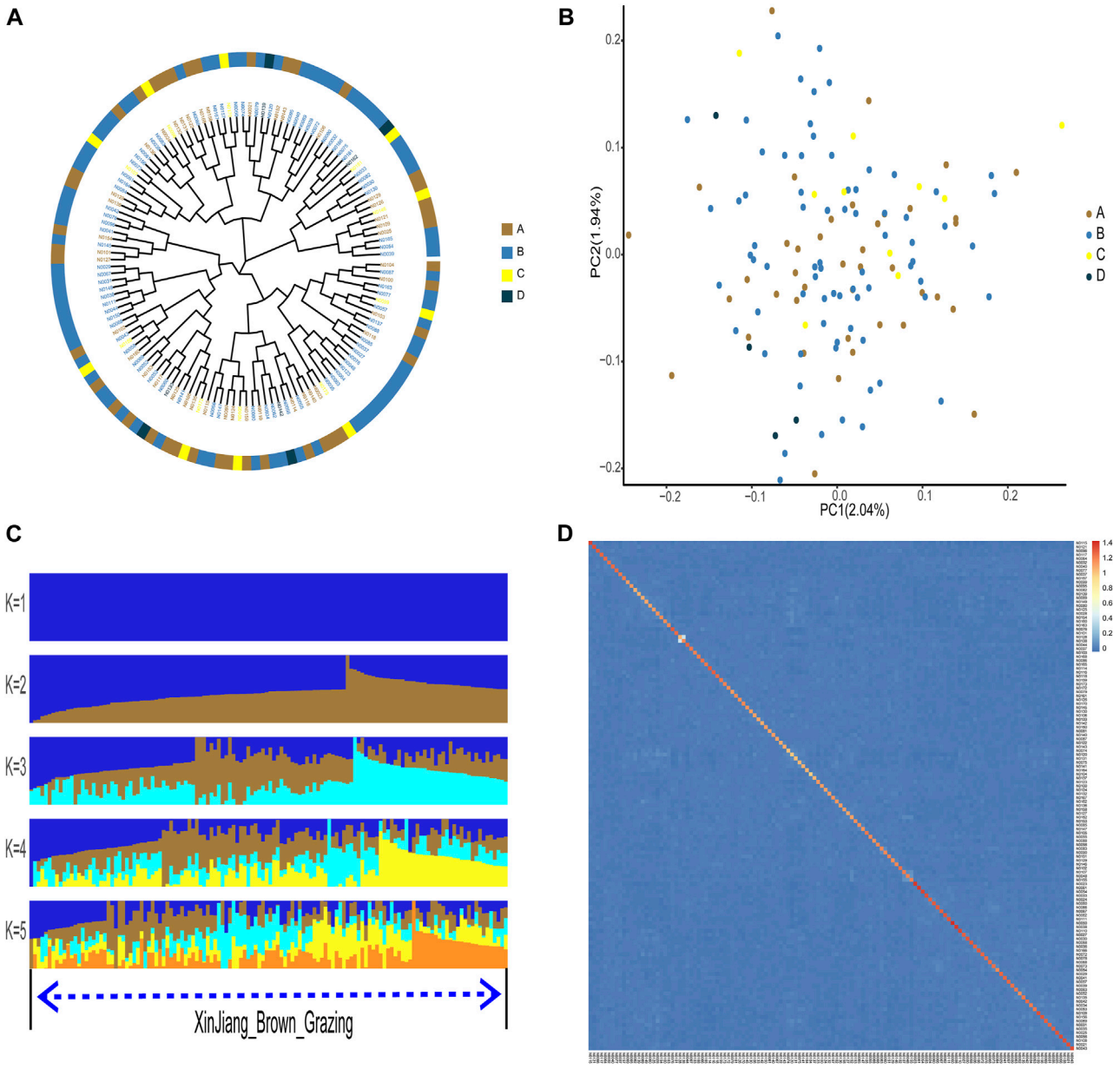



FIGURE 3
Phylogenetic relationship, population structure, and individual kinship of the Xinjiang Brown cattle-grazing type (XBG) cattle among the four cow coat color groups in this study. **(A)** Neighbor-joining phylogenetic tree constructed from single-nucleotide variant data among the four color groups. **(B)** Principal component analysis for the first two PCs of the 130 XBG cattle. **(C)** ADMIXTURE analysis with five presumed ancestral groups to one presumed ancestral group (K = from 1 to 5). **(D)** Heat map of kinship values for the 130 individual XBG cattle.

**TABLE 1** Genetic diversity of the grazing type of Xinjiang Brown cattle among the four cow coat color groups.

| Group | Ho | He | Pic | MAF |
|---|---|---|---|---|
| A | 0.2371 | 0.2872 | 0.2355 | 0.2047 |
| B | 0.2270 | 0.2878 | 0.2364 | 0.2040 |
| C | 0.2297 | 0.2691 | 0.2191 | 0.2228 |
| D | 0.2519 | 0.2431 | 0.1956 | 0.2651 |
| Average | 0.2364 | 0.2718 | 0.2216 | 0.2241 |

Abbreviations: He, expected heterozygosity; Ho, observed heterozygosity; MAF, minor allele frequency; Pic, polymorphism information content.

10.3389/fgene.2023.1348329

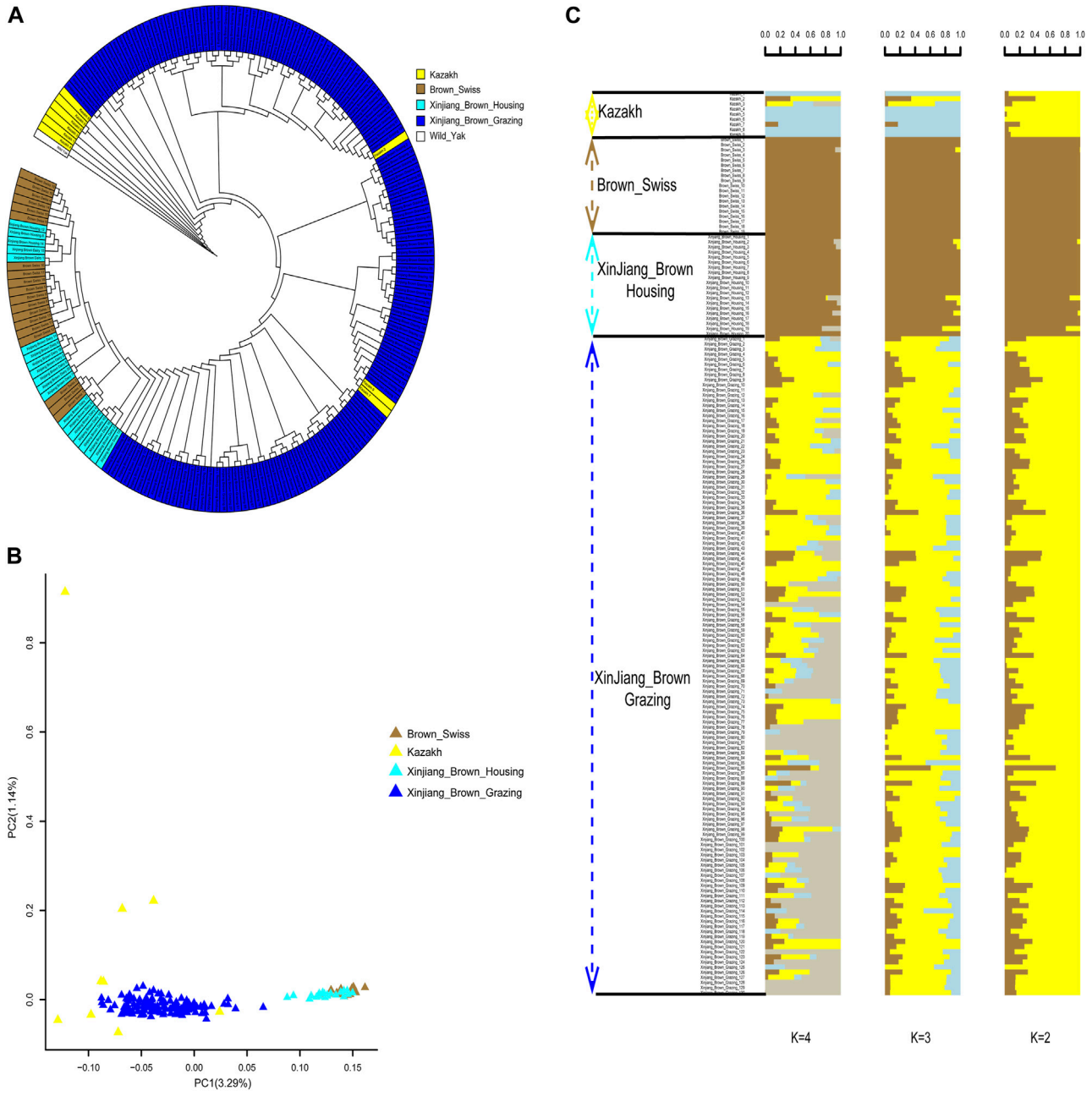



FIGURE 4
Phylogenetic relationship and population structure of the Xinjiang Brown cattle-grazing type (XBG) cattle and the other three breeds evaluated in this study. **(A)** Neighbor-joining phylogenetic tree constructed from single-nucleotide variant data among four populations. **(B)** Principal component analysis for the first two PCs of the 177 studied cattle. **(C)** ADMIXTURE analysis with four presumed ancestral groups to two presumed ancestral groups (K = from 2 to 4).

taurine reference genome (*B. taurus* ARS-UCD 1.2) (Rosen et al., 2020) and merged with 129 SLAF genomes for genotyping. A total of 22,708,388 raw SNPs were detected in 177 samples; the data were then filtered using Dp4-miss0.3-maf0.01 for the WGS and SLAF samples respectively, resulting in 17,201,439 SNPs from the former and 167,936 from the latter. The two SNP datasets were then merged to yield 104,163 common loci, which were once again filtered using the Dp4-miss0.3-maf0.01 condition. Ultimately, 99,933 high-quality SNP loci were obtained for subsequent analysis.

## 3.4 Population genetic structures of Xinjiang Brown cattle and their ancestor species

Phylogenetic trees yielded clear genetic structure, with XBG shown to be most closely related to KZ. In contrast, XBH and BS were found to be genetically indistinguishable. The closest relatives of the XBG were a mixture of KZ and BS cattle (Figure 4A). PCA also mainly distinguished the two clusters along PC1 (i.e., with two clusters being brought forth: XBG & KZ and XBH & BS). KZ were further separated from XBG populations along PC2, but the separation was incomplete (Figure 4B).

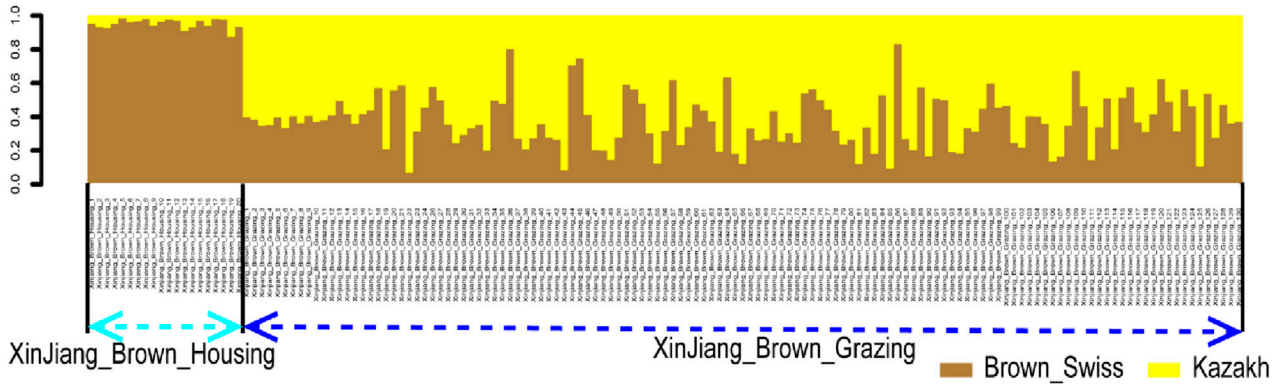



FIGURE 5
Ancestry proportion of the 129 XBG and 20 XBH individuals inferred using RFMix, as based on the reference panels of Kazakh and Brown Swiss cattle.

Furthermore, ADMIXTURE analysis assuming ancestral number K from 2 to 8 was performed (Figure 4C; Supplementary Figure S6), and we found that the cross-validation error rate was lowest when K = 2 (Supplementary Figure S7), allowing for an inference of the genetic structure and admixture specifically for the two cattle clusters.

## 3.5 Global and local ancestry proportion

We determined the global ancestry proportions of the cattle using ADMIXTURE analysis for XBG, XBH, and their parents based on the breeding procedure. The findings revealed that when K = 2, ancestor 1 and ancestor 2 were used, the average proportions in XBG were 19.05% and 80.95%, in XBH they were 97.96% and 2.04%, in BS they were 99.97% and 0.03%, and in KZ they were 10.21% and 89.78% (Figure 4C).

Additionally, a rapid and reliable forward-backward technique in RFMix was used to perform local ancestry inference in the context of XBG and XBH. Thus, in the context of BS and KZ lineages, the average proportions of XBG were 37.22% and 62.78%, respectively. In addition, these two bloodlines had average XBH proportions of 95.14% and 4.86%, respectively (Figure 5).

## 3.6 Population genetic diversity of Xinjiang Brown cattle and their ancestor species

The KZ cattle had the highest Ho, He, Nei's genetic diversity index (Nei), and Pic values, indicating that this population had the highest genetic diversity. The XBH population had the lowest Ho, He, Nei, and Pic values, indicating that this population had the lowest genetic diversity (Table 2).

## 3.7 Identification and functional annotation of candidate genes using a selective sweep test

Pairwise comparisons of the four cattle breeds were calculated using selective sweep approaches. The calculated FST (Supplementary Figure S9) and θπ (Supplementary Figure S8) values were used to detect the selective sweep signals. The intersection of selected parts of the FST and θπ is presented in Figure 6; Supplementary Figure S10. In total, 846 genes were identified in candidate intervals within the SNP corresponding to the gene (Supplementary Table S8). Gene annotation and pathway analysis showed that enrichment of GO terms was significantly concentrated between XBH and XBG cattle, with the three novel genes selected for XBG cattle being enriched in the 18 GO terms with Q < 0.05 (adjusted *p*-value) and concentrated between BS and XBH cattle, with the 23 genes selected for XBH being enriched in the 1 GO term (adjusted *p*-value) (Figure 7; Supplementary Table S9). KEGG analysis showed that nine genes were enriched in five pathways with Q < 0.05 (adjusted *p*-value). Most of the highly enriched pathways included cell adhesion molecules (bta04514), osteoclast differentiation (bta04380), and the insulin secretion pathway (bta04911) (Table 3; Supplementary Table S10).

## 3.8 Core SNP marker information statistics

Regarding the core loci of the resulting 8379 SNPs, further analysis of SNP distribution in the genome revealed that 88.24% and 1.86% were located in intronic and exonic regions, respectively, and 3.58% and 0.33% were located in intronic and exonic non-coding RNA, respectively; 2.03% and 2.49% were observed in the 5 kb regions upstream and downstream of the transcription start site, respectively, with 0.30% and 1.05% in the 5′ and 3′ UTRs, respectively, and 0.02% in the splice junctions (Supplementary Table S11). Specific information has been provided for these SNP core loci (Supplementary Table S12).

# 4 Discussion

The GBS method is simple, quick, extremely specific, highly reproducible, and may reach important regions of the genome that are inaccessible to sequence capture approaches. GBS libraries based on reducing genome complexity with restriction enzymes (REs), making it feasible for species with high genetic diversity and large

**TABLE 2** Genetic diversity in the four studied cattle breeds.

| Population (breed) | Ho | He | Nei | Pic |
|---|---|---|---|---|
| Kazakh | 0.2496 | 0.2368 | 0.2368 | 0.2512 |
| Xinjiang_Brown_Grazing | 0.2066 | 0.2204 | 0.2204 | 0.2214 |
| Brown_Swiss | 0.1919 | 0.1810 | 0.1810 | 0.1861 |
| Xinjiang_Brown_Housing | 0.1812 | 0.1745 | 0.1754 | 0.1831 |

Abbreviations: He, expected heterozygosity; Ho, observed heterozygosity; Nei, Nei's diversity index; Pic, polymorphism information content.

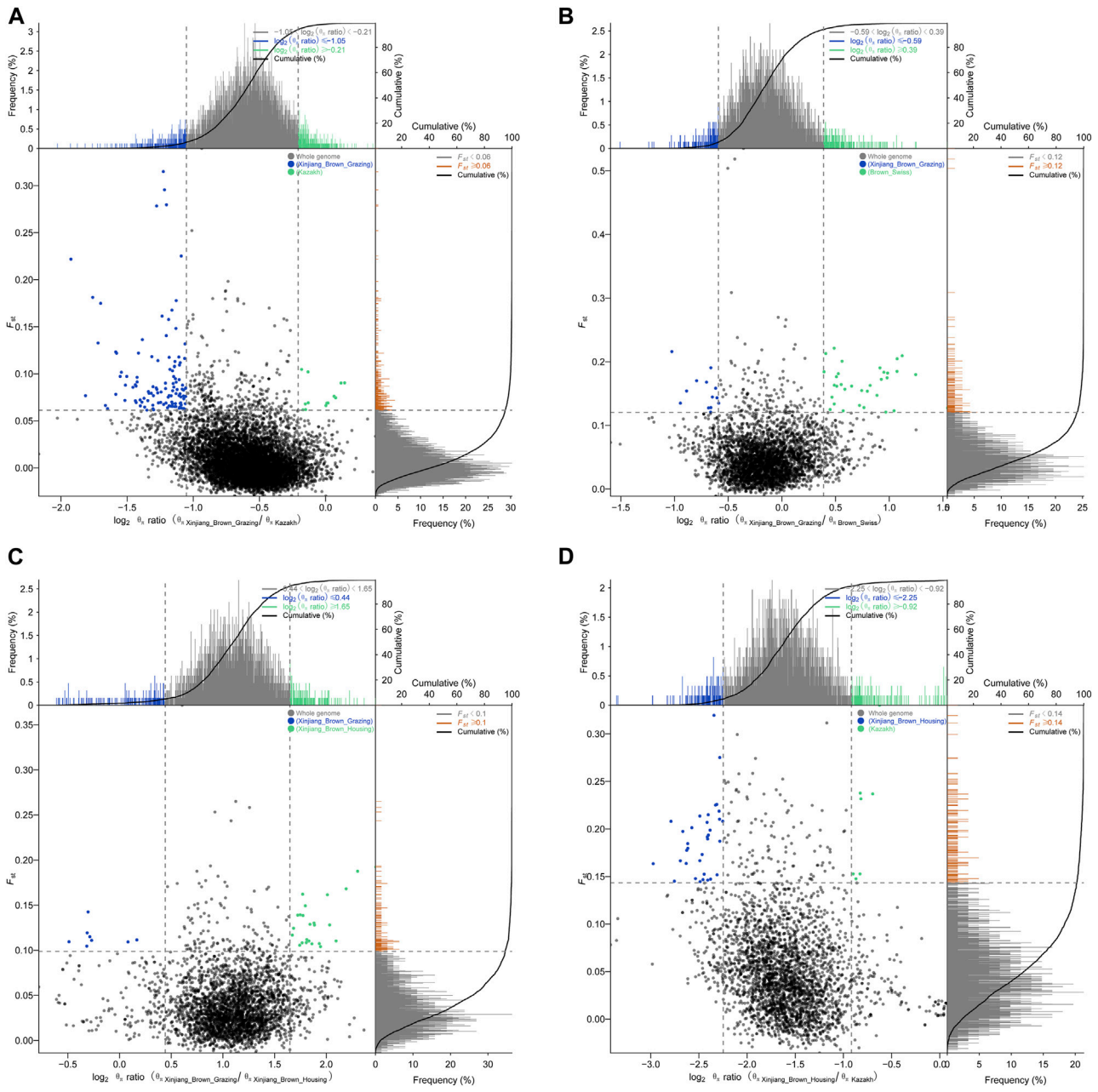



FIGURE 6
Schematic diagrams of selection signals. **(A)** Xinjiang Brown cattle-grazing type vs. Kazakh cattle. **(B)** Xinjiang Brown cattle-grazing type vs. Brown Swiss cattle. **(C)** Xinjiang Brown cattle-grazing type vs. Xinjiang Brown cattle-housing type. **(D)** Xinjiang Brown cattle-housing type vs. Kazakh cattle.

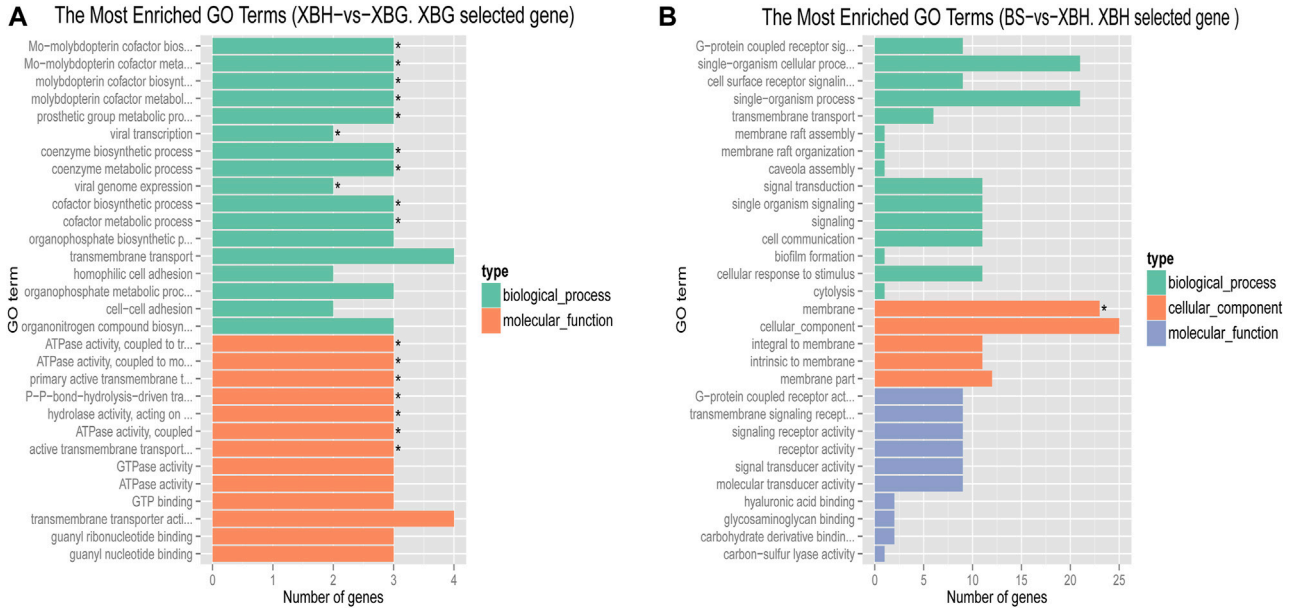



FIGURE 7
Bar chart of the distribution of candidate genes in different GO categories. **(A)** The most enriched GO terms of a Xinjiang Brown cattle-grazing type (XBG) selected gene in Xinjiang Brown cattle-housing type (XGH) vs. XBG. **(B)** The most enriched GO terms of an XBH selected gene in Brown Swiss (BS) cattle vs. XBG.

**TABLE 3** Pathways significantly enriched between breeds and candidate genes within those pathways.

| XBH vs. XBG, XBG selected gene | | | | |
|---|---|---|---|---|
| Pathway ID | Pathway name | Pathway candidate gene entry ID/KO | *p*-value[a] | Q-value[b] |
| bta04514 | Cell adhesion molecules | bta:525796/KO6797 | 0.006352 | 0.006352 |
| KZ vs. XBH, XBH selected gene | | | | |
| Pathway ID | Pathway name | Pathway candidate gene entry ID/KO | *p*-value[a] | Q-value[b] |
| bta04380 | Osteoclast differentiation | bta:536097/KO6551, bta:529990/KO6551 | 0.000289 | 0.018238 |
| bta04911 | Insulin secretion | bta:282573/KO4946, bta:505740/KO8045 | 0.000486 | 0.018238 |
| | | bta:538996/KO5032, bta:532060/KO5004 | | |
| XBH vs. KZ, KZ selected gene | | | | |
| Pathway ID | Pathway name | Pathway candidate gene entry ID/KO | *p*-value[a] | Q-value[b] |
| bta04340 | Hedgehog signaling pathway | bta:522467/KO0312 | 0.002477 | 0.036980 |
| bta05217 | Basal cell carcinoma | bta:511308/KO0312 | 0.002958 | 0.036980 |

[a]Statistical test method: hypergeometric test/Fisher's exact test.
[b]FDR, correction method: Benjamini and Hochberg.
Abbreviations: KZ, Kazakh; XBG, Xinjiang Brown cattle-grazing type; XBH, Xinjiang Brown cattle-housing type.

genomes (Elshire et al., 2011). SLAF-seq techniques share similarities with GBS in their principles and methods in that they are both part of reduced-representation genome sequencing. SLAF-seq methods, however, are not the same as GBS methods in a few respects. For example, one tag is identified by SLAF-seq roughly every 10 K, the uniform distribution of SLAF tags guarantees that significant chromosomal segments are not overlooked, and SLAF-seq is a cost-effective method since it avoids repeating sequences (Chen et al., 2022b). While both SLAF-seq and WGS can identify SNPs, they greatly differ in terms of cost-effectiveness and missing data. Although SLAF-seq is speedy and inexpensive, its DNA fragmentation stage causes a significant amount of missed data. WGS may be better for small sample sizes since it can yield more information; however, most labs find that a large-scale genotyping project quickly becomes unnecessarily expensive. In contrast, SLAF-seq is an affordable substitute that can yield the same outcomes and

detect the same genetic links for a significantly smaller price (Friel et al., 2021). As we gathered a sizable sample size of 130 XBG cattle, the SLAF-seq approach was selected to make the distinction between XBG and XBH cattle clear.

The results show that the four groups of 130 XBG we collected were not clustered in the phylogenetic tree nor the PCA and ADMIXTURE analyses, which is a strong indication that the XBG formed a single population with stable genetic performance after nearly a century of cross breeding. The results also show that the 130 individual cattle are distantly related, indicating that the sample size was both large and representative. These findings represent a strong basis for obtaining reliable genetic analysis results through future analyses.

The ancestors of the Xinjiang Brown cattle, BS and KZ cattle, have been clearly delineated in the global classification of cattle breeds. The world's cattle breeds are divided into five main categories: European, Eurasian, and East Asian taurines and Chinese and Indian indicines. The BS and KZ breeds are both Eurasian taurines, and their offspring, Xinjiang Brown cattle, also fall within this category (Chen et al., 2018). Our research focused on clarifying the similarities and differences between the two different types of Xinjiang Brown cattle and the genetic relationships of the ancestral species, so no other cattle breeds were included. In utilizing new methods of data merging from the past 2 years, we successfully combined the power of SLAF-seq and WGS methods to exploit a wider range of data sources for research purposes in a cost-effective manner. The results show that XBG and XBH cattle are two different cattle that can be divided into separate groups as per the phylogenetic tree as well as PCA and ADMIXTURE analyses. Furthermore, regarding XBG cattle, the percentage of KZ ancestry was greater than that of BS ancestry; for XBH cattle, the opposite was noted. Software unsuited for unequal sampling (Puechmaille, 2016) and close ancestral populations (Liu et al., 2013; Uren et al., 2020) may have produced less rigorous results, even though the global ancestry proportion was inferred using classical ADMIXTURE. Thus, the local ancestry proportions were estimated by utilizing the fitting software RFMix. One of the two ancestors of the KZ accounts for 10.21% of the inferred ADMIXTURE ancestry. This is most likely because KZ and BS cattle are of the same Eurasian taurine breed or that KZ cattle were inadvertently contaminated with BS genes during the breeding process of the Xinjiang Brown cattle. It was deduced by ADMIXTURE and RFMix that over 95% of the blood in XBH cattle comes from one of the two ancestors and BS. This outcome is in line with the findings of earlier WGS research (Chen et al., 2022a) and indicates that the inferred genetic structure results remained accurate after merging with SLAF-seq data, despite the high loss of SNP loci. However, XBG cattle have a BS genetic lineage of no higher than 40% and a more than 60% KZ genetic ancestry. Having a greater KZ genetic ancestry could explain the better adaptability and grazing abilities of XBG cattle with regards to the local environment compared to XBH cattle. This is consistent with frequent crosses with frozen semen of BS cattle having been performed in the last 20 years, as these crosses were used to improve the productive performance of the minority of XBH. Under housing conditions, XBH cattle are significantly more productive than XBG cattle, but the former's grazing adaptability is significantly reduced. The majority of XBG cattle are not inseminated using frozen semen from BS because of the grazing conditions and nonetheless maintain a strong grazing performance. Therefore, the XBG breed is more representative of the overall Xinjiang Brown cattle breed in terms of genetic background, genetic diversity, adaptability, and population size. As such, the XBG breed should be protected; otherwise, like the XBH breed, it will eventually lose its uniqueness.

The GO terms identified in this study are mainly related to the biosynthetic and metabolic processes of the molybdenum cofactor. Cattle that consumed fodder high in molybdenum during the 1930s developed a crippling illness. Since molybdate is a common trace element, inducing a dietary molybdenum shortage in plants or animals is difficult. For some animals, particularly sheep and cattle, large molybdenum intakes can produce secondary copper insufficiency, making molybdenum extremely hazardous (López-Alonso and Miranda, 2020). Most of the world's molybdenum mines are in China, with Xinjiang's molybdenum resources mainly located in Yining, Bole, and Tacheng (Li et al., 2023). These locations overlap closely with the distribution and grazing pastures of XBG cattle. As such, we hypothesize that XBG cattle may be better adapted to high molybdenum environments than XBH cattle.

When comparing differentially expressed genes between XBH and XBG cattle, cadherin 4 (*CDH4*) was enriched in cell adhesion molecules pathways for the selective sweep in XBG cattle. Cell adhesion molecules are (glyco-) proteins expressed on the cell surface that critically influence a wide array of biological processes that include hemostasis, immune responses, inflammation, embryogenesis, and neuronal tissue development (Montoya et al., 2002; Muller, 2003). The transcriptional level of *CDH4* may serve as an effective diagnostic and prognostic biomarker for renal cell carcinoma patients (Zhou et al., 2020), as it is a novel determinant of osteosarcoma tumorigenesis and metastasis (Tang et al., 2018). This level can downregulate impairments through *in vivo* infiltration and malignancies in patient-derived glioblastoma cells (Ceresa et al., 2019). Moreover, *CDH4* may function as a potential tumor suppressor gene in lung cancer (Li et al., 2017). In conclusion, *CDH4* helps confer disease resistance and may explain the higher level of disease resistance in XBG cattle.

When comparing KZ and XBH cattle, the differentially expressed genes adenylate cyclase 5 (*ADCY5*), ATP binding cassette subfamily C member 8 (*ABCC8*), potassium inwardly rectifying channel subfamily J member 11 (*KCNJ11*), and potassium calcium-activated channel subfamily M alpha 1 (*KCNMA1*) were enriched in insulin secretion pathways for the selective sweep in XBH cattle. Regarding insulin secretion, pancreatic beta cells are specialized endocrine cells that continuously sense the levels of blood sugar and other substrates and, in response, secrete insulin to maintain normal metabolic homeostasis. Glucose-induced insulin secretion and its potentiation constitute the principal mechanism of insulin release (Seino et al., 2010; Rorsman and Braun, 2013). A functional regulatory variant associated with type 2 diabetes is located at the *ADCY5* locus in a pancreatic islet enhancer (Roman et al., 2017). The most frequent genetic cause of hyperinsulinism and neonatal diabetes is pathogenic mutations in *KCNJ11* and *ABCC8*; the subunits of the β-cell ATP-sensitive potassium channel, a crucial element of the glucose-stimulated insulin secretion pathway, are encoded by these genes. Dysregulated insulin secretion results from

mutations in these two genes (De Franco et al., 2020; Timmers et al., 2021). Exercise and diet influence insulin sensitivity and secretion (Ding et al., 2019) in XBH cattle with low exercise and high feed energy levels; this may reflect genetic alterations that have occurred to adapt to housing management. In addition, the osteoclast differentiation pathways associated with were also enriched; differential genes included signal regulatory protein β1 (*SIRPB1*), signal-regulatory protein alpha (*SIRPα*), and signal-regulator protein gamma (*SIRPG*). Osteoclasts, multinucleated cells originating from the hematopoietic monocyte-macrophage lineage, are responsible for bone resorption (Nakashima and Takayanagi, 2009; Takayanagi, 2010). A member of the immunoglobulin superfamily, SIRPB1 is a signal regulatory protein that can control receptor tyrosine kinase-coupled signaling. SIRPB1 is a potential oncogene capable of activating Akt signaling to stimulate prostate cancer proliferation (Song et al., 2020), and the *SIRPB1* gene confers susceptibility to Crohn's disease (Tang et al., 2023). The tumor micro-environment features a marked expression of *SIRPα*, an inhibitory receptor present on myeloid cells, as well as its widely distributed counter-receptor *CD47* (De Vlaminck et al., 2021). Genetically, both *SIRPB1* and *SIRPα* are associated with disease resistance and immunity.

Ultimately, the development of core SNPs for XBG cattle provides a basis for the next step of customizing a solid-phase or liquid-phase gene microarray dedicated to Xinjiang brown cattle for germplasm resource identification, genome-wide association studies research, and genomic selection.

Overall, we comprehensively evaluated the genetic relationship and diversity of XBG cattle compared with two ancestral breeds and another type of the same breed (XBH). Our findings provide new insights into the historical contribution of foreign BS and Chinese KZ breeds to Xinjiang Brown cattle. These findings will help develop a reliable and sustainable strategy for the conservation and improvement of Xinjiang Brown cattle. This study's results convey that SLAF-seq initially provides very few loci and even fewer loci following data merging, resulting in few enriched GO terms and KEGG pathways. After determining the representativeness and breed significance of XBG cattle, WGS was required to obtain additional loci and information to detail germplasm characteristics.

## 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 Animal Welfare and Ethics Committee of Xinjiang Agricultural University (approval number: 2021100). Written informed consent was obtained from the owners for the participation of their animals in this study.

## Author contributions

XW: Conceptualization, Visualization, Writing–original draft. ZM: Data curation, Writing–review and editing. LG: Investigation, Resources, Writing–review and editing. LY: Writing–review and editing. ZY: Investigation, Writing–review and editing. FC: Investigation, Writing–review and editing. XG: Resources, Writing–review and editing. WL: Writing–review and editing. XY: Funding acquisition, Investigation, Project administration, Supervision, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Xinjiang Uygur Autonomous Region Science and Technology Major Project (No. 2022A02001-1), Xinjiang Brown Cattle Joint Breeding Group Improvement and Enhancement Action Plan Issues (No. 2023XJHN-14), and Xinjiang Agriculture Research System (No. XJARS-XM-08).

## Acknowledgments

We are grateful to Dr. Yanyan Zhou from the School of Life Sciences, Zhengzhou University, and Dr. Zhen Liu from the College of Animal Science and Technology, China Agricultural University, for their technical guidance. We would also like to thank Editage (www.editage.cn) for English language editing.

## 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



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1348329/full#supplementary-material

# References


Aerts, J., Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., et al. (2013). SLAF-seq: an efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8, e58700. doi:10.1371/journal.pone.0058700

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109

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. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi:10.1111/mec.12354

Ceresa, D., Alessandrini, F., Bosio, L., Marubbi, D., Reverberi, D., Malatesta, P., et al. (2019). Cdh4 down-regulation impairs *in vivo* infiltration and malignancy in patients derived glioblastoma cells. *Int. J. Mol. Sci.* 20, 4028. doi:10.3390/ijms20164028

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat. Commun.* 9, 2337. doi:10.1038/s41467-018-04737-0

Chen, Q., Xu, L., Zhang, M., Zhang, T., Yan, M., Zhai, M., et al. (2022a). Whole genome resequencing reveals the genetic contribution of Kazakh and Swiss Brown cattle to a population of Xinjiang Brown cattle. *Gene* 839, 146725. doi:10.1016/j.gene.2022.146725

Chen, Z., He, Y., Iqbal, Y., Shi, Y., Huang, H., and Yi, Z. (2022b). Investigation of genetic relationships within three Miscanthus species using SNP markers identified with SLAF-seq. *BMC Genom* 23, 43. doi:10.1186/s12864-021-08277-8

Cingolani, P., Platts, A., Wang Le, 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. (Austin)* 6, 80–92. doi:10.4161/fly.19695

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 Franco, E., Saint-Martin, C., Brusgaard, K., Knight Johnson, A. E., Aguilar-Bryan, L., Bowman, P., et al. (2020). Update of variants identified in the pancreatic β-cell KATP channel genes KCNJ11 and ABCC8 in individuals with congenital hyperinsulinism and diabetes. *Hum. Mutat.* 41, 884–905. doi:10.1002/humu.23995

De Vlaminck, K., Romão, E., Puttemans, J., Pombo Antunes, A. R., Kancheva, D., Scheyltjens, I., et al. (2021). Imaging of glioblastoma tumor-associated myeloid cells using nanobodies targeting signal regulatory protein alpha. *Front. Immunol.* 12, 777524. doi:10.3389/fimmu.2021.777524

Ding, C., Chooi, Y. U. C., Chan, Z., Lo, J., Choo, J., Ding, B. T. K., et al. (2019). Dose-dependent effects of exercise and diet on insulin sensitivity and secretion. *Med. Sci. Sports Exerc.* 51, 2109–2116. doi:10.1249/mss.0000000000002020

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

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* 17, 27–32. doi:10.1111/1755-0998.12509

Friel, J., Bombarely, A., Fornell, C. D., Luque, F., and Fernández-Ocaña, A. M. (2021). Comparative analysis of genotyping by sequencing and whole-genome sequencing methods in diversity studies of *Olea europaea* L. *Eur. L. Plants* 10, 2514. doi:10.3390/plants10112514

Gao, L., Jia, J., and Kong, X. (2016). A SNP-based molecular barcode for characterization of common wheat. *PLoS One* 11, e0150947. doi:10.1371/journal.pone.0150947

Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nat. Genet.* 3, 266–272. doi:10.1038/ng0393-266

Graebner, R. C., Hayes, P. M., Hagerty, C. H., and Cuesta-Marcos, A. (2015). A comparison of polymorphism information content and mean of transformed kinships as criteria for selecting informative subsets of barley (Hordeum vulgare L. s. l.) from the USDA Barley Core Collection. *Genet. Resour. Crop Evol.* 63, 477–482. doi:10.1007/s10722-015-0265-z

Jones, J. C., Fan, S., Franchini, P., Schartl, M., and Meyer, A. (2013). The evolutionary history of Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Mol. Ecol.* 22, 2986–3001. doi:10.1111/mec.12269

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Li, H., and Durbin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009b). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, Y., Zhou, P., Shen, X., and Zhao, K. (2023). Molybdenum fertilizer improved antioxidant capacity of Chinese Merino sheep under compound contamination. *Biol. Trace Elem. Res.* 201, 1717–1725. doi:10.1007/s12011-022-03266-8

Li, Z., Su, D., Ying, L., Yu, G., and Mao, W. (2017). Study on expression of CDH4 in lung cancer. *World J. Surg. Oncol.* 15, 26. doi:10.1186/s12957-016-1083-2

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., and Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Hum. Genom.* 7, 1. doi:10.1186/1479-7364-7-1

López-Alonso, M., and Miranda, M. (2020). Copper supplementation, a challenge in cattle. *Anim. (Basel)* 10, 1890. doi:10.3390/ani10101890

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

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi:10.1016/j.ajhg.2013.06.020

Mauki, D. H., Tijjani, A., Ma, C., Ng'ang'a, S. I., Mark, A. I., Sanke, O. J., et al. (2022). Genome-wide investigations reveal the population structure and selection signatures of Nigerian cattle adaptation in the sub-Saharan troPics. *BMC Genom* 23, 306. doi:10.1186/s12864-022-08512-w

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, 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

Montoya, M. C., Sancho, D., Vicente-Manzanares, M., and Sánchez-Madrid, F. (2002). Cell adhesion and polarity during immune interactions. *Immunol. Rev.* 186, 68–82. doi:10.1034/j.1600-065x.2002.18607.x

Muller, W. A. (2003). Leukocyte-endothelial-cell interactions in leukocyte transmigration and the inflammatory response. *Trends Immunol.* 24 (6), 327–334. doi:10.1016/s1471-4906(03)00117-0

Nakashima, T., and Takayanagi, H. (2009). Osteoimmunology: crosstalk between the immune and bone systems. *J. Clin. Immunol.* 29 (5), 555–567. doi:10.1007/s10875-009-9316-6

Pan, Y., Wang, X., Sun, G., Li, F., and Gong, X. (2016). Application of RAD sequencing for evaluating the genetic diversity of domesticated Panax notoginseng (Araliaceae). *PLoS One* 11, e0166419. doi:10.1371/journal.pone.0166419

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol. Ecol. Resour.* 16, 608–627. doi:10.1111/1755-0998.12512

Roman, T. S., Cannon, M. E., Vadlamudi, S., Buchkovich, M. L., Wolford, B. N., Welch, R. P., et al. (2017). A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the ADCY5 locus. *Diabetes* 66, 2521–2530. doi:10.2337/db17-0464

Rorsman, P., and Braun, M. (2013). Regulation of insulin secretion in human pancreatic islets. *Annu. Rev. Physiol.* 75, 155–179. doi:10.1146/annurev-physiol-030212-183754

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 (3), giaa021. doi:10.1093/gigascience/giaa021

Sambrook, J., and Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol:chloroform. *Cold Spring Harb. Protoc.* 2006, pdb.prot4455. doi:10.1101/pdb.prot4455

Seino, S., Shibasaki, T., and Minami, K. (2010). Pancreatic beta-cell signaling: toward better understanding of diabetes and its treatment. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 86 (6), 563–577. doi:10.2183/pjab.86.563

Song, Q., Qin, S., Pascal, L. E., Zou, C., Wang, W., Tong, H., et al. (2020). SIRPB1 promotes prostate cancer cell proliferation via Akt activation. *Prostate* 80, 352–364. doi:10.1002/pros.23950

Takayanagi, H. (2010). New immune connections in osteoclast formation. *Ann. N. Y. Acad. Sci.* 1192, 117–123. doi:10.1111/j.1749-6632.2009.05303.x

Tang, J., Wan, X., Zhang, J., Diao, N., Zhang, C., Gao, X., et al. (2023). A frameshift variant in the SIRPB1 gene confers susceptibility to Crohn's disease in a Chinese population. *Front. Genet.* 14, 1130529. doi:10.3389/fgene.2023.1130529

Tang, Q., Lu, J., Zou, C., Shao, Y., Chen, Y., Narala, S., et al. (2018). CDH4 is a novel determinant of osteosarcoma tumorigenesis and metastasis. *Oncogene* 37, 3617–3630. doi:10.1038/s41388-018-0231-2

Tian, H. L., Wang, F. G., Zhao, J. R., Yi, H. M., Wang, L., Wang, R., et al. (2015). Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. *Mol. Breed.* 35, 136. doi:10.1007/s11032-015-0335-0

Timmers, M., Dirinck, E., Lauwers, P., Wuyts, W., and De Block, C. (2021). ABCC8 variants in MODY12: review of the literature and report of a case with severe complications. *Diabetes Metab. Res. Rev.* 37, e3459. doi:10.1002/dmrr.3459

Uren, C., Hoal, E. G., and Möller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genet.* 21, 40. doi:10.1186/s12863-020-00845-3

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. doi:10.1101/gr.073585.107

Yan, X.-M., Zhang, Z., Liu, J.-B., Li, N., Yang, G.-W., Luo, D., et al. (2021). Genome-wide identification and analysis of long noncoding RNAs in longissimus muscle tissue from Kazakh cattle and Xinjiang brown cattle. *Anim. Biosci.* 34, 1739–1748. doi:10.5713/ajas.20.0317

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

Yin, R., Kwoh, C. K., and Zheng, J. (2019). "Whole genome sequencing analysis," in *Encyclopedia of bioinformatics and computational biology*. Editors S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press).

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. doi:10.1186/gb-2010-11-2-r14

Yurchenko, A., Yudin, N., Aitnazarov, R., Plyusnina, A., Brukhin, V., Soloshenko, V., et al. (2017). Genome-wide genotyping uncovers genetic profiles and history of the Russian cattle breeds. *Heredity* 120, 125–137. doi:10.1038/s41437-017-0024-3

Zhang, M., Luo, H., Xu, L., Shi, Y., Zhou, J., Wang, D., et al. (2022). Genomic selection for milk production traits in Xinjiang Brown cattle. *Animals* 12, 136. doi:10.3390/ani12020136

Zhang, Y. C., Kuang, M., Yang, W. H., Xu, H. X., Zhou, D. Y., Wang, Y. Q., et al. (2013). Construction of a primary DNA fingerprint database for cotton cultivars. *Genet. Mol. Res.* 12, 1897–1906. doi:10.4238/2013.January.30.3

Zhou, J., Liu, L., Chen, C. J., Zhang, M., Lu, X., Zhang, Z., et al. (2019). Genome-wide association study of milk and reproductive traits in dual-purpose Xinjiang Brown cattle. *BMC Genom* 20, 827. doi:10.1186/s12864-019-6224-x

Zhou, X., Huang, H., Cui, W., Wang, Y., Luo, W., Matskova, L., et al. (2020). Expression and prognostic significance of cadherin 4 (CDH4) in renal cell carcinoma. *Med. Sci. Monit.* 26, e922836. doi:10.12659/msm.922836

Zhou, Y., and Pan, H. (2023). Specific-locus amplified fragment sequencing (SLAF-Seq). *Methods Mol. Biol.* 2638, 165–171. doi:10.1007/978-1-0716-3024-2_11

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

frontiers | Frontiers in Genetics

TYPE Correction
PUBLISHED 28 February 2024
DOI 10.3389/fgene.2024.1372841



OPEN ACCESS

EDITED AND REVIEWED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

*CORRESPONDENCE
Wujun Liu,
lwj_ws@163.com
Xiangmin Yan,
yanxiangmin1014@sohu.com

RECEIVED 18 January 2024
ACCEPTED 01 February 2024
PUBLISHED 28 February 2024

CITATION
Wang X, Ma Z, Gao L, Yuan L, Ye Z, Cui F, Guo X, Liu W and Yan X (2024), Corrigendum: Genome-wide survey reveals the genetic background of Xinjiang Brown cattle in China. *Front. Genet.* 15:1372841. doi: 10.3389/fgene.2024.1372841



# Corrigendum: Genome-wide survey reveals the genetic background of Xinjiang Brown cattle in China

Xiao Wang[1,2], Zhen Ma[3], Liang Gao[2], Lixin Yuan[3], Zhibing Ye[3], Fanrong Cui[3], Xiaoping Guo[4], Wujun Liu[1]* and Xiangmin Yan[3]*

[1]College of Animal Science, Xinjiang Agricultural University, Urumqi, China, [2]Yili Vocational and Technical College, Yili, China, [3]Institute of Animal Science, Xinjiang Academy of Animal Science, Urumqi, China, [4]Yili Kazakh Autonomous Prefecture General Animal Husbandry Station, Yili, China

KEYWORDS

**Xinjiang Brown cattle, specific-locus amplified fragment-sequencing, genetic structure, genetic diversity, candidate genes, ancestry proportion**

A Corrigendum on
Genome-wide survey reveals the genetic background of Xinjiang Brown cattle in China

by Wang X, Ma Z, Gao L, Yuan L, Ye Z, Cui F, Guo X, Liu W and Yan X (2024). Front. Genet. 14: 1348329. doi: 10.3389/fgene.2023.1348329

In the published article, there was an error in the legend for Figure 4 as published. In the phrase "(B) Principal component analysis for the first two PCs of the 178 studied cattle", 178 needs to be replaced with 177 in order to be consistent with the numbers in the text. The corrected legend appears below.

"Figure 4. Phylogenetic relationship and population structure of the Xinjiang Brown cattle-grazing type (XBG) cattle and the other three breeds evaluated in this study. **(A)** Neighbor-joining phylogenetic tree constructed from single-nucleotide variant data among four populations. **(B)** Principal component analysis for the first two PCs of the 177 studied cattle. **(C)** ADMIXTURE analysis with four presumed ancestral groups to two presumed ancestral groups (K = from 2 to 4)."

In the published article, there was an error in the legend for Figure 5 as published. In the phrase "Ancestry proportion of the 130 XBG and 20 XBH individuals inferred using RFMix, as based on the reference panels of Kazakh and Brown Swiss cattle", 130 needs to be replaced with 129, in order to be consistent with the numbers in the text**.** The corrected legend appears below.

"Figure 5. Ancestry proportion of the 129 XBG and 20 XBH individuals inferred using RFMix, as based on the reference panels of Kazakh and Brown Swiss cattle."

In the published article, there was an error in the **Funding**. "National Agricultural Science and Technology Special Project of China (No. NK2022130302)" is a secret item and its number needs to be deleted. The correct **Funding** statement appears below.

"The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Xinjiang Uygur Autonomous

Region Science and Technology Major Project (No. 2022A02001-1), Xinjiang Brown Cattle Joint Breeding Group Improvement and Enhancement Action Plan Issues (No. 2023XJHN-14), and Xinjiang Agriculture Research System (No. XJARS-XM-08).”

The authors apologize for these errors and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

## Publisher’s note



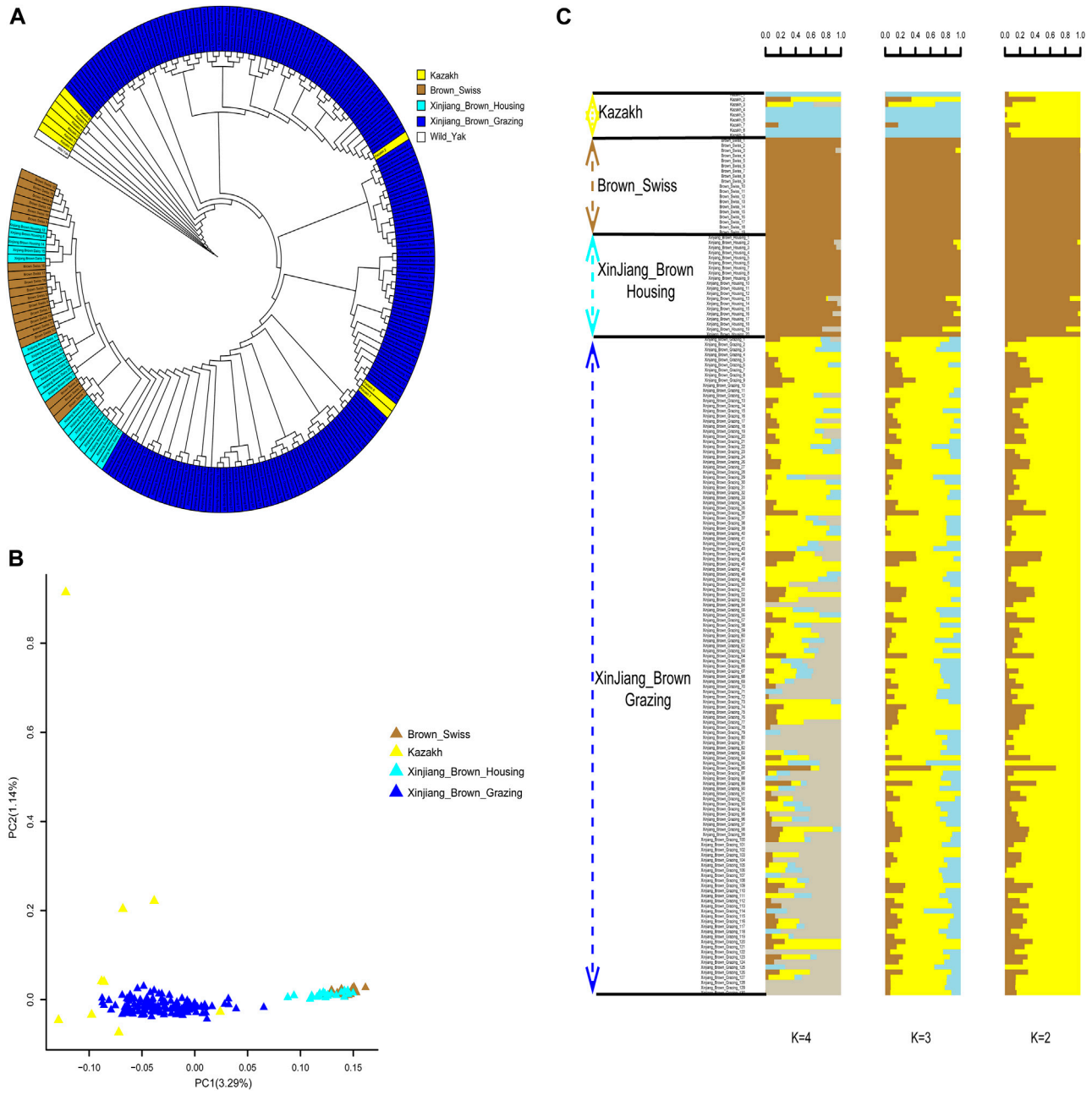



FIGURE 4
Phylogenetic relationship and population structure of the Xinjiang Brown cattle-grazing type (XBG) cattle and the other three breeds evaluated in this study. **(A)** Neighbor-joining phylogenetic tree constructed from single-nucleotide variant data among four populations. **(B)** Principal component analysis for the first two PCs of the 177 studied cattle. **(C)** ADMIXTURE analysis with four presumed ancestral groups to two presumed ancestral groups (K = from 2 to 4).

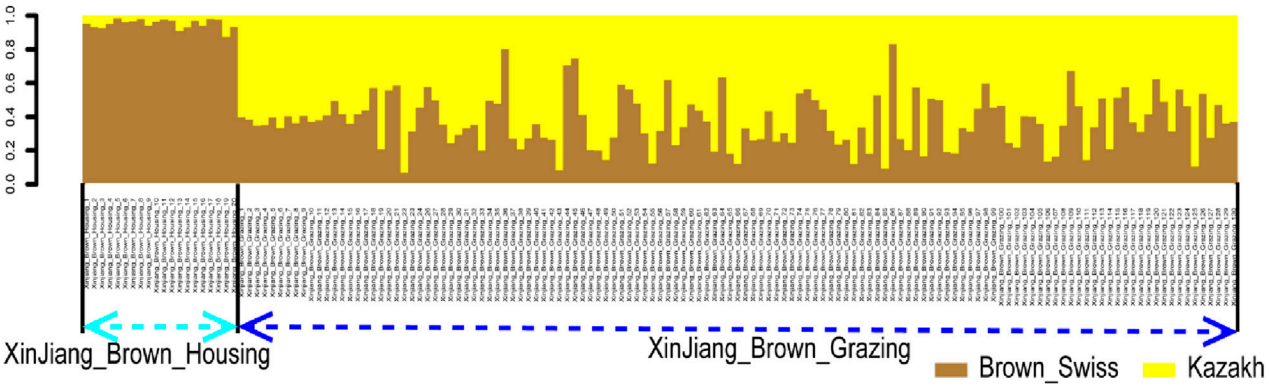



FIGURE 5
Ancestry proportion of the 129 XBG and 20 XBH individuals inferred using RFMix, as based on the reference panels of Kazakh and Brown Swiss cattle.

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 22 January 2024
DOI 10.3389/fvets.2024.1320484



OPEN ACCESS

EDITED BY
Ana Fabrícia Braga Magalhães,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Tainã Figueiredo Cardoso,
Brazilian Agricultural Research Corporation (EMBRAPA), Brazil
Qianfu Gan,
Fujian Agriculture and Forestry University, China

*CORRESPONDENCE
Shi-Yi Chen
✉ sychensau@gmail.com

RECEIVED 12 October 2023
ACCEPTED 08 January 2024
PUBLISHED 22 January 2024

CITATION
Jia X, Kang Z, Wang G, Zhang K, Fu X, Li C, Lai S and Chen S-Y (2024) Long-read sequencing-based transcriptomic landscape in longissimus dorsi and transcriptome-wide association studies for growth traits of meat rabbits.
*Front. Vet. Sci.* 11:1320484.
doi: 10.3389/fvets.2024.1320484



# Long-read sequencing-based transcriptomic landscape in longissimus dorsi and transcriptome-wide association studies for growth traits of meat rabbits

Xianbo Jia[1], Zhe Kang[1], Guozhi Wang[1], Kai Zhang[2], Xiangchao Fu[2], Congyan Li[3], Songjia Lai[1] and Shi-Yi Chen[1]*

[1]Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, China, [2]Sichuan Academy of Grassland Sciences, Chengdu, China, [3]Animal Breeding and Genetics Key Laboratory of Sichuan Province, Sichuan Animal Science Academy, Chengdu, China

Rabbits are an attractive meat livestock species that can efficiently convert human-indigestible plant biomass, and have been commonly used in biological and medical researches. Yet, transcriptomic landscape in muscle tissue and association between gene expression level and growth traits have not been specially studied in meat rabbits. In this study Oxford Nanopore Technologies (ONT) long-read sequencing technology was used for comprehensively exploring transcriptomic landscape in Longissimus dorsi for 115 rabbits at 84 days of age, and transcriptome-wide association studies (TWAS) were performed for growth traits, including body weight at 84 days of age and average daily gain during three growth periods. The statistical analysis of TWAS was performed using a mixed linear model, in which polygenic effect was fitted as a random effect according to gene expression level-based relationships. A total of 18,842 genes and 42,010 transcripts were detected, among which 35% of genes and 47% of transcripts were novel in comparison with the reference genome annotation. Furthermore, 45% of genes were widely expressed among more than 90% of individuals. The proportions (±SE) of phenotype variance explained by genome-wide gene expression level ranged from $0.501 \pm 0.216$ to $0.956 \pm 0.209$, and the similar results were obtained when explained by transcript expression level. In contrast, neither gene nor transcript was detected by TWAS to be statistically significantly associated with these growth traits. In conclusion, these novel genes and transcripts that have been extensively profiled in a single muscle tissue using long-read sequencing technology will greatly improve our understanding on transcriptional diversity in rabbits. Our results with a relatively small sample size further revealed the important contribution of global gene expression to phenotypic variation on growth performance, but it seemed that no single gene has an outstanding effect; this knowledge is helpful to include intermediate omics data for implementing genetic evaluation of growth traits in meat rabbits.

KEYWORDS

**TWAS, variance components, body weight, average daily gain, growth performance**

## Introduction

Domestic rabbits (*Oryctolagus cuniculus*) are a very prolific and small herbivorous livestock with a global population of ~600 million (1). Rabbits are mainly raised in China, North Korea, and some European countries for providing meat, wool, fur, as well as the laboratory animal. Rabbit meat is characterized by excellent nutritional characteristics, such as high protein content, high percentage of unsaturated fatty acids, high content of essential amino acids, low fat content, and low cholesterol and sodium level (2, 3). On the other hand, rabbits, as well as other herbivorous livestock, can efficiently utilize plant fiber fractions that are indigestible to human (4). In this context, it is economically necessary to improve growth performance of meat rabbits through genetic selection approaches, especially in the era of genomics (5). Of course, reproductive performance is another important contribution to lifetime productivity in rabbits (6, 7). In comparison with other common livestock species, however, we remain less-known about genomic architecture and transcriptomic landscape underlying phenotypic variation on economically important traits in meat rabbits.

The live body weight (BW) at various ages and average daily gain of BW (ADG) post-weaning are two main types of traits that have been commonly used for measuring individual growth performance in rabbits. In Gabali rabbits, the estimates of heritability ranged from 0.06 to 0.26 for individual BW at 28 to 84 days of age, with the highest estimate at 56 days of age (8). Through divergent selection on individual BW, the estimate of heritability was 0.22 for BW at 63 days of age (9). Overall, moderate heritabilities of BW at various ages have been reported in meat rabbits (10). Like BW, the heritability estimates were generally moderate for post-weaning ADG. In two rabbit lines, the estimates of heritability were 0.19 and 0.22 for ADG under *ad libitum* and restricted feeding systems, respectively (11). Piles and Tusell (12) estimated the genetic correlation between growth and fertility in rabbits, and reported the heritability of 0.15 for ADG. Therefore, the moderate heritabilities of measured traits in relation to growth performance in meat rabbits could facilitate the genetic selection and improvement, for which García and Argente (13) provided a comprehensive review on the advances of genetic improvements achieved in meat rabbits.

In human and livestock, genome-wide association studies (GWAS) have been widely used for identifying quantitative trait loci (QTL) and causal genes/variants significantly affecting complex traits (14). Using 320 K genome-wide single-nucleotide polymorphisms (SNPs), Yang et al. (15) performed GWAS of BW at seven different ages in meat rabbits and suggested the significant candidate QTL and genes. Instead of using independent BW records measured at a specific age, GWAS was alternatively performed in meat rabbits via combining the fitting of growth curve based on multiple BW measurements and estimation of SNP effects into a single-step nonlinear mixed model (16). In rabbits, GWAS have been also applied to other production traits, such as feed efficiency (17), number of teats (18), and coat colour (19). Because of the relatively low SNP density used in these studies, however, it is difficult to identify causal genes and variants within the large genomic regions revealed by significant association signals.

Instead of genetic variant-trait association, transcriptome-wide association studies (TWAS) have been increasingly used during the past years to identify the association between gene expression levels and complex traits, which may effectively improve the power for identifying causal genes (20). TWAS also could fill up the gaps between significant variants and finally manifest phenotype that are mediated by transcriptional regulation. Because of the high cost and technological limitations, gene expression data involved in TWAS have been commonly obtained via computational imputation approaches based on both a small reference set of gene expression data and a large number of genotyped individuals (21). Due to the increasing throughput and decreasing cost, single-molecule long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), are becoming increasingly routine approaches for transcriptome profiling (22, 23). In rabbits, a transcriptome atlas was successfully revealed using PacBio sequencing technology (24). In the present study, we aimed to: (1) comprehensively profile transcriptomic landscape in the muscle tissue using ONT RNA sequencing technology, (2) estimate the phenotypic variance of growth traits explained by genome-wide gene expression levels, and (3) perform TWAS with these traits in meat rabbits.

## Materials and methods

### Animals and phenotypes

A crossbred population of Zika rabbits and Sichuan White rabbits was used in this study, and all of them were F1 offspring of four males and 14 females. After weaning at 35 days of age, all rabbits were fed a routine commercial pellet diet (labelled as: digestible energy = 10.5 MJ, protein = 15.5%, and crude fiber = 16.5%) and housed in cages of 50 × 40 × 40 cm in size until 84 days of age (two and one rabbits per cage before and after 70 days of age, respectively). The air conditioning control system was used when indoor temperature was higher than 25°C. The individual BW was measured at 35 (BW35), 56 (BW56), 70 (BW70), and 84 (BW84) days of age, respectively; and 119 rabbits were successfully collected for BW records at the four time points. These BW records were quality controlled by removing the outliers that reside outside the median ± 3 × median absolute deviation (MAD) at each time point (25), after which a total of 115 rabbits were finally retained. Based on these BW records, individual ADG for three growth periods were derived as between BW70 and BW84 (ADG70), between BW56 and BW84 (ADG56), and between BW35 and BW84 (ADG35), respectively.

### Samples and transcriptome sequencing

All rabbits at 84 days of age were slaughtered by electrical shock after fasting for 24 h, and Longissimus dorsi tissue was collected and snap-frozen in liquid nitrogen for total RNA extraction and ONT transcriptome sequencing for each individual. Total RNA was extracted using RNASimple Total RNA Kit (Tiangen Biotech, Beijing, China) following the manufacturer's instruction. RNA concentration and RNA integrity number (RIN) were analyzed using Nanodrop 2000C (Thermo Fisher Scientific, Waltham, United States) and Agilent 2,100 Bioanalyzer (Agilent Technologies, Santa Clara, United States), respectively (Supplementary Table S1). The sequencing libraries were prepared using ~1 μg of quantified RNA sample and cDNA-PCR Sequencing Kit (SQK-PCS109, Oxford Nanopore

Technologies). In brief, the full-length cDNAs were enriched using template switching activity of reverse transcriptase. PCR adapters were directly added to both ends of the first-strand cDNAs. After 14 rounds of PCR amplification using LongAmp Tag (NEB), ONT sequencing adaptors were ligated to PCR products using T4 DNA ligase (NEB). DNA purification was performed using Agencourt AMPure XP beads (Beckman, CA, United States). The final cDNA libraries were added to FLO-MIN109 flowcells and run on PromethION platform at Biomarker Technology Company (Beijing, China).

## Assembly and quantification of transcripts

The raw sequencing reads were first subjected to quality controls (QC) for identifying, orienting, and reusing full-length Nanopore cDNA reads using Pychopper software with default parameters.[1] During this QC process, we discarded the short (<50 bp) or low quality (mean base quality <7.0) reads that accounted for 2.3% of raw sequencing reads on average. These qualified reads were aligned to rabbit reference genome sequences (UM_NZW_1.0, with only the autosome sequences) using minimap2 software with parameters of "-ax splice-p 0.9-N 1" (26). Herein, both reconstruction of transcripts and quantitation of gene/isoform expression levels were simultaneously performed for all individuals using IsoQuant software with the default parameters (23), which employs the intron graphs for reconstructing transcripts with reference genome annotation. The novel mono-exonic transcripts were not used.

Regarding the novel transcripts that have not been annotated yet, protein-coding potential was predicted using CPC2 software with the default parameters (27). The gene expression was quantified by directly counting the uniquely aligned reads but not requiring the necessary consistency with its isoform(s), while the preset parameters specifically regarding ONT long reads was used for matching read-to-isoform relationship (23). After the raw read counts were normalized using TMM method (28), both gene and transcript expression levels were finally measured as counts per million reads (CPM) using edgeR R package (29).

## Transcriptome-wide association studies

To avoid the bias resulting from lowly expressed genes in association studies, we only retained genes that had been effectively expressed (raw read count ≥2) within more than 30% of individuals. The associations between gene/transcript expression level and the trait of interest were analyzed using OSCA software (30) and mixed linear model (MLM) as:

$$\mathrm{y} = \mathrm{w}_i b_i + \mathrm{Xb} + \mathrm{Zu} + \mathrm{e},$$

where y is an $n \times 1$ vector of each trait (i.e., BW84, ADG70, ADG56, and ADG35) with $n$ being the sample size; $b_i$ is the estimated effect of gene $i$ on the trait with its expression level vector $\mathrm{w}_i$; X is an $n \times 2$ incidence matrix for the two covariates of sex (two levels) and birth season (three levels) with the effect vector b; Z is an $n \times m$ matrix containing the normalized expression levels of $m$ genes; u is the an $m \times 1$ vector of joint effect of all genes (also termed the polygenic effect) on the trait with $\mathrm{u} \sim \mathrm{N}\left(0, \mathrm{A}\sigma_o^2\right)$, in which A is the expression level-based relationship matrix (ORM) and $\sigma_o^2$ is the proportion of phenotype variance explained by all genes; e is an $n \times 1$ vector of residuals with $\mathrm{e} \sim \mathrm{N}\left(0, \mathrm{I}\sigma_e^2\right)$. The element of A matrix between individual $j$ and $k$ was computed as (30):

$$A_{jk} = \frac{1}{m}\sum_i \left(x_{ij} - \mu_i\right) / \left(x_{ik} - \mu_i\right) / \sigma_i^2,$$

where $x_{ij}$ and $x_{ik}$ are the normalized expression level of gene $i$ in the individual $j$ and $k$, respectively; $\mu_i$ and $\sigma_i^2$ are the mean and variance of expression level for gene $i$ across all individuals, respectively. To avoid the double fitting problem of one target gene simultaneously considered as both fixed and random effects in the MLM, the MOMENT (multi-component MLM-based omic association excluding the target) module implemented in OSCA software (30) was used with the default parameters. The variance components of $\sigma_o^2$ and $\sigma_e^2$ were estimated using Restricted Maximum Likelihood (REML) algorithm in OSCA software (30). The multiple comparison adjustment was performed using Bonferroni approach (31), therefore, the $P$ threshold of 0.05 divided by total number of genes/transcripts was used for defining the genome-wide significant gene/transcript. If no genome-wide significant gene/transcript was found, we alternatively listed the top 20 protein-coding genes that have the lowest $p$ values as the suggestive candidates. The genomic inflation factor ($\lambda$) and 95% confidence interval were further computed for checking if there was potential population stratification problem (32).

## Functional analyses of candidate genes

For the candidate genes proposed, functional enrichment analyses were conducted using the g:GOSt function of the g:Profiler web server (33), including the target data sets of the GO terms (34) and KEGG pathways (35). The default parameters and methods for adjusting for multiple hypotheses testing (i.e., the build-in g:SCS method) were used, targeting an adjusted 5% level of significance.

# Results

## Transcriptome profiling in longissimus dorsi

An average of 6.6 million raw ONT long reads per individual were initially obtained with the mean length of 1,138 bp (Supplementary Table S1). During our QC process, the autotuned parameter of $q$ value that determines the stringency of primer alignment was 0.1724 (Supplementary Figure S1), by which up to 97.7% of raw reads passed the QC steps. These qualified reads were aligned against all 21 autosomes with the average mapping ratio of 83.55%, ranging from 75.72 to 88.20% (Supplementary Table S1).

1 www.github.com/epi2me-labs/pychopper

A total of 18,842 genes and 42,010 transcripts were detected among all the studied individuals, and 6,531 genes (35%) and 19,949 transcripts (47%) of them were novel in comparison with the reference genome annotation. On average there were 2.85 and 1.07 transcripts per gene for the known and novel genes identified, respectively (Figure 1A); therefore, most of these novel genes (97%) were single-transcript genes. The average number of exons was 9.17 and 4.42 for the known and novel transcripts, and the median sequence length was 2,546 bp and 1,229 bp, respectively (Figure 1B). Among the novel transcripts, 7,535 transcripts (38%) were predicted to be protein-coding. The mean and median lengths of the first exons of transcripts were 889.9 bp and 359.0 bp, respectively (Figure 1C). Based on the raw counts of mapped reads, 45% of genes were widely expressed among more than 90% of individuals studied, while 21% of genes were restrictively expressed among less than 10% of individuals (Figure 1D).

## Transcriptome-wide association studies

The descriptive statistics of all four traits are shown in Table 1. The average BW84 was 1982 g and had the higher variability among individuals than other three traits. The average ADG decreased from 25.59 g/day between 35 and 84 days of age (ADG35) to 19.84 g/day between 70 and 84 days of age (ADG70), and the decreased variability was also observed. The phenotypic correlations of BW84 with ADG35, ADG56, and ADG70 were 0.780, 0.644, and 0.447, respectively. Among the three ADG traits, the phenotypic correlations ranged from 0.578 between ADG35 and ADG70 to 0.743 between ADG35 and ADG56 (Supplementary Figure S2).

After removing the lowly expressed genes, 10,290 genes were remained for the association analyses; and there was no obvious population stratification according to the global gene expression level (Figure 2). The estimates of variance components of MLM are shown in Table 2. Beside ADG70 that was not converged successfully, the phenotype variances explained by gene expression level (± standard error, SE) were 0.659 ± 0.198 for BW84, 0.956 ± 0.209 for ADG35, and 0.501 ± 0.216 for ADG56, respectively. The association analyses are shown in Figure 3, for which the genomic inflation factors (95% CI) were 0.947 (1.133–0.761) for BW84, 0.942 (1.123–0.760) for ADG35, and 1.005 (1.190–0.821) for ADG56. As no genome-wide significant gene was found by the association analyses, the top 20 protein-coding genes with the lowest $p$ values are shown in Table 3. Among them, three genes were overlapped among three traits (TAR RNA binding protein 1, *TARBP1*) or between two traits (Kelch repeat and BTB domain containing 12, *KBTBD12* and Leukotriene A4 hydrolase, *LTA4H*). Regarding these candidate genes, functional enrichment analysis revealed a significant GO term of "non-membrane spanning protein tyrosine kinase activity" (the adjusted $p$ value = 0.0071), in which three genes of *LYN* (LYN proto-oncogene), *PKDCC* (protein kinase domain containing cytoplasmic), *JAK1* (Janus kinase 1) were involved. No significant KEGG was found.

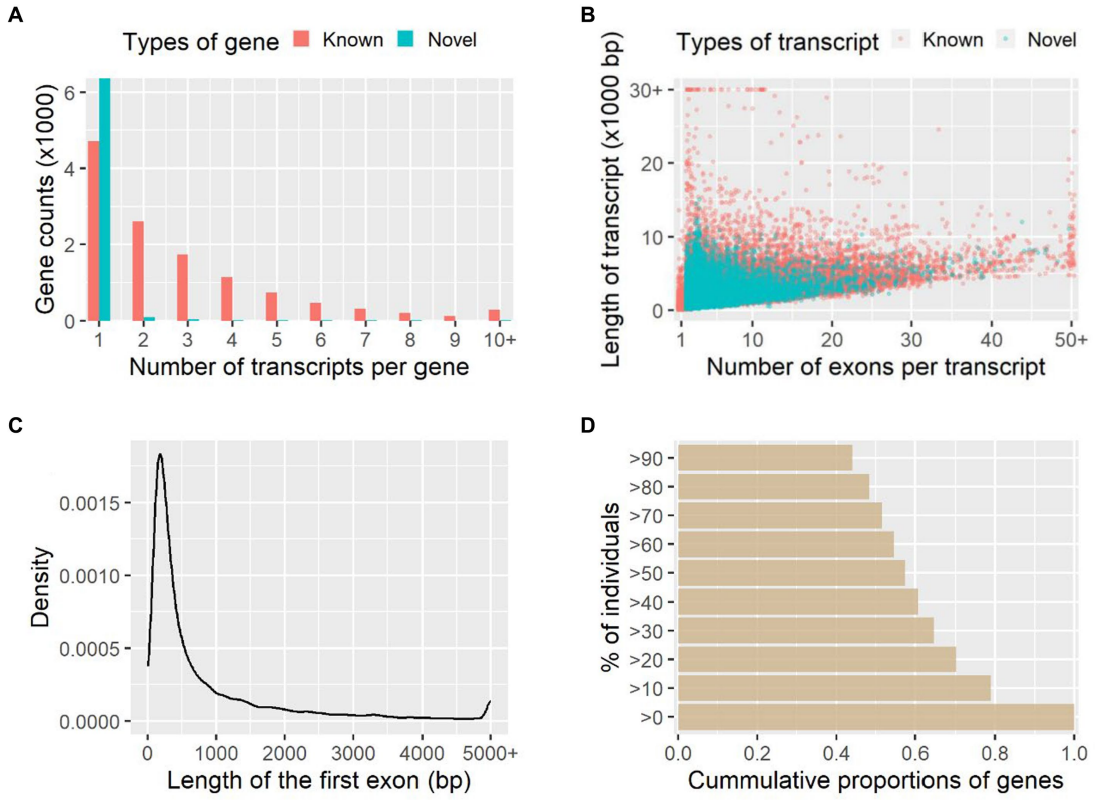



FIGURE 1
Transcript assembly and gene expression quantification. The numbers of transcripts per gene and numbers of exons per transcript are shown in **(A)** and **(B)**, respectively. Length distribution of the first exon of transcripts **(C)** and the cumulative proportions of genes expressed among individuals **(D)** are further shown.

TABLE 1 Descriptive statistics of the growth traits.

| Traits | Mean | SD | Min | Max | CV |
|---|---|---|---|---|---|
| BW84 (g) | 1981.61 | 289.84 | 1.120 | 2.689 | 6.84 |
| ADG35 (g/day) | 25.59 | 4.51 | 13.97 | 36.98 | 5.67 |
| ADG56 (g/day) | 24.35 | 7.12 | 9.08 | 44.84 | 3.42 |
| ADG70 (g/day) | 19.84 | 7.01 | 5.91 | 40.91 | 2.83 |

BW84, body weight at 84 days of age; ADG35, average daily gain between 35 and 84 days of age; ADG56, average daily gain between 56 and 84 days of age; ADG70, average daily gain between 70 and 84 days of age; SD, standard deviation; Min, minimum value; Max, maximum value; CV, coefficient of variation.

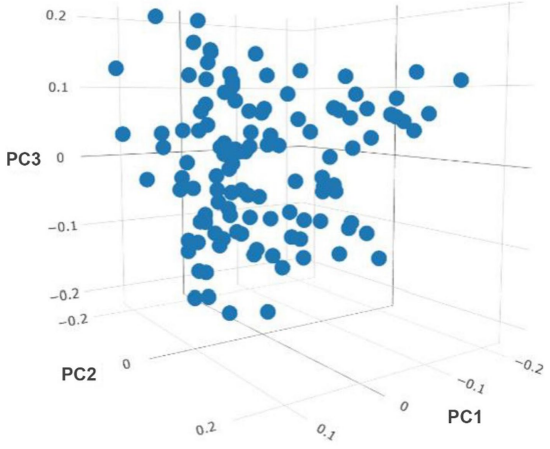



FIGURE 2
Sample clustering based on the principal component (PC) analyses of genome-wide gene expression level.

The transcript expression level-based association studies were further conducted for a total of 16,601 transcripts. As shown in Table 2, the phenotype variances explained by transcript expression levels (±SE) were 0.510 ± 0.142 for BW84 and 0.658 ± 0.223 for ADG56, respectively (MLM failed to converge for the other two traits). Like the gene-based association results, there was no transcript that showed the statistically significant association with these growth traits (Supplementary Figure S3). Based on the top 20 suggestive genes and transcripts with the lowest $p$ values, no gene was overlapped between them.

## Discussion

The phenotypic variation of complex traits in human and livestock has been determined significantly by gene transcriptional regulation (36, 37). Transcriptional diversity may be referred to the varied expression level and spatio-temporal transcription. However, both genomic architecture and transcriptomic landscape underlying economically important traits have not been extensively explored in rabbits in comparison with other livestock species, such as cattle, pigs, and chicken (5). In this study, therefore, we established an experimental cross population between Zika rabbit and Sichuan White rabbit, both of them are raised for producing meat. To investigate the transcriptional diversity extensively and accurately, we focused on a single muscle tissue among 115 individuals and also

TABLE 2 Estimates of variance components (±standard error).

| Items | Traits | $\sigma_o^2$ | $\sigma_e^2$ | $\rho^2$ |
|---|---|---|---|---|
| Genes | BW84 | 45,263 ± 29,635 | 23,354 ± 11,277 | 0.659 ± 0.198 |
| | ADG35 | 25.41 ± 16.22 | 1.16 ± 5.41 | 0.956 ± 0.209 |
| | ADG56 | 18.19 ± 17.43 | 18.09 ± 5.53 | 0.501 ± 0.216 |
| | ADG70 | Not converged | | |
| Transcripts | BW84 | 33,838 ± 21,398 | 32,478 ± 7,235 | 0.510 ± 0.142 |
| | ADG35 | Not converged | | |
| | ADG56 | 25.65 ± 19.99 | 13.33 ± 7.19 | 0.658 ± 0.223 |
| | ADG70 | Not converged | | |

BW84, body weight at 84 days of age; ADG35, average daily gain between 35 and 84 days of age; ADG56, average daily gain between 56 and 84 days of age; ADG70, average daily gain between 70 and 84 days of age.

employed the more robust long-read sequencing technology. We further investigated the proportion of phenotypic variance of growth traits explained by global gene expression variation, and performed association analyses between gene expression levels and growth traits using MLM approach (38). However, we acknowledge that the sample size involved in this study is relatively small in the context of TWAS (20).

In the past decade, the comprehensive profiling of transcriptome has been largely facilitated by the enormous advances of short-read high-throughput sequencing of RNA molecules (39). The accuracies of transcript assembly and expression quantification have been further improved due to the later single-molecule long-read RNA sequencing technologies (40). Therefore, long-read sequencing technologies are expecting to be increasingly used in transcriptome studies (41), such as in cattle (42), pigs (43), and chicken (44). In rabbit, Chen et al. (24) first used PacBio long-read sequencing technology for exploring transcriptomic landscape and revealed a large proportion of novel genes and transcripts using a pooling sample of multiple organ tissues sampled at different ages. In the present study, ONT long-read sequencing technology was similarly used for investigating transcriptomic landscape in meat rabbits for a single muscle tissue and in a large set of individuals, by which considerable numbers of genes and transcripts were revealed to be novel. Hence, these findings indicate that current reference genome of rabbit has not been well annotated yet, whereas a comprehensive annotation is required for biomedical researches when using rabbit as animal model (45).

As a monogastric herbivore, rabbit can efficiently utilize plant fiber fractions that are indigestible to human, which means that raising rabbits for meat, fur and wool can be considered as an effective contribution to achieving global food security (46). In context of meat rabbits, the improved growth appearance is economically significant and could be achieved through genetic selection approaches. Recently, García and Argente (13) provided a good review on the estimated genetic parameters for various growth traits that have traditionally been used as selection criteria in meat rabbits, and the moderate to high heritabilities were reported regarding these traits. In livestock, the genome-wide SNPs, as an alternative to pedigree records, have been increasingly used for estimating heritabilities, genetic correlations, and individual breeding values for various production traits of interest (5). However, there are obvious gaps between significant genetic

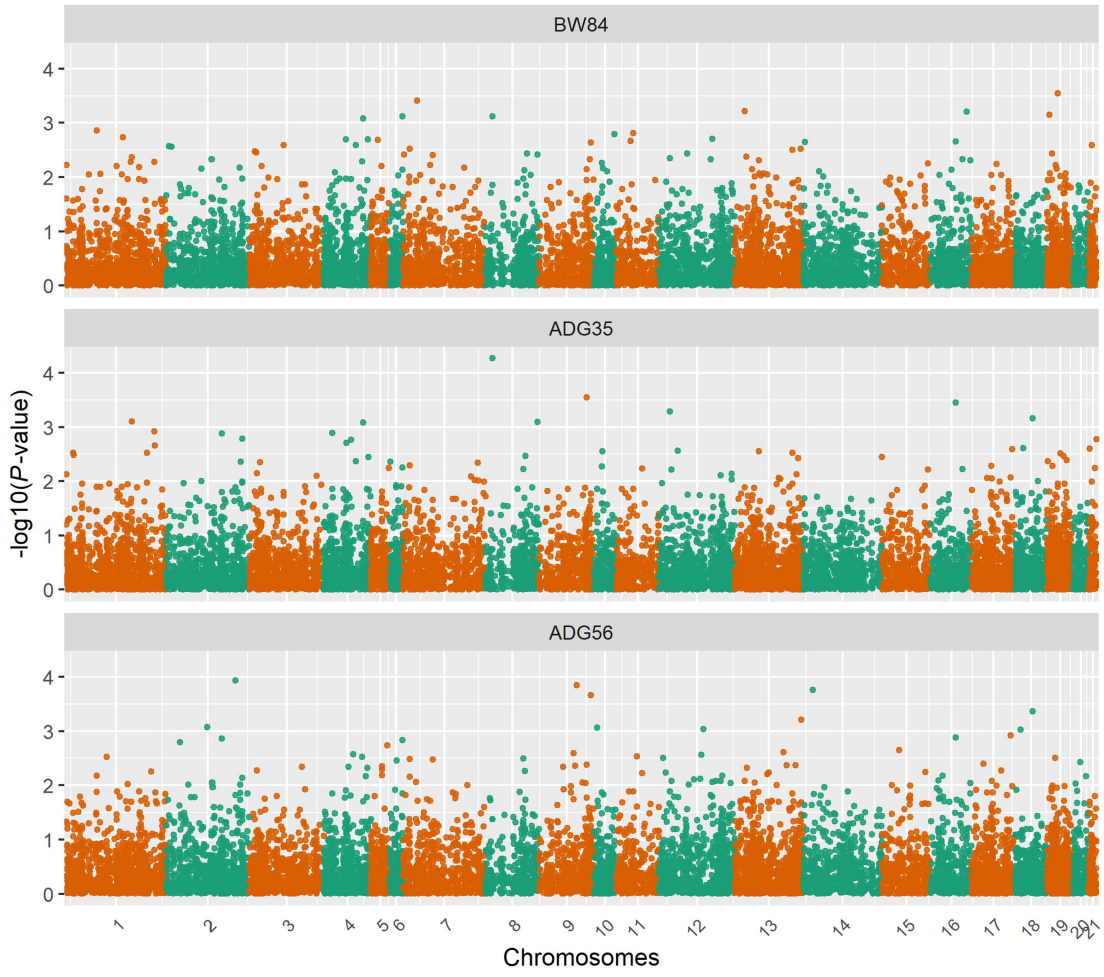



FIGURE 3
Manhattan plots for the gene expression level-based association analyses. BW84, body weight at 84 days of age; ADG35, average daily gain between 35 and 84 days of age; ADG56, average daily gain between 56 and 84 days of age.

variants identified and the finally manifested phenotype, mainly caused by the transcriptional and post-transcriptional variability (47). Therefore, genome-wide gene expression profile has been alternatively used for explaining the complex diseases in human and production traits in livestock (36, 48). In this context, we obtained transcriptome profile using long-read sequencing technology in a single population, by which individual relationships were measured and hence used in MLM for estimating different variance components of growth traits in meat rabbits (30). We first found that more than half of phenotypic variances could be explained by the genome-wide gene expression variation, which are higher than the traditional pedigree-based estimation of heritability (13). These results suggest that there is considerable contribution of gene expression level to inter-individual variation on growth performance. However, our estimates had some large SE mainly because of the relatively smaller sample size than that of pedigree- or SNP-based estimation. Using GTEx data, more recently, the proportion of phenotypic variance explained by gene expression level was estimated to be 0.68 (SE = 0.06) for body mass index in human, but the varied transcriptomic variances were observed across different tissues (49). To our best knowledge, few studies have been reported for partitioning phenotypic variance by the transcriptomic expression data regarding these economically important traits in livestock.

In addition to partitioning phenotypic variance by genome-wide gene expression level, it is possible and necessary to identify the potential causal genes that significantly affect individual phenotype using TWAS approach (20). Because of difficulty in sampling of appropriate biological issues and high cost of RNA sequencing for a large set of samples, the gene expression data used in TWAS are always imputed indirectly from a small reference data set of gene expression data (21, 48). Unfortunately, high-quality reference transcriptome data sets are not routinely available in livestock until the recent releases of FarmGTEx project for cattle and pig.[2] In Huaxi cattle, the gene expression levels in Longissimus dorsi were imputed using a reference transcriptome data set containing 120 individual RNA sequencing data, and the TWAS successfully revealed some genes significantly

2 www.farmgtex.org

TABLE 3 The suggestive candidate genes by transcriptome-wide association studies.

| Gene name | Chr | Start | End | *p* value | Description |
|---|---|---|---|---|---|
| ***BW84*** | | | | | |
| *TAF15* | 19 | 22,434,285 | 22,471,504 | 2.88E-4 | TATA-box binding protein associated factor 15 |
| *LOC100356444* | 7 | 29,654,685 | 29,720,143 | 3.90E-4 | Cytochrome P450 20A1 |
| *VAMP4* | 13 | 21,381,832 | 21,426,615 | 6.08E-4 | Vesicle associated membrane protein 4 |
| *PFKFB3* | 16 | 74,850,845 | 74,874,653 | 6.17E-4 | 6-phosphofructo-2-kinase/fructose-2 |
| *LOC127484429* | 19 | 4,468,701 | 4,473,178 | 7.14E-4 | DNA-directed RNA polymerase III subunit RPC10-like |
| *CNPY4* | 6 | 27,597,738 | 27,603,097 | 7.64E-4 | Canopy FGF signaling regulator 4 |
| *LTA4H* | 4 | 82,722,807 | 82,762,965 | 8.30E-4 | Leukotriene A4 hydrolase |
| *LOC100348955* | 1 | 63,804,078 | 63,804,605 | 1.38E-3 | 40S ribosomal protein S20 |
| *LOC100346443* | 4 | 46,872,687 | 46,880,703 | 2.04E-3 | Retinol dehydrogenase 5 |
| *CSNK2A2* | 5 | 14,124,537 | 14,163,573 | 2.04E-3 | Casein kinase 2 alpha 2 |
| *TARBP1* | 16 | 51,620,152 | 51,697,033 | 2.22E-3 | TAR (HIV-1) RNA binding protein 1 |
| *OXNAD1* | 14 | 3,497,868 | 3,560,799 | 2.25E-3 | Oxidoreductase NAD binding domain containing 1 |
| *KBTBD12* | 9 | 109,057,898 | 109,122,809 | 2.29E-3 | Kelch repeat and BTB domain containing 12 |
| *MYF6* | 4 | 67,515,503 | 67,517,420 | 2.57E-3 | Myogenic factor 6 |
| *LOC103345082* | 21 | 6,413,652 | 6,454,720 | 2.58E-3 | Transmembrane protein 120B |
| *LYN* | 3 | 71,263,078 | 71,404,233 | 2.60E-3 | LYN proto-oncogene |
| *CC2D2A* | 2 | 6,505,464 | 6,657,081 | 2.72E-3 | Coiled-coil and C2 domain containing 2A |
| *PACRGL* | 2 | 11,836,569 | 11,876,015 | 2.77E-3 | Parkin coregulated like |
| *GPN2* | 13 | 138,518,582 | 138,528,748 | 3.00E-3 | GPN-loop gtpase 2 |
| *TESK2* | 13 | 121,249,440 | 121,412,510 | 3.16E-3 | Testis associated actin remodelling kinase 2 |
| ***ADG35*** | | | | | |
| *ATRIP* | 9 | 99,898,363 | 99,916,757 | 2.84E-4 | ATR interacting protein |
| *TARBP1* | 16 | 51,620,152 | 51,697,033 | 3.52E-4 | TAR (HIV-1) RNA binding protein 1 |
| *HIVEP2* | 12 | 22,386,395 | 22,601,396 | 5.22E-4 | HIVEP zinc finger 2 |
| *MED17* | 1 | 137,049,828 | 137,074,367 | 7.93E-4 | Mediator complex subunit 17 |
| *LOC100339309* | 8 | 108,875,028 | 108,890,309 | 8.05E-4 | Thioredoxin-dependent peroxide reductase |
| *LTA4H* | 4 | 82,722,807 | 82,762,965 | 8.31E-4 | Leukotriene A4 hydrolase |
| *IMMP1L* | 1 | 184,529,319 | 184,628,501 | 1.21E-3 | Inner mitochondrial membrane peptidase subunit 1 |
| *HSPA12B* | 4 | 18,241,628 | 18,260,505 | 1.29E-3 | Heat shock protein family A (Hsp70) member 12B |
| *MAPRE3* | 2 | 159,296,622 | 159,346,159 | 1.67E-3 | Microtubule associated protein RP/EB family member 3 |
| *PTPRR* | 4 | 57,205,791 | 57,492,990 | 1.73E-3 | Protein tyrosine phosphatase receptor type R |
| *LOC100352320* | 4 | 47,568,888 | 47,570,921 | 1.96E-3 | SPRY domain-containing protein 4 |
| *EIF3M* | 1 | 185,772,686 | 185,792,687 | 2.18E-3 | Eukaryotic translation initiation factor 3 subunit M |
| *LRRTM3* | 18 | 16,304,216 | 16,475,497 | 2.44E-3 | Leucine rich repeat transmembrane neuronal 3 |
| *LOC100355489* | 21 | 2,089,119 | 2,090,689 | 2.51E-3 | CCHC-type zinc finger nucleic acid binding protein |
| *PCSK6* | 17 | 85,347,007 | 85,515,126 | 2.55E-3 | Proprotein convertase subtilisin/kexin type 6 |
| *KIAA0408* | 12 | 38,832,741 | 38,863,007 | 2.74E-3 | KIAA0408 ortholog |
| *TRIM33* | 13 | 49,949,922 | 50,093,346 | 2.83E-3 | Tripartite motif containing 33 |
| *ALAD* | 1 | 14,134,039 | 14,144,172 | 3.02E-3 | Aminolevulinate dehydratase |
| *NASP* | 13 | 121,088,515 | 121,121,163 | 3.02E-3 | Nuclear autoantigenic sperm protein |
| *FAR1* | 1 | 169,240,935 | 169,311,723 | 3.02E-3 | Fatty acyl-coa reductase 1 |
| ***ADG56*** | | | | | |
| *PKDCC* | 2 | 144,389,511 | 144,406,466 | 1.18E-4 | Protein kinase domain containing |

*(Continued)*

TABLE 3 (Continued)

| Gene name | Chr | Start | End | *p* value | Description |
|---|---|---|---|---|---|
| *MITF* | 9 | 78,687,674 | 78,933,570 | 1.43E-4 | Melanocyte inducing transcription factor |
| *OSBPL10* | 14 | 20,072,024 | 20,374,605 | 1.74E-4 | Oxysterol binding protein like 10 |
| *KBTBD12* | 9 | 109,057,898 | 109,122,809 | 2.18E-4 | Kelch repeat and BTB domain containing 12 |
| *ECRG4* | 2 | 86,711,966 | 86,728,075 | 8.53E-4 | ECRG4 augurin precursor |
| *GSDME* | 10 | 6,653,179 | 6,714,011 | 8.70E-4 | Gasdermin E |
| *TARBP1* | 16 | 51,620,152 | 51,697,033 | 1.33E-3 | TAR (HIV-1) RNA binding protein 1 |
| *TAF6* | 6 | 27,603,174 | 27,609,994 | 1.47E-3 | TATA-box binding protein associated factor 6 |
| *UGDH* | 2 | 29,316,754 | 29,343,983 | 1.60E-3 | UDP-glucose 6-dehydrogenase |
| *MAF* | 5 | 33,701,831 | 34,050,433 | 1.86E-3 | MAF bzip transcription factor |
| *LOC103350727* | 15 | 36,973,168 | 37,042,794 | 2.25E-3 | Rho gtpase-activating protein 20 |
| *JAK1* | 13 | 101,582,720 | 101,723,625 | 2.46E-3 | Janus kinase 1 |
| *LOC100346455* | 9 | 72,727,244 | 72,728,366 | 2.57E-3 | Non-histone chromosomal protein HMG-14-like |
| *GLIPR1* | 4 | 61,890,236 | 61,906,387 | 2.66E-3 | GLI pathogenesis related 1 |
| *BCKDHB* | 12 | 88,528,385 | 88,800,448 | 2.74E-3 | Branched chain keto acid dehydrogenase E1 subunit beta |
| *MRPL42* | 4 | 80,285,035 | 80,324,028 | 2.97E-3 | Mitochondrial ribosomal protein L42 |
| *OMD* | 1 | 84,305,301 | 84,316,047 | 3.00E-3 | Osteomodulin |
| *ZDHHC14* | 12 | 8,041,888 | 8,299,592 | 3.11E-3 | Zinc finger DHHC-type palmitoyltransferase 14 |
| *BLTP2* | 19 | 17,099,589 | 17,135,496 | 3.16E-3 | Bridge-like lipid transfer protein family member 2 |
| *PTMS* | 8 | 78,854,643 | 78,859,308 | 3.19E-3 | Parathymosin |

BW84, body weight at 84 days of age; ADG35, average daily gain between 35 and 84 days of age; ADG56, average daily gain between 56 and 84 days of age; Chr, chromosome.

associated with productive ability (50). In the present study, we did not detect gene or transcript that is significantly associated with growth traits in meat rabbits by TWAS approach, despite it was observed that the large proportions of phenotypic variance could be explained by genome-wide gene expression level. There are two likely explanations for the observed negative results in our TWAS. First, the growth of meat rabbits is controlled by an extremely polygenic architecture and, more importantly, no gene has an outstanding effect on the phenotype in the sense of gene expression level. Second, the relatively small sample size involved in the present study, which is expected to be increased in future studies due to a significant decrease in long-read sequencing cost, might compromise the statistical detection power of TWAS. Alternatively, we analyzed some top candidate genes with the lowest *p* values and found that they were significantly enriched into the health-related pathway. Among them, *TARBP1* was suggested to be involved in multiple cancers in human by regulating immune function (51); the significantly differential expression was observed in human Colorectal adenoma for the gene of *KBTBD12* (52); it was reported that *LTA4H* can modulate the susceptibility to Mycobacterial infection in zebrafish and humans (53). Furthermore, the significantly enriched GO term of "non-membrane spanning protein tyrosine kinase activity" in this study was previously reported to be involved in regulation of tumor immune microenvironment in glioma (54). These findings may be reasonably explained as a healthy rabbit will have the greater growth performance.

There are three practical implications of results obtained in this study. First, the comprehensively explored transcripts and their expression levels in the muscle tissue have enhanced our understanding on transcriptomic landscape associated with growth performance in meat rabbits. Second, our observation that the high proportion of phenotypic variance could be explained by global gene expression variation suggests the possibility to predict the complex and hard-to-measured production traits, such as meat quality, using transcriptomic expression data. Third, the genetic effect of single gene expression level on complex traits may be smaller than we have expected, which suggest that a large enough sample size is required for successfully identifying the significant genes. On the other hand, we acknowledged some limitations of this study. The most obvious limitation is the relatively small sample size used in TWAS, which limited the detection power for identifying the significantly associated genes. Another limitation is the absence of genome-wide SNPs that could be used for analyzing QTL affecting gene expression level.

## Conclusion

In this study, many novel genes and transcripts have been comprehensively explored in longissimus dorsi of meat rabbits using long-read RNA sequencing technology, which hence contributed to the improved annotation of rabbit genome. We also revealed that the large proportions of phenotypic variance on growth performance in meat rabbits could be explained by variation of genome-wide gene expression levels, whereas the transcriptome-wide association studies did not find gene or transcript that is statistically significantly associated with the growth traits studied.

## Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://ngdc.cncb.ac.cn/gsa/; CRA013219.

## Ethics statement

The animal study was approved by Institutional Animal Care and Use Committee of Sichuan Agricultural University. The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

XJ: Formal analysis, Methodology, Writing – original draft. ZK: Data curation, Methodology, Writing – original draft. GW: Methodology, Writing – review & editing. KZ: Resources, Writing – review & editing. XF: Resources, Writing – review & editing. CL: Resources, Writing – review & editing. SL: Resources, Writing – review & editing. S-YC: Formal analysis, Funding acquisition, Methodology, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was financially supported by National Natural Science Foundation of China (32072684), Earmarked Fund for China Agriculture Research System (CARS-44-A-2), and Science & Technology Department of Sichuan Province (2021YFYZ0033).

## 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2024.1320484/full#supplementary-material

## References

1. FAOSTAT (2023). The statistics division of the FAO. Available at: www.fao.org/faostat/en/#data (Accessed on August 2023).

2. Li S, Zeng W, Li R, Hoffman LC, He Z, Sun Q. Rabbit meat production and processing in China. *Meat Sci*. (2018) 145:320–8. doi: 10.1016/j.meatsci.2018.06.037

3. Siddiqui SA, Gerini F, Ikram A, Saeed F, Feng X. Rabbit meat-production, consumption and consumers attitudes and behavior. *Sustainability*. (2023) 15:2008. doi: 10.3390/su15032008

4. Giorgino A, Raspa F, Valle E, Bergero D, Cavallini D, Gariglio M. Effect of dietary organic acids and botanicals on metabolic status and milk parameters in mid-late lactating goats. *Animals*. (2023) 13:797. doi: 10.3390/ani13050797

5. Misztal I, Lourenco D, Legarra A. Current status of genomic evaluation. *J Anim Sci*. (2020) 98:skaa101. doi: 10.1093/jas/skaa101

6. Castellini C, Dal Bosco A, Arias-Álvarez M, Lorenzo PL, Cardinali R, Rebollar PG. The main factors affecting the reproductive performance of rabbit does: a review. *Anim Reprod Sci*. (2010) 122:174–82. doi: 10.1016/j.anireprosci.2010.10.003

7. Pollesel M, Tassinari M, Frabetti A, Fornasini D, Cavallini D. Effect of does parity order on litter homogeneity parameters. *Ital J Anim Sci*. (2020) 19:1188–94. doi: 10.1080/1828051X.2020.1827990

8. El-Deghadi AS. Selection indices for improving body weight in Gabali rabbits. *Egypt Poult Sci*. (2018) 38:1115–26. doi: 10.21608/EPSJ.2018.22904

9. Larzul C, Gondret F, Combes S, De Rochambeau H. Divergent selection on 63-day body weight in the rabbit: response on growth, carcass and muscle traits. *Genet Sel Evol*. (2005) 37:105. doi: 10.1186/1297-9686-37-1-105

10. Abdel-Kafy ES, El-Deighadi AS, Shabaan HM, Ali WH, Sabra ZEAA, Farid A. Genetic evaluation for growth traits in new synthetic rabbit line in Egypt. *Open J Agri Res*. (2021) 1:62–73. doi: 10.31586/ojar.2021.119

11. Drouilhet L, Gilbert H, Balmisse E, Ruesche J, Tircazes A, Larzul C. Genetic parameters for two selection criteria for feed efficiency in rabbits. *J Anim Sci*. (2013) 91:3121–8. doi: 10.2527/jas.2012-6176

12. Piles M, Tusell L. Genetic correlation between growth and female and male contributions to fertility in rabbit. *J Anim Breed Genet*. (2012) 129:298–305. doi: 10.1111/j.1439-0388.2011.00975.x

13. García M. L., Argente M. J. (2020). The genetic improvement in meat rabbits Lagomorpha characteristics . (Rijeka: IntechOpen)

14. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR. Genome-wide association studies. *Nat Rev Methods Primers*. (2021) 1:59. doi: 10.1038/s43586-021-00056-9

15. Yang X, Deng F, Wu Z, Chen S-Y, Shi Y, Jia X, et al. A genome-wide association study identifying genetic variants associated with growth, carcass and meat quality traits in rabbits. *Animals*. (2020) 10:1068. doi: 10.3390/ani10061068

16. Liao Y, Wang Z, Glória LS, Zhang K, Zhang C, Yang R. Genome-wide association studies for growth curves in meat rabbits through the single-step nonlinear mixed model. *Front Genet*. (2021) 12:750939. doi: 10.3389/fgene.2021.750939

17. Sánchez JP, Legarra A, Velasco-Galilea M, Piles M, Sánchez A, Rafel O. Genome-wide association study for feed efficiency in collective cage-raised rabbits under full and restricted feeding. *Anim Genet*. (2020) 51:799–810. doi: 10.1111/age.12988

18. Bovo S, Schiavo G, Utzeri VJ, Ribani A, Schiavitto M, Buttazzoni L. A genome-wide association study for the number of teats in European rabbits (Oryctolagus cuniculus) identifies several candidate genes affecting this trait. *Anim Genet*. (2021) 52:237–43. doi: 10.1111/age.13036

19. Zhang K, Wang G, Wang L, Wen B, Fu X, Liu N. A genome-wide association study of coat color in Chinese rex rabbits. *Front Vet Sci*. (2023) 10:1184764. doi: 10.3389/fvets.2023.1184764

20. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. (2019) 51:592–9. doi: 10.1038/s41588-019-0385-z

21. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. (2015) 47:1091–8. doi: 10.1038/ng.3367

22. Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S. ESPRESSO: robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci Adv*. (2023) 9:eabq5072. doi: 10.1126/sciadv.abq5072

23. Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol*. (2023) 41:915–8. doi: 10.1038/s41587-022-01565-y

24. Chen S-Y, Deng F, Jia X, Li C, Lai S-J. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci Rep*. (2017a) 7:7648. doi: 10.1038/s41598-017-08138-z

25. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol*. (2013) 49:764–6. doi: 10.1016/j.jesp.2013.03.013

26. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. (2018) 34:3094–100. doi: 10.1093/bioinformatics/bty191

27. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. (2017) 45:W12–6. doi: 10.1093/nar/gkx428

28. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. (2010) 11:R25. doi: 10.1186/gb-2010-11-3-r25

29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616

30. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol*. (2019) 20:107. doi: 10.1186/s13059-019-1718-z

31. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis*. (2017b) 9:1725–9. doi: 10.21037/jtd.2017.05.34

32. van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *J Anim Breed Genet*. (2019) 136:418–29. doi: 10.1111/jbg.12419

33. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H. G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. (2019) 47:W191–8. doi: 10.1093/nar/gkz369

34. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. (2019) 47:D330–8. doi: 10.1093/nar/gky1055

35. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. (2019) 47:D590–5. doi: 10.1093/nar/gky962

36. Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res*. (2020) 30:790–801. doi: 10.1101/gr.250704.119

37. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*. (2020) 369:1318–30. doi: 10.1126/science.aaz1776

38. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. (2010) 42:565–9. doi: 10.1038/ng.608

39. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. (2009) 10:57–63. doi: 10.1038/nrg2484

40. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ. Comprehensive comparison of Pacific biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. *F1000Research*. (2017) 6:100. doi: 10.12688/f1000research.10571.2

41. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*. (2019) 20:631–56. doi: 10.1038/s41576-019-0150-2

42. Ren Y, Tseng E, Smith TP, Hiendleder S, Williams JL. Long read isoform sequencing reveals hidden transcriptional complexity between cattle subspecies. *BMC Genomics*. (2023) 24:108. doi: 10.1186/s12864-023-09212-9

43. Shu Z, Wang L, Wang J, Zhang L, Hou X, Yan H. Integrative analysis of nanopore and illumina sequencing reveals alternative splicing complexity in pig longissimus dorsi muscle. *Front Genet*. (2022) 13:877646. doi: 10.3389/fgene.2022.877646

44. Li D, Zhong C, Sun Y, Kang L, Jiang Y. Identification of genes involved in chicken follicle selection by ONT sequencing on granulosa cells. *Front Genet*. (2023) 13:1090603. doi: 10.3389/fgene.2022.1090603

45. Mapara M, Thomas BS, Bhat KM. Rabbit as an animal model for experimental research. *Dent Res J*. (2012) 9:111–8. doi: 10.4103/1735-3327.92960

46. Hoque M, Mondal S, Adusumilli S. *Sustainable livestock production and food security in emerging issues in climate smart livestock production: Biological tools and techniques*. Cambridge, Massachusetts: Academic Press (2022).

47. Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet*. (2013) 14:168–78. doi: 10.1038/nrg3404

48. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. (2016) 48:245–52. doi: 10.1038/ng.3506

49. Jullian Fabres P, Lee SH. Phenotypic variance partitioning by transcriptomic gene expression levels and environmental variables for anthropometric traits using GTEx data. *Genet Epidemiol*. (2023) 47:465–74. doi: 10.1002/gepi.22531

50. Liang M, An B, Deng T, Du L, Li K, Cao S. Incorporating genome-wide and transcriptome-wide association studies to identify genetic elements of longissimus dorsi muscle in Huaxi cattle. *Front Genet*. (2023) 13:982433. doi: 10.3389/fgene.2022.982433

51. Manning M, Jiang Y, Wang R, Liu L, Rode S, Bonahoom M. Pan-cancer analysis of RNA methyltransferases identifies FTSJ3 as a potential regulator of breast cancer progression. *RNA Biol*. (2020) 17:474–86. doi: 10.1080/15476286.2019.1708549

52. Wang B, Wang X, Tseng Y, Huang M, Luo F, Zhang J, et al. Distinguishing colorectal adenoma from hyperplastic polyp by WNT2 expression. *J Clin Lab Anal*. (2021) 35:e23961. doi: 10.1002/jcla.23961

53. Tobin DM, Vary JC, Ray JP, Walsh GS, Dunstan SJ, Bang ND. The lta4h locus modulates susceptibility to mycobacterial infection in zebrafish and humans. *Cell*. (2010) 140:717–30. doi: 10.1016/j.cell.2010.02.013

54. Xue J, Gao HX, Sang W, Cui WL, Liu M, Zhao Y, et al. Identification of core differentially methylated genes in glioma. *Oncol Lett*. (2019) 18:6033–45. doi: 10.3892/ol.2019.10955

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 February 2024
DOI 10.3389/fgene.2024.1335591



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Zishuai Wang,
Chinese Academy of Agricultural Sciences (CAAS), China
Hyago Passe Pereira,
Brazilian Agricultural Research Corporation (EMBRAPA), Brazil

*CORRESPONDENCE
Yandong Zhan,
zhanyandong@lcu.edu.cn
Changfa Wang,
wangchangfa@lcu.edu.cn

RECEIVED 09 November 2023
ACCEPTED 26 January 2024
PUBLISHED 09 February 2024

CITATION
Wang X, Peng Y, Liang H, Zahoor Khan M, Ren W, Huang B, Chen Y, Xing S, Zhan Y and Wang C (2024), Comprehensive transcriptomic analysis unveils the interplay of mRNA and LncRNA expression in shaping collagen organization and skin development in Dezhou donkeys.
*Front. Genet.* 15:1335591.
doi: 10.3389/fgene.2024.1335591



# Comprehensive transcriptomic analysis unveils the interplay of mRNA and LncRNA expression in shaping collagen organization and skin development in Dezhou donkeys

Xinrui Wang, Yongdong Peng, Huili Liang, Muhammad Zahoor Khan, Wei Ren, Bingjian Huang, Yinghui Chen, Shishuai Xing, Yandong Zhan* and Changfa Wang*

Liaocheng Research Institute of Donkey High-Efficiency Breeding, Liaocheng University, Liaocheng, China

The primary focus of donkey hide gelatin processing lies in the dermal layer of donkey hide due to its abundant collagen content. However, the molecular mechanism involved in collagen organization and skin development in donkey skin tissue across various developmental stages remains incomplete. The current study aims to investigate the transcriptomic screening of lncRNAs and mRNA associated with skin development and collagen organization across different ages in Dezhou donkeys' skin. In the pursuit of this objective, we used nine skin tissue samples obtained from Dezhou donkeys at various ages including 8-month fetal stage, followed by 2 and 8 years. RNA-seq analysis was performed for the transcriptomic profiling of differentially expressed genes (DEGs) and lncRNAs associated with skin development in different age groups. Our investigation revealed the presence of 6,582, 6,455, and 405 differentially expressed genes and 654, 789, and 29 differentially expressed LncRNAs within the skin tissues of Dezhou donkeys when comparing young donkeys (YD) vs. middle-aged donkeys (MD), YD vs. old donkeys (OD), and MD vs. OD, respectively. Furthermore, we identified *Collagen Type I Alpha 1 Chain* (*COL1A1*), *Collagen Type III Alpha 1 Chain* (*COL3A1*), and *Collagen Type VI Alpha 5 Chain* (*COL6A5*) as key genes involved in collagen synthesis, with *COL1A1* being subject to cis-regulation by several differentially expressed LncRNAs, including ENSEAST00005041187, ENSEAST00005038497, and MSTRG.17248.1, among others. Interestingly, collagen organizational and skin development linked pathways including Protein digestion and absorption, metabolic pathways, Phosphatidylinositol 3-Kinase-Protein Kinase B signaling pathway (PI3K-Akt signaling pathway), Extracellular Matrix-Receptor Interaction (ECM-receptor interaction), and Relaxin signaling were also reported across different age groups in Dezhou donkey skin. These findings enhance our comprehension of the molecular

mechanisms underlying Dezhou donkey skin development and collagen biosynthesis and organization, thus furnishing a solid theoretical foundation for future research endeavors in this domain.

KEYWORDS

**Dezhou donkey, hide gelatin, collagen, RNA-seq, lncRNA, mRNA, KEGG, genetic markers**

# 1 Introduction

In the context of modern agriculture and the evolving role of donkeys in agricultural production, this discourse delves into the increasing utilization of donkey products, with particular emphasis on the medicinal value attributed to ejiao, derived from donkey hide (Maigari et al., 2020; Goodrum et al., 2022; Yan et al., 2023). It is of particular note that the dermal layer of donkey hide, enriched with collagen, assumes a pivotal role in the preparation of ejiao. The popularity of ejiao in China and among practitioners of traditional Chinese medicine has led to a surge in demand, subsequently causing a substantial decline in the donkey population within China and the consequential need to import raw materials from other regions such as Africa, South America, and Australia (Maigari et al., 2020; Goodrum et al., 2022). Given the growing demand for ejiao and the diminishing supply of donkey hides, it becomes imperative to explore strategies to address this supply gap. To this end, the identification of candidate genes associated with collagen deposition in donkey skin emerges as a pivotal avenue of investigation. By cultivating new strains of Dezhou donkeys with enhanced skin performance through selective breeding, it may be possible to mitigate the scarcity of donkey hides and sustain the production of ejiao.

With the rapid development of molecular biology and related disciplines, animal breeding has moved from conventional breeding to molecular breeding. Marker-assisted selection and genomic selection have become mainstream practices in molecular breeding of livestock (Yang et al., 2017). Complex traits such as diseases, production parameters and skin development in animals are controlled by several genes. While RNA-seq is considered an emerging molecular technique utilizing for screening genes associated with complex traits (Wickramasinghe et al., 2014; Song et al., 2019). Consistently, by utilizing RNA-seq as a tool, several studies have been conducted in donkeys to screened key genes associated with skin thickness (Wang et al., 2022), skeletal muscles development (Li et al., 2022; Chai et al., 2023), and skin coat color (Wang et al., 2020).

So far very little information is available regarding the molecular mechanisms involve in the development of skin and collagen organization in Dezhou Donkeys. Thus, the current study endeavors to bridge this knowledge gap by employing RNA-Seq technology to analyze lncRNAs and mRNAs in the skin of Dezhou donkeys across different age groups. Furthermore, this study documented some key pathways like Protein digestion and absorption, PI3K-Akt signaling pathway, ECM-receptor interaction, and Relaxin signaling which have strong association with collagen restructuring and skin development in Dezhou donkey. Interestingly, our findings documented genes like *COL1A1* and *COL3A1* that were involved in the regulation of above mentioned pathways. Moreover, the role of LncRNAs (ENSEAST00005041187 and ENSEAST00005038497) showed a key role in regulation of *COL1A1* gene. Overall, our findings provided the foundational model for skin biology and collagen synthesis and organization in Dezhou donkeys' skin.

# 2 Materials and methods

## 2.1 Ethical statement

The research conducted in this study adhered to stringent ethical guidelines and received the necessary approvals from the Animal Welfare and Ethics Committee of the Institute of Animal Sciences, Liaocheng University (Approval No. LC 2019-1). All aspects of the experimental procedures, including the use of experimental animals, were conducted in full compliance with local animal welfare laws, guidelines, and ethical codes. Our foremost commitment was to minimize any potential suffering experienced by the experimental animals throughout the study.

## 2.2 Experimental animals and sample collection

In this investigation, a cohort of nine male donkeys sourced from a reputable donkey farm located in Dezhou City, Shandong Province, was the subject of our study. To ensure a comprehensive analysis, the donkeys were stratified into three distinct age groups: 8-month-old fetuses, 2-year-olds, and 8-year-olds. Each age group was represented by three biological replicates. The 8-month-old fetuses were selected due to miscarriages resulting from external pressure, and skin samples were promptly collected within 1 h following miscarriage. In contrast, minimally invasive skin sampling techniques were employed for the 2-year-old and 8-year-old donkeys. All sampling locations were carefully chosen at the midpoint of the left dorsal region, situated between the sixth and seventh thoracic vertebrae. Before sampling, rigorous preparation of the sampling sites was conducted using a specialized skin preparatory instrument. Furthermore, to alleviate any potential pain, procaine was administered at the designated sampling points. Subsequently, a 5 mm skin sampler (Acuderm, United States) was employed to procure skin tissue samples, which were meticulously washed with phosphate-buffered saline (PBS) and then stored in cryotubes before rapid freezing in liquid nitrogen. These tissue samples were subsequently preserved at −80°C, awaiting RNA-Seq analysis. Following the sampling process, sterile gauze was applied to staunch any bleeding, and anti-inflammatory drugs, in

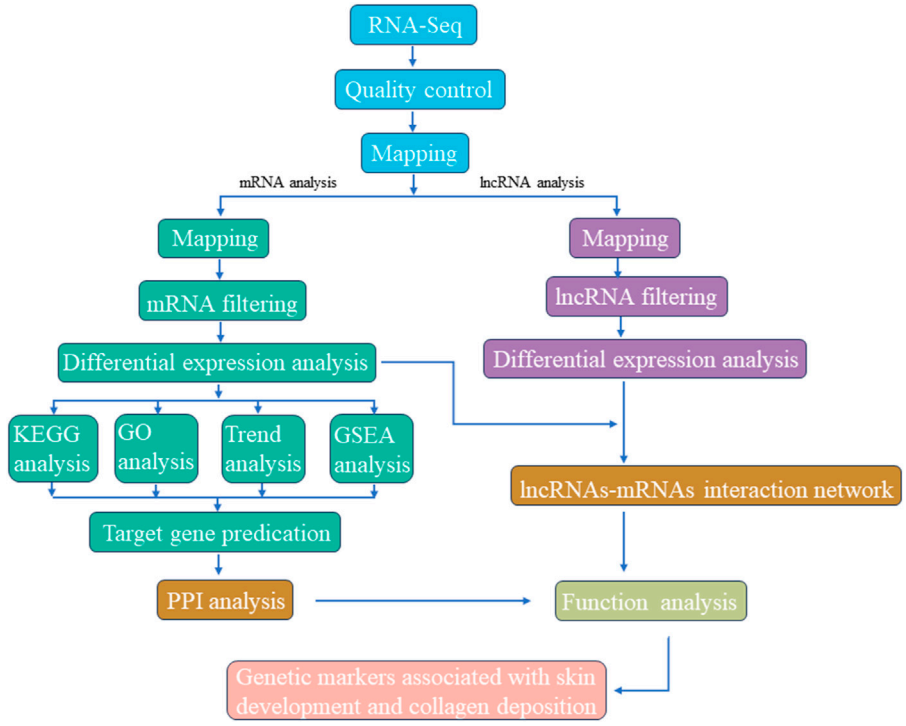



FIGURE 1
Graphical presentation of methodology adopted for screening candidate genetic markers associated with skin development and collagen deposition.

conjunction with iodine, were administered for treatment purposes. It is essential to emphasize that all the donkeys included in this study were in a healthy condition and exhibited favorable prognoses.

## 2.3 RNA extraction and sequencing

To initiate the molecular analysis, the skin tissue samples were initially ground to a fine powder in liquid nitrogen. Total RNA extraction was then carried out using the Trizol reagent kit (Invitrogen, Carlsbad, CA, United States) as per the manufacturer's guidelines. The quality of extracted RNA was assessed using the Agilent 2,100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States), with RNase-free agarose gel electrophoresis employed for additional quality evaluation. Subsequent to total RNA extraction, eukaryotic mRNA was selectively enriched through the use of Oligo (dT) beads. The construction of cDNA libraries was accomplished by Gene *Denovo* Biotechnology Co. (Guangzhou, China), and Illumina Novaseq6000 was the chosen platform for sequencing.

## 2.4 Sequencing data quality control and read mapping

The overall RNA-seq protocol adopted for this study has been summarized in Figure 1. To ensure the reliability of our data, a series of rigorous quality control measures were instituted. Initial filtering of raw reads was accomplished using fastp (Chen et al., 2018) (version 0.18.0). Reads containing adapters, those with more than 10% unknown nucleotides (N), and those characterized by low quality, defined as having more than 50% bases with a quality score (q-value) of ≤20, were systematically removed from the dataset. Furthermore, Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.8) was deployed to eliminate reads marked as rRNA, thereby resulting in a collection of high-quality clean reads ready for subsequent assembly and analysis. HISAT2.2.4 (Kim et al., 2015) was subsequently employed to align the paired-end clean reads with the reference genome of the Dezhou donkey (ASM1607732v2). The assembled reads from each sample were consolidated using StringTie v1.3.1 (Pertea et al., 2015; Pertea et al., 2016). The calculation of FPKM (fragment per kilobase of transcript per million mapped reads) values to quantify gene expression levels was facilitated by RSEM (Li and Dewey, 2011). Correlation analysis was executed using R, while principal component analysis (PCA) was carried out utilizing the gmodels package (http://www.rproject.org/). Differential expression analysis was undertaken using DESeq2 (Love et al., 2014) software, which was employed to identify differentially expressed genes (DEGs) meeting the criteria of a fold change ≥2.00 and an adjusted $p$-value of 0.05.

## 2.5 Identification and prediction of differentially expressed LncRNAs

Identification of potential lncRNAs was accomplished through a comprehensive multi-step process. Firstly,

Gffcompare was employed to retain transcripts classified as class "u," signifying intergenic transcripts (Shi et al., 2019a). Subsequently, based on the merged GTF file, only transcripts exhibiting characteristics such as more than one exon and a length exceeding 200bp were retained (Chen et al., 2019a; Mishra and Wang, 2021). Coding potential assessment of non-coding transcripts was executed using CPC2, CNCI, LGC, and PLEK, with the intersection of predictions being designated as novel lncRNAs (Chen et al., 2019; Mishra and Wang, 2021). To further enhance specificity, transcripts translated in all six possible frames were subjected to scrutiny using HMMER. Any transcripts displaying homology with known protein family domains in the Pfam database were excluded. Furthermore, the BLASTX program was harnessed to filter out transcripts bearing similarity to known proteins present in the NCBI nr and UniRef90 databases (E Value < 1e-5). Finally, transcripts with FPKM values exceeding zero in at least one sample were retained.

## 2.6 Potential target gene prediction and network construction

The identification of potential target genes was facilitated through the utilization of BEDTools (version 2.17.0) (Quinlan and Hall, 2010). Adjacent protein-coding genes situated within a 100 kb radius of each lncRNA locus were selected. Additionally, protein-coding genes exhibiting a Pearson correlation coefficient exceeding 0.95 with differentially expressed lncRNAs were considered as potential target genes. These target genes were subsequently integrated with the DEGs dataset. For a more comprehensive understanding of the interactions between DEGs and various groups, a protein-protein interaction (PPI) network was constructed employing the STRING database (https://string-db.org/) (Szklarczyk et al., 2015). Furthermore, an lncRNA-mRNA network was formulated based on targeting relationships, with visualization achieved through the Cytoscape software (V3.9.0) (The Cytoscape Consortium, United States), employing default parameters and the "layout = attribute circle layout" setting (Saito et al., 2012).

## 2.7 Functional enrichment analysis of DEGs and DELs

Functional enrichment analysis was conducted to gain insights into the biological relevance of DEGs and DELs. The GO database (Ashburner et al., 2000) was employed to predict molecular functions, cellular components, and biological processes associated with DEGs and DELs. Comparison with the Gene Ontology database (http://www.geneontology.org/) enabled mapping of all DEGs and DELs to their respective GO terms. Additionally, KEGG annotation (http://www.genome.jp/kegg) (Kanehisa and Goto, 2000) was used to subject DEGs and DELs to KEGG enrichment analysis. False discovery rate (FDR) correction, specifically employing the Benjamini–Hochberg adjustment method, was applied. GO and KEGG terms boasting $p$ values below 0.05 were identified as significantly enriched.

# 3 Results

## 3.1 Overview of sequencing data in Dezhou donkey skin

A total of nine cDNA libraries were constructed, each representing a distinct time point or developmental stage. The raw data obtained from these libraries yielded an average of approximately 46,860,812 reads per sample. These raw reads were subjected to rigorous quality control measures, resulting in an average of 46,637,055 clean reads per sample. Notably, the clean reads constituted approximately 99.37% of the raw data, underscoring the high quality of the sequencing data. Furthermore, the quality assessment revealed that the clean reads possessed exceptional accuracy, with Q20 and Q30 percentages exceeding 97.15% and 92.26%, respectively (Supplementary Table S1). These metrics are indicative of the reliability and precision of the sequencing process. Subsequently, a critical step in the analysis involved mapping the clean reads to the reference genome of the Dezhou donkey (ASM1607732v2). Impressively, over 92.86% of the clean reads were successfully mapped to the reference genome. Among these mapped reads, approximately 76.11% aligned to the exon regions, approximately 12.75% to the intron regions, and approximately 11.14% to the intergenic regions. These mapping results provided essential information regarding the distribution of sequenced reads across different genomic regions (Table 1). To further elucidate the transcriptome profile, we employed StringTie to assemble transcripts for each library. Subsequently, all assembled transcripts were synthesized into a nonredundant transcript dataset using StringTie-Merge. This comprehensive transcriptome dataset served as a foundation for downstream analyses. Additionally, the study identified a subset of 539 putative long non-coding RNAs (lncRNAs) based on the criteria illustrated in Figure 2.

## 3.2 Identification of differentially expressed mRNAs (DEMs) and long non-coding RNAs (DELs) in the Dezhou donkey

The overall data has been provided in Supplementary Table S2. In this section, we delve into the analysis of differentially expressed mRNAs (DEMs) and long non-coding RNAs (DELs) in the skin tissues of Dezhou donkeys across various comparative groups. Our approach involved stringent criteria for identifying these differential expressions based on fold change and statistical significance. Based on the criteria of $|\log_2 FC(\text{fold change})| > 1$ and q-value <0.05 normalized expression, total of 6,582, 6,455, and 405 differentially expressed mRNAs (DEMs) (Figures 3C–E) respectively. In addition, 654, 789, and 29 differentially expressed lncRNAs (DELs) were presented in the skin tissues of Dezhou donkeys among YD vs. MD, YD vs. OD, and MD vs. OD, respectively (Figures 3F–H). Furthermore, Venn diagram analysis reveals 182 differential genes were commonly shared among the three groups (YD vs. MD, YD vs. OD, and MD vs. OD) (Figures 3A, B).

To provide a comprehensive visualization of the differential expression patterns in both mRNAs and lncRNAs, clustered heatmaps were constructed. These heatmaps offer a clear

TABLE 1 Quality assessment of sequencing data.

| Sample | RawDatas | CleanData (%) | Q20 (%) | Q30 (%) | Total_Mapped (%) | exon (%) | intron (%) | intergenic (%) |
|---|---|---|---|---|---|---|---|---|
| YD-1 | 45633838 | 99.46 | 97.59 | 93.20 | 96.34 | 74.81 | 14.34 | 10.85 |
| YD-2 | 44955666 | 99.37 | 97.15 | 92.26 | 95.72 | 74.22 | 14.71 | 11.07 |
| YD-3 | 48310130 | 99.48 | 97.49 | 92.98 | 96.17 | 75.90 | 13.48 | 10.62 |
| MD-1 | 45978156 | 99.57 | 97.76 | 93.61 | 93.15 | 79.82 | 9.94 | 10.24 |
| MD-2 | 50951222 | 99.56 | 97.66 | 93.37 | 92.85 | 76.79 | 12.10 | 11.11 |
| MD-3 | 47430242 | 99.57 | 97.87 | 93.86 | 93.20 | 77.35 | 11.55 | 11.10 |
| OD-1 | 43964952 | 99.62 | 97.42 | 92.75 | 92.63 | 75.48 | 13.24 | 11.28 |
| OD-2 | 46192542 | 99.50 | 97.66 | 93.31 | 92.83 | 76.06 | 12.23 | 11.71 |
| OD-3 | 48330560 | 99.56 | 97.73 | 93.45 | 92.97 | 74.59 | 13.16 | 12.24 |

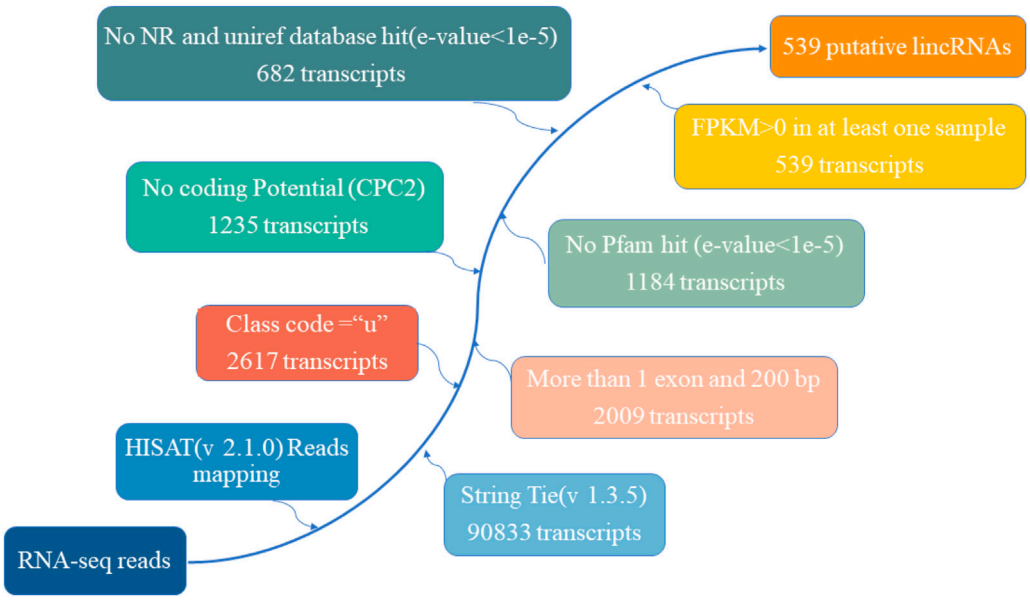



FIGURE 2
The pipeline for the identification of putative lncRNAs in this study, while the frames in the direction of the arrow show the filtering process and the number of screened transcripts.

representation of the distinct expression profiles among the identified DEMs and DELs. The clustering methodology utilized here groups genes with comparable expression patterns, as visually depicted in Figure 3I (mRNAs) and Figure 3J (lncRNAs).

## 3.3 GO and KEGG enrichment analysis of DEGs

The overall data obtained for pathways and functional processes [(Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC)] has been presented in Supplementary Table S3. To investigate the functions and pathways of DEGs, we conducted separate analyses using GO terms and KEGG pathways. According to the analysis of GO terms for the DEMs, we identified a total of 182 DEMs that were significantly enriched in 36 GO terms. These terms were further categorized into 24 Biological Processes (BP), 10 Molecular Functions (MF), and 2 Cellular Components (CC). Notably, the Biological Processes were primarily associated with cellular processes, metabolic processes, and biological regulation. In the Molecular Functions category, the top terms included binding, catalytic activity, and molecular function regulation. Our GO analysis of Cellular Components indicated that the DEMs were enriched in two specific GO terms, namely, cellular anatomical entity and protein-containing complex (Figure 4A). Furthermore, the KEGG pathway analysis unveiled a total of 168 pathways. Among these, 45 pathways demonstrated significant enrichment ($p$-value <0.05) (Figure 4B). In addition, based on bioinformatics analysis, we selected the pathways and biological function processes associated with protein metabolism and skin development.

10.3389/fgene.2024.1335591

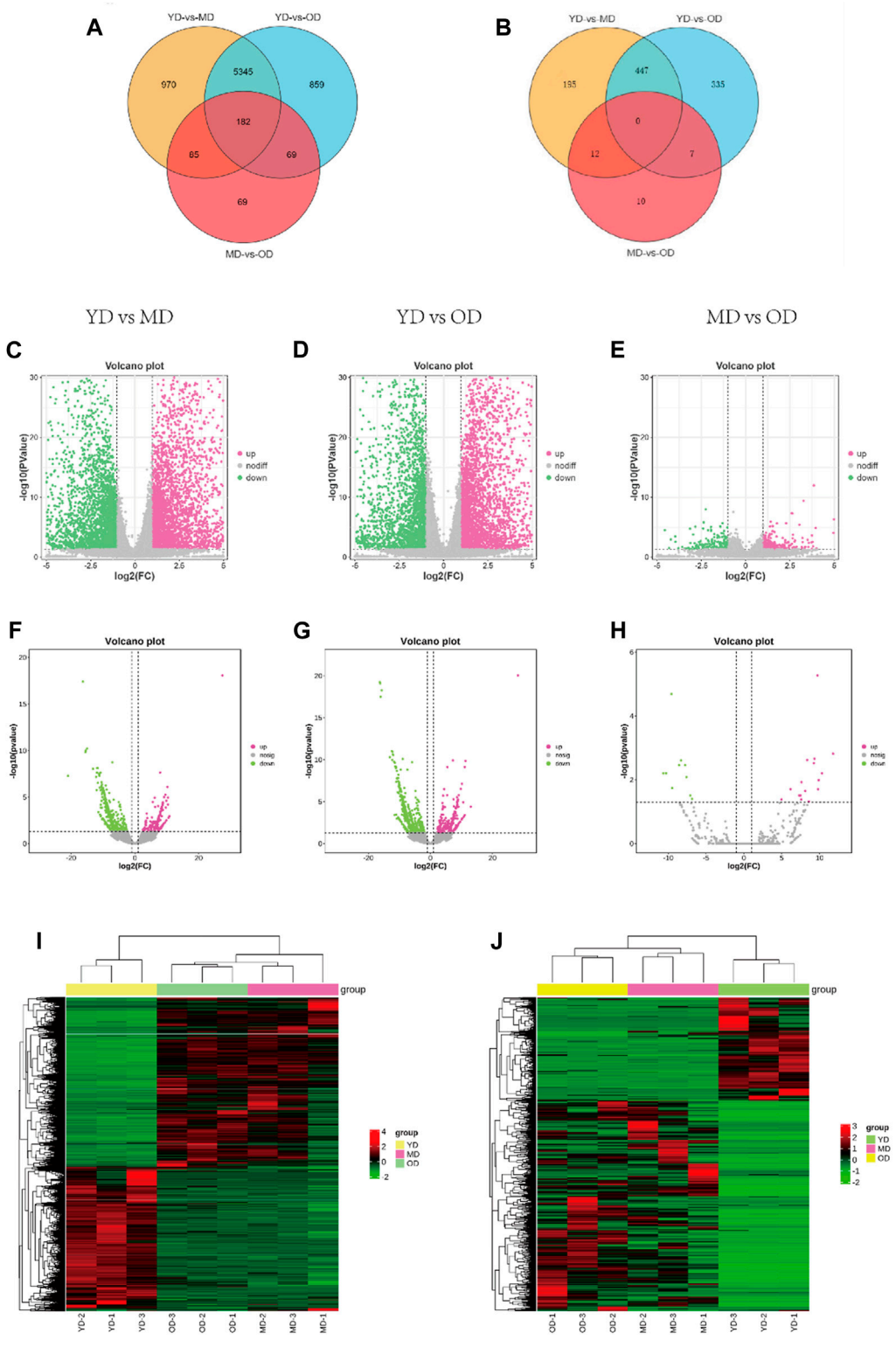



FIGURE 3
Analysis of DEMs and DELs expression profiles in the Dezhou donkey. **(A, B)** are the Venn diagrams of differentially expressed mRNAs and lncRNAs, respectively. **(C–H)** are the volcano plots of mRNAs and lncRNAs, respectively. **(I, J)** are the heatmaps of differentially expressed mRNAs and lncRNAs, respectively.

In order to effectively address the limitations of traditional enrichment analysis in mining relevant information for low-effect genes, we conducted Gene Set Enrichment Analysis (GSEA) analysis on three groups of differentially expressed genes that were co-expressed (Figure 5). Through this analysis, we identified significant enrichment of GO terms

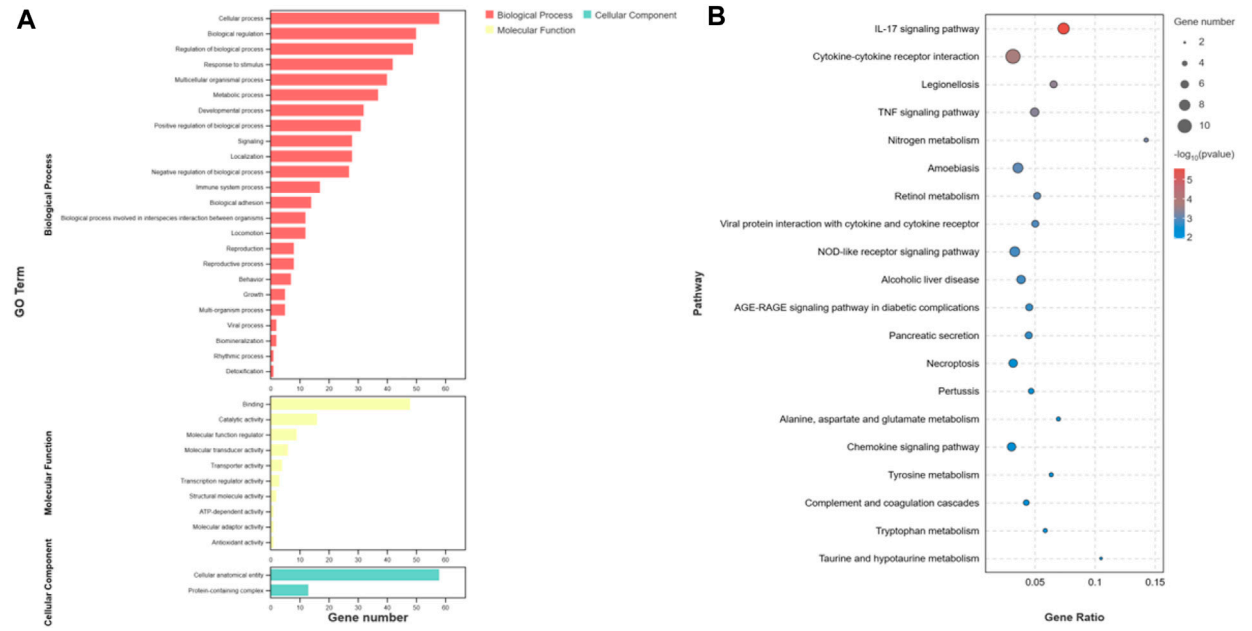



FIGURE 4
Functional analysis of differentially expressed mRNAs. **(A)** is the bar graph showing the enrichment of GO terms and the number of genes involved in differentially expressed mRNAs. **(B)** represents the top 20 enriched KEGG pathways for differentially expressed mRNAs.

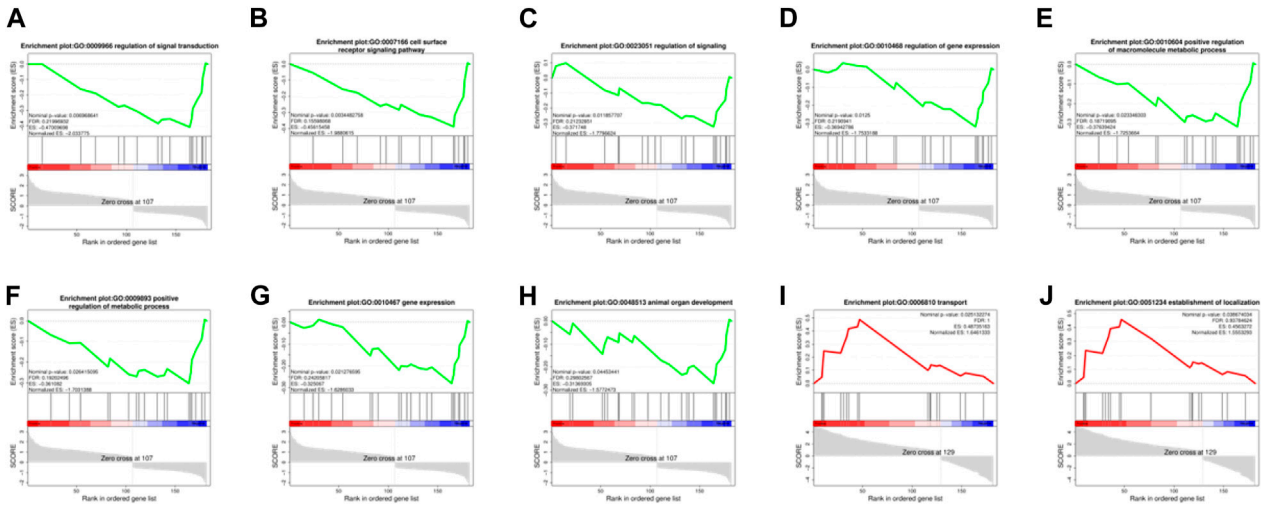



FIGURE 5
Enrichment Analysis (GSEA) analysis on three groups of 182 differentially expressed genes. **(A–H)**: The significantly enriched GO terms related to protein synthesis during the MD vs. OD period (nominal *p*-value<0.05). **(I, J)**: The significantly enriched GO terms related to protein synthesis during the YD vs. OD period (nominal *p*-value<0.05).

related to protein synthesis (Supplementary Table S4). At the same time, a trend analysis was performed on the target gene group (Figure 6; Supplementary Table S5), which was classified into six profiles (excluding profile 3 and 4). Enrichment analysis was conducted on these profiles, and the results indicated a significant association between the expression of numerous genes and pathways such as Cell growth and death, Signal transduction, metabolic process, immune system, and signal molecules and interaction.

## 3.4 PPI analysis of the DEMs related to protein deposition in the Dezhou donkey skin

In order to gain comprehensive insights of the biological processes associated with protein synthesis candidate genes, we concentrated on the posttranslational protein levels of the candidate genes. Therefore, we constructed protein-protein interaction networks (PPIs) by the STRING (https://string-db.

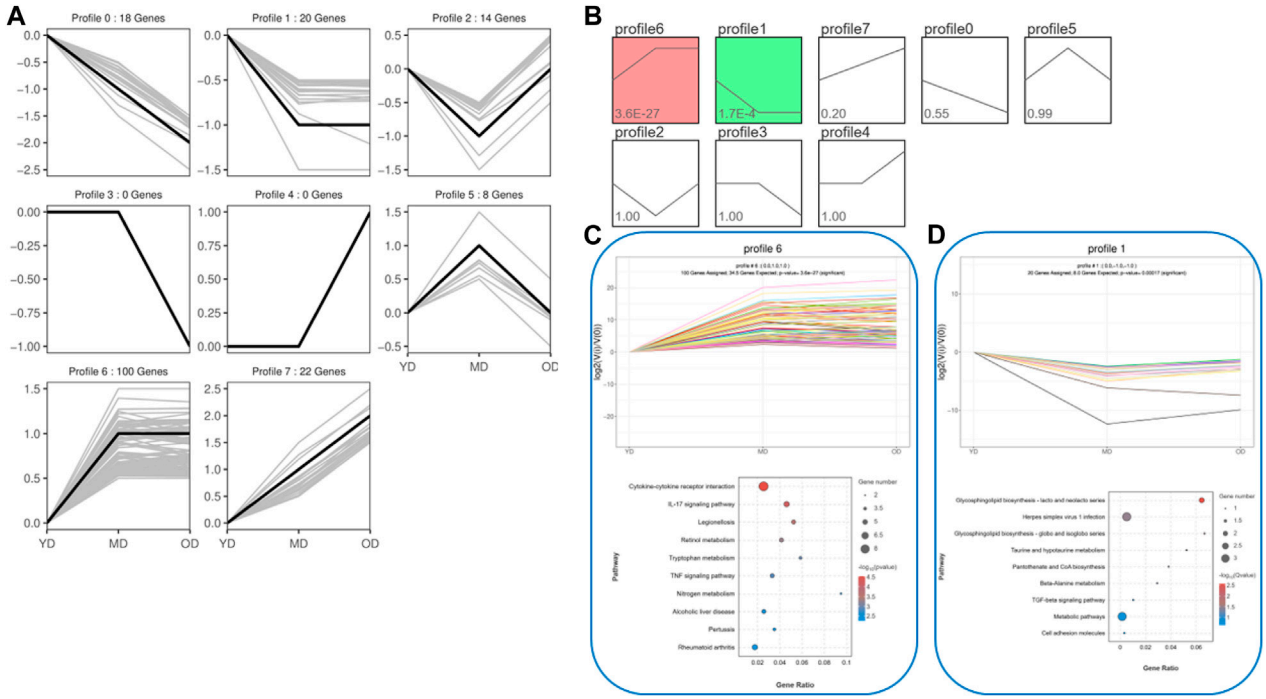



FIGURE 6
The trend analysis was performed on the target gene group. **(A)** Actual trend of gene mutations. **(B)** Trend graph fitted based on the predetermined number of trends and the genetic change trend. **(C)** KEGG functional analysis of profile 6 genes reveals the top 10 significantly enriched pathways. **(D)** KEGG functional analysis of profile 1 genes reveals the top 10 significantly enriched pathways. By conducting the aforementioned analysis on the DEMs, several candidate genes (*ELAPOR1, FOSL1, MEP1B, PAX9, PPP1R1B, SIX4, ZEP36, CD14, COL1A1, COL3A1, COL6A5, EGFL6, PGLYRP4, SERPINB13,* and *Spink6*) associated with protein synthesis in the skin were identified.

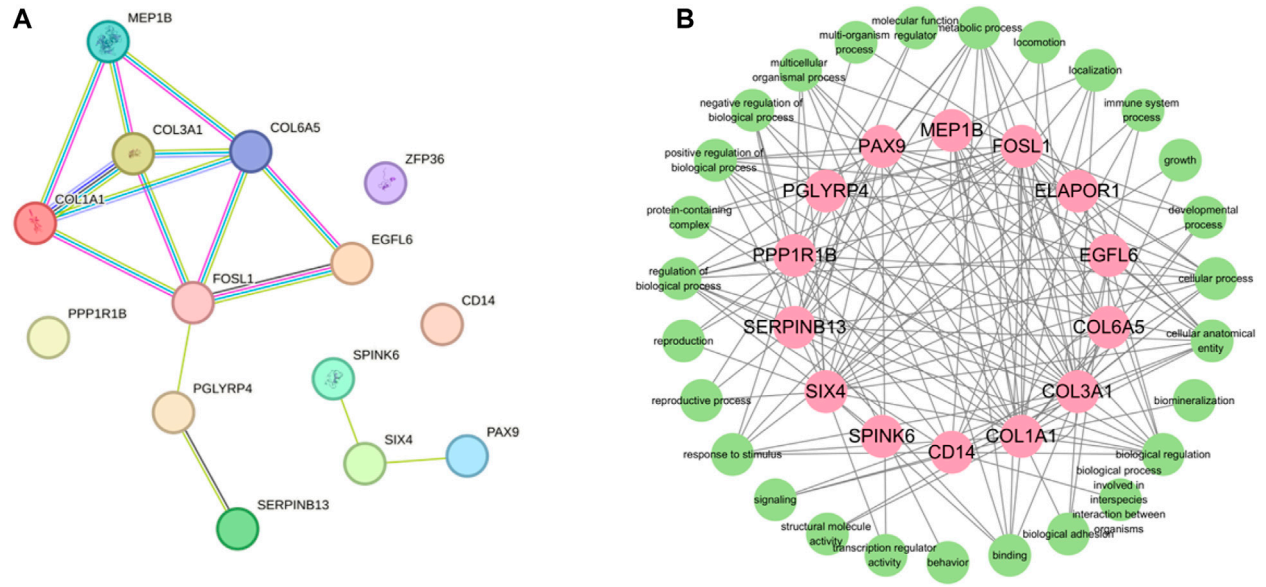



FIGURE 7
Protein-protein interactin (PPI)network. **(A)** PPI network was constructed by the STRING database using the protein synthesis candidate genes. **(B)** GO terms association map of protein synthesis candidate genes.

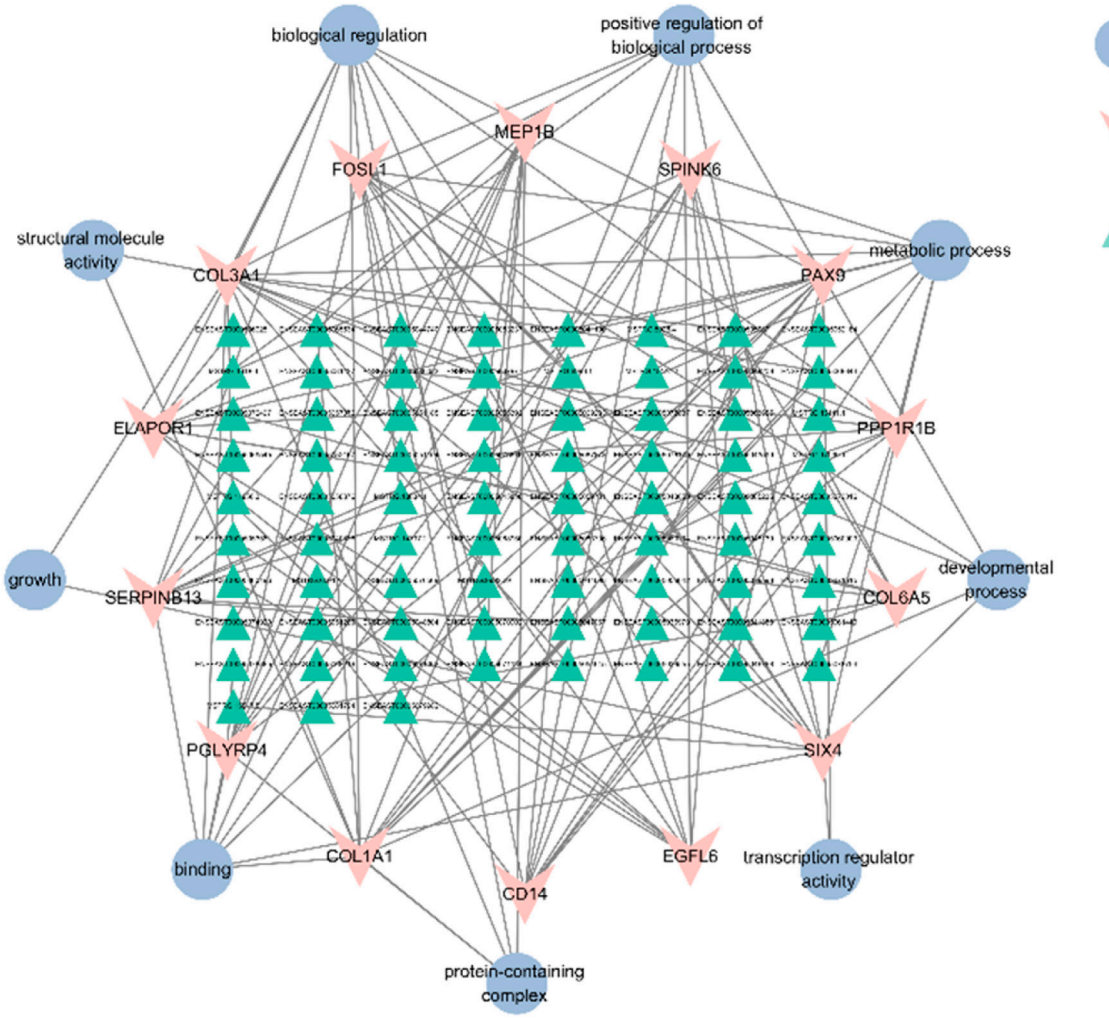



FIGURE 8
The interaction networks of "pathways-differentially expressed lncRNAs-target genes". The triangle is for lncRNA, the "V" represents the target gene, and the ellipse is for the GO terms of the target gene, respectively. A total of 93 DELs were regulated with 14 mRNAs in the interactive network.

org/STRING, accessed on 10 April 2022). There were 14 nodes and 15 edges in the PPI network (Figure 7). According to the results of the PPI network analysis, *COL3A1* was the hup gene (degree = 14). In addition, PPI network analysis of candidate genes was significantly enriched in GO terms, including fibrillar collagen trimer and collagen-containing extracellular matrix. *FOSL1, PGLYRP4,* and *SIX4* were the key degrees of the PPI network with other proteins.

## 3.5 Analysis of the targeting relationship between lncRNAs and mRNAs

The DEMs and DELs with Pearson correlation coefficients above 0.95 were selected to construct the lncRNA-mRNA co-expression network using Cytoscape v.3.10.1. In this network, a total of 739 nodes and 2,428 lncRNA-mRNA pairs were identified based on the results (Supplementary Table S6). There were 75 lncRNAs corresponding to 14 target gene in Figure 8. Enrichment analysis using GO terms was conducted to understand the functional implications of these lncRNAs and their associated target genes. From Figure 8, it could be observed that the target genes were mainly enriched in protein synthesis and metabolic, such as binding, protein-containing complex, biological regulation, and metabolic process. This indicated that lncRNAs could interfere with target genes, thereby regulating protein synthesis function in donkey skin. LncRNAs played a crucial role in regulating the functions of some genes in different pathways. We observed that 15 target genes (*ELAPOR1, FOSL1, MEP1B, PAX9, PPP1R1B, SIX4, ZEP36, CD14, COL1A1, COL3A1, COL6A5, EGFL6, PGLYRP4, SERPINB13,* and *Spink6*.) were regulated by lncRNAs to function in protein synthesis and metabolic, although these DEMs participated in other pathways. The collagen proteins encoded by *COL1A1* and *COL3A1* were involved in the synthesis and repair processes of the extracellular matrix in the extracellular matrix remodeling pathway. Enrichment analysis of lncRNA-related mRNAs was performed through GO terms to understand the functions of these lncRNAs and target genes. It can be seen from Figure 8, the target genes are mainly enriched in protein synthesis and metabolic. This indicates that lncRNA can interfere with target

TABLE 2 Selected Pathways/Biological processes associated with metabolism, collagen organization and skin development.

| Pathways | Genes count | *p*-value | Genes name |
|---|---|---|---|
| Protein digestion and absorption | 4 | 1.72E-04 | COL1A1, COL3A1, MEP1B, COL6A5 |
| Relaxin signaling pathway | 3 | 0.008505048 | COL1A1, COL3A1, RXFP1 |
| Metabolic pathways | 6 | 0.018227888 | B3GALT2, CTH, FMO2, FUT2, LOC106840130, GLUL |
| Glycosphingolipid biosynthesis - lacto and neolacto series | 2 | 0.032318823 | B3GALT2, FUT2 |
| Alanine, aspartate and glutamate metabolism | 2 | 0.034689964 | LOC106840130, GLUL |
| Cysteine and methionine metabolism | 2 | 0.038261718 | CTH, LOC106840130 |
| Biosynthesis of amino acids | 2 | 0.048155606 | CTH, GLUL |
| ECM-receptor interaction | 2 | 0.043557123 | COL1A1, COL6A5 |
| Biological Processes | | | |
| skin development | 2 | 0.005881289 | COL1A1, COL3A1 |
| collagen fibril organization | 2 | 0.007658935 | COL1A1, COL3A1 |
| cellular response to amino acid stimulus | 2 | 0.007658935 | COL1A1, COL3A1 |
| extracellular matrix organization | 2 | 0.025318973 | COL1A1, COL3A1 |

genes, thereby regulating the protein synthesis function in donkey skin.

## 4 Discussion

In this study we have considered rigorous exploration of the transcriptome expression profiles, encompassing both mRNA and lncRNA, within Dezhou donkey skin, with a particular focus on collagen deposition at distinct developmental stages. The employment of RNA-seq technology has facilitated this comprehensive investigation. This rigorous inquiry has culminated in the identification of a total of 182 differentially expressed genes that were commonly altered across the three experimental groups. Some of these identified genes are postulated to wield substantial influence in governing the process of skin collagen deposition and development. Collagen, being a ubiquitous protein constituent across mammalian species, assumes a multifaceted role, being deposited within diverse organs and tissues, while simultaneously playing a pivotal role in both structural and functional aspects (Franchi et al., 2007; Devos et al., 2023). Consistently, it has been well-established that the process of collagen deposition is orchestrated through a complex interplay of gene regulatory networks (Reilly and Lozano, 2021).

In this study, we also observed that the 182 DEMs exhibit significant enrichment across 168 distinct pathways. Notably, several of these pathways hold direct relevance to the intricate process of skin collagen deposition and development. Of these, Protein digestion and absorption, Metabolic pathways, PI3K-Akt signaling pathway, ECM-receptor interaction, and Relaxin signaling showed their pivotal roles in orchestrating collagen deposition within the dermal matrix, skin development, extracellular matrix and collagen fibril organization of Dezhou donkey skin (Table 2). Interestingly, a recent study found several important signaling pathways related to wool growth and bending, such as ECM-receptor interaction, PI3K-Akt signaling pathway, Relaxin signaling pathway, protein digestion and absorption, and metabolic pathways in Zhongwei goats (He et al., 2020; Liu et al., 2022). In addition, a recent experimental trial has shown that PI3K-Akt signaling pathway, Protein digestion and absorption and ECM–receptor interaction signaling pathways were associated with regulation of actin filament-based process, regulation of actin cytoskeleton organization, cell-matrix adhesion and collagen fibril

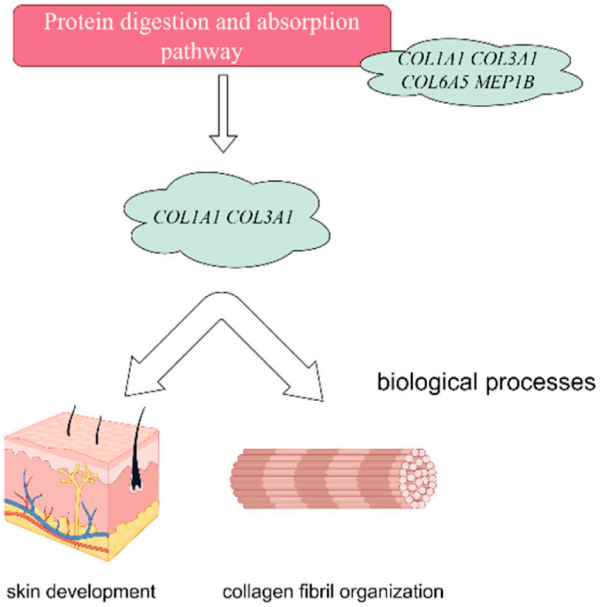



FIGURE 9
Collagen Alpha family genes regulate protein digestion and absorption signaling pathway. Furthermore, the protein digestion and absorption signaling pathway showed their association with skin development and collagen organization in Dezhou skin. The figure has generated via Figdraw.

organization (Wang and Gu, 2021). Furthermore, they assumed that the aging of skin might be associated with extracellular matrices depletion, which plays a key role in many cellular processes including migration, differentiation, and survival (Bonnans et al., 2014). Combined with published article data and our current findings it was proved that these pathways may exert vital roles in skin development, collagen and extracellular matrix organization.

In this investigation, we have delved into the intricate landscape of collagen synthesis and its associated genetic regulators. Specifically, *COL1A1*, *COL3A1*, *COL6A5* and *LOC106840130* have been identified as pivotal players in regulation of important pathways including ECM-receptor interaction, Relaxin signaling pathway, and protein digestion and absorption and biological function processes (skin development, extracellular matrix and collagen fibril organization) as shown in Figure 9. Collagen, a prominent member of the extracellular matrix (ECM), is characterized by its triple-helix structure (Devos et al., 2023). Extensive research has showcased the widespread presence of collagen I, collagen II, and collagen III across various anatomical domains in numerous animal species, encompassing skin, skeletal structures, blood vessels, tendons, and internal organs (Rahkonen et al., 2004; Ben Amor et al., 2013; Wang et al., 2017). Notably, prior investigations have unveiled the profound impact of collagen peptide and vitamin C derivative co-treatment on upregulating *COL1A1* and *Has2* expression, thereby averting skin thinning (Shibuya et al., 2014). Furthermore, single nucleotide polymorphisms within the *COL3A1* and *COL6A5* genes have been associated with atopic dermatitis, revealing distinct genotype-phenotype relationships (Szalus et al., 2023). These genetic variations within these collagen-related genes may exert substantial influence on collagen expression, structure, and function, thereby modulating the pathogenesis and clinical manifestations of atopic dermatitis (Söderhäll et al., 2007). Consistently, it has been revealed that collagen alpha family genes play a key role in skin aging in human (Wang and Gu, 2021), and hair follicular stem cell development in goat (Reilly and Lozano, 2021) by regulating extracellular matrix and collagen organization respectively.

Concurrently, our study has unveiled the role of the *LOXL2* gene among the differentially expressed genes, shedding light on its involvement in collagen synthesis and modification. The multifaceted functions of *LOXL2* encompass gene transcription, cell motility, migration, adhesion, angiogenesis, and differentiation (Fujimoto and Tajima, 2009). Intriguingly, the present investigation has observed a significant downregulation of *COL1A1*, *COL3A1*, and *LOXL2* genes with increasing age, signifying the potential involvement of myriad molecular mechanisms, including epigenetic regulation, alterations in intracellular signaling pathways, and fluctuations in hormone levels (Wang et al., 2021; Wang and Gu, 2021). However, the precise mechanistic underpinnings necessitate further comprehensive research to validate and elucidate.

It is imperative to acknowledge that previous studies have elucidated the regulatory influence of lncRNAs in a cis manner, directly impacting gene expression (Zou et al., 2017; Shi et al., 2019b; Chen et al., 2019c; Zhu et al., 2020). In this study, DEL ENSEAST00005067788 was observed to exert an influence on *COL1A1* and *COL3A1* genes, with significantly higher expression levels in the YD period, consistent with the observed expression trends. This implies that lncRNA may modulate target gene expression by interfering with regulatory regions on the same chromosome. Particularly noteworthy is the involvement of other DELs, including ENSEAST00005041187, ENSEAST00005038497, MSTRG.17248.1, and other lncRNAs, in the cis-regulation of COL1A1 gene expression, emphasizing their potential roles in protein binding or the regulation of protein-containing complexes (Liu et al., 2016). Furthermore, the *LOXL2* gene, highly expressed in the YD period, was similarly cis-regulated by lncRNAs such as MSTRG.7943.1 and ENSEAST00005067788. Mutations within the *LOXL2* gene locus have been linked to decreased elastin renewal, underscoring its pivotal role in skin physiology (Gambichler et al., 2019). Moreover, hypermethylation of the *LOXL2* gene promoter region has been associated with elastolytic conditions, further highlighting its significance (Gambichler et al., 2016).

In summary, this research has elucidated the regulatory interplay between DELs and DEMs in the context of skin biology, with potential implications for collagen deposition in donkey skin. The identified genetic and epigenetic factors, including *COL1A1*, *COL3A1*, *LOXL2*, and various lncRNAs, may serve as key orchestrators in modulating the intricate processes underpinning collagen synthesis and deposition. This comprehensive investigation not only enhances our understanding of the molecular mechanisms governing collagen homeostasis but also paves the way for further research into therapeutic interventions and clinical applications in the field of dermatology and tissue engineering. Although our study provides important insights into the candidate genes related to collagen deposition in the skin of Dezhou donkeys, our study still has some limitations. First, our study mainly focuses on the analysis at the gene level, and we have not been able to deeply explore how these genes affect the deposition of collagen at the protein level. Secondly, our study was only conducted in Dezhou donkeys and did not consider other breeds of donkeys. For future research, we suggest further studying the role of these candidate genes at the protein level and their role in other donkey breeds.

## 5 Conclusion

In conclusion, the investigation into the transcriptome of Dezhou donkey skin during the Young Donkey (YD), Middle Donkey (MD), and Old Donkey (OD) periods has unveiled notable variations in the regulation of both mRNAs and lncRNAs specifically associated with collagen deposition. Our analysis, guided by GO annotations and KEGG, has highlighted a substantial proportion of DEGs such as coolagen alpha family genes *COL1A1*, *COL3A1*) and *LOXL2* are intricately linked to important signaling pathways (Protein digestion and absorption, PI3K-Akt signaling pathway, ECM-receptor interaction, and Relaxin pathways) and biological function processes (skin development, extracellular matrix and collagen fibril organization). In addition, Furthermore, the construction of an interaction network encompassing lncRNA genes has illuminated the potential roles of certain lncRNAs (ENSEAST00005041187, ENSEAST00005038497, MSTRG.17248.1) in modulating target genes (*COL1A*), thereby contributing to the intricate process of

collagen deposition within the skin. The findings presented herein not only expand our comprehension of the regulatory networks governing collagen organization and skin development but also furnish a foundational framework upon which further research endeavors in this domain can be grounded. This research, thus, plays a vital cornerstone for future explorations aimed at advancing our knowledge of skin biology and collagen synthesis and organization in Dezhou donkeys.

## Data availability statement

The data presented in the study are deposited in the SRA repository, accession number PRJNA1069658.

## Ethics statement

The animal study was approved by the Experiments and animals care in this study were conducted by the Animal Welfare and Ethics Committee of Institute of Animal Sciences, Liaocheng University (No. LC2019-1). The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

XW: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing–original draft, Writing–review and editing. YP: Data curation, Formal Analysis, Investigation, Software, Writing–review and editing. HL: Data curation, Formal Analysis, Investigation, Software, Writing–review and editing. MK: Data curation, Formal Analysis, Methodology, Writing–review and editing. WR: Data curation, Formal Analysis, Investigation, Software, Writing–review and editing. BH: Data curation, Investigation, Software, Writing–review and editing. YC: Data curation, Formal Analysis, Investigation, Software, Writing–review and editing. SX: Data curation, Software, Writing–review and editing. YZ: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. CW: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Shandong Province Modern Agricultural Industrial Technology System Project (grant number SDAIT-27), Construction of Molecular ID Card for Donkey and Camel Species, Livestock and Poultry Seed Industry Project of Ministry of Agriculture and Rural Affair (grant number 19211162), Key R&D Project of Shandong Province: Innovation and Demonstration of Key Technologies for the Integrated Development of Donga Black Donkey Industry (grant number 2021TZXD012) and the National Key R&D Program of China (grant number 2022YFD1600103).

## 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.2024.1335591/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. The Gene Ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Ben Amor, I. M., Roughley, P., Glorieux, F. H., and Rauch, F. (2013). Skeletal clinical characteristics of osteogenesis imperfecta caused by haploinsufficiency mutations in COL1A1. *J. Bone Mineral Res.* 28 (9), 2001–2007. doi:10.1002/jbmr.1942

Bonnans, C., Chou, J., and Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. *Nat. Rev. Mol. Cell Biol.* 15, 786–801. doi:10.1038/nrm3904

Chai, W., Qu, H., Ma, Q., Zhu, M., Li, M., Zhan, Y., et al. (2023). RNA-seq analysis identifies differentially expressed gene in different types of donkey skeletal muscles. *Anim. Biotechnol.* 34 (5), 1786–1795. doi:10.1080/10495398.2022.2050920

Chen, G., Cheng, X., Shi, G., Zou, C., Chen, L., Li, J., et al. (2019b). Transcriptome analysis reveals the effect of long intergenic noncoding RNAs on pig muscle growth and fat deposition. *BioMed Res. Int.* 2019, 2951427–2951515. doi:10.1155/2019/2951427

Chen, L., Shi, G., Chen, G., Li, J., Li, M., Zou, C., et al. (2019a). Transcriptome analysis suggests the roles of long intergenic non-coding RNAs in the growth performance of weaned piglets. *Front. Genet.* 10, 196. doi:10.3389/fgene.2019.00196

Chen, L., Shi, G., Chen, G., Li, J., Li, M., Zou, C., et al. (2019c). Transcriptome analysis suggests the roles of long intergenic non-coding RNAs in the growth performance of weaned piglets. *Front. Genet.* 10, 196. doi:10.3389/fgene.2019.00196

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–i890. doi:10.1093/bioinformatics/bty560

Devos, H., Zoidakis, J., Roubelakis, M. G., Latosinska, A., and Vlahou, A. (2023). Reviewing the regulators of COL1A1. *Int. J. Mol. Sci.* 24 (12), 10004. doi:10.3390/ijms241210004

Franchi, M., Trirè, A., Quaranta, M., Orsini, E., and Ottani, V. (2007). Collagen structure of tendon relates to function. *TheScientificWorldJournal* 7, 404–420. doi:10.1100/tsw.2007.92

Fujimoto, E., and Tajima, S. (2009). Reciprocal regulation of LOX and LOXL2 expression during cell adhesion and terminal differentiation in epidermal keratinocytes. *J. dermatological Sci.* 55 (2), 91–98. doi:10.1016/j.jdermsci.2009.03.010

Gambichler, T., Mahjurian-Namari, M., Reininghaus, L., Schmitz, L., Skrygan, M., Schulze, H. J., et al. (2019). Lysyl oxidase-like-2 mutations and reduced mRNA and protein expression in mid-dermal elastolysis. *Clin. Exp. dermatology* 44 (1), 47–51. doi:10.1111/ced.13652

Gambichler, T., Skrygan, M., Reininghaus, L., Schulze, H. J., Schaller, J., Hessam, S., et al. (2016). Lysyl oxidase-like 2 promoter hypermethylation in mid-dermal elastolysis. *Br. J. Dermatology* 175 (6), 1354–1356. doi:10.1111/bjd.14666

Goodrum, F., Theuri, S., Mutua, E., and Carder, G. (2022). The donkey skin trade: challenges and opportunities for policy change. *Glob. Policy* 13 (2), 304–309. doi:10.1111/1758-5899.13072

He, N., Su, R., Wang, Z., Zhang, Y., and Li, J. (2020). Exploring differentially expressed genes between anagen and telogen secondary hair follicle stem cells from the Cashmere goat (*Capra hircus*) by RNA-Seq. *PLoS One* 15 (4), e0231376. doi:10.1371/journal.pone.0231376

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

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

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi:10.1038/nmeth.1923

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* 12, 323–416. doi:10.1186/1471-2105-12-323

Li, Y., Ma, Q., Shi, X., Yuan, W., Liu, G., and Wang, C. (2022). Comparative transcriptome analysis of slow-twitch and fast-twitch muscles in dezhou donkeys. *Genes* 13 (9), 1610. doi:10.3390/genes13091610

Liu, J., Luo, C., Yin, Z., Li, P., Wang, S., Chen, J., et al. (2016). Downregulation of let-7b promotes COL1A1 and COL1A2 expression in dermis and skin fibroblasts during heat wound repair. *Mol. Med. Rep.* 13 (3), 2683–2688. doi:10.3892/mmr.2016.4877

Liu, Y., Ding, Y., Liu, Z., Chen, Q., Li, X., Xue, X., et al. (2022). Integration analysis of transcriptome and proteome reveal the mechanisms of goat wool bending. *Front. Cell Dev. Biol.* 10, 836913. doi:10.3389/fcell.2022.836913

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550–621. doi:10.1186/s13059-014-0550-8

Maigari, M., Dantani, U., Yelwa, M., and Ibrahim, A. (2020). Scavenging for Ejiao's raw material and the extinction of donkeys in Nigeria. *Glob. J. Sociol. Curr. Issues* 10 (2), 71–87. doi:10.18844/gjs.v10i2.5102

Mishra, S. K., and Wang, H. (2021). Computational analysis predicts hundreds of coding lncRNAs in zebrafish. *Biology* 10 (5), 371. doi:10.3390/biology10050371

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 (9), 1650–1667. doi:10.1038/nprot.2016.095

Pertea, M., Pertea, 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

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi:10.1093/bioinformatics/btq033

Rahkonen, O., Su, M., Hakovirta, H., Koskivirta, I., Hormuzdi, S. G., Vuorio, E., et al. (2004). Mice with a deletion in the first intron of the Col1a1 gene develop age-dependent aortic dissection and rupture. *Circulation Res.* 94 (1), 83–90. doi:10.1161/01.RES.0000108263.74520.15

Reilly, D. M., and Lozano, J. (2021). Skin collagen through the lifestages: importance for skin health and beauty. *Plast. Aesthetic Res.* 8 (2), 2. doi:10.20517/2347-9264.2020.153

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. methods* 9 (11), 1069–1076. doi:10.1038/nmeth.2212

Shi, G., Chen, L., Chen, G., Zou, C., Li, J., Li, M., et al. (2019a). Identification and functional prediction of long intergenic non-coding RNAs related to subcutaneous adipose development in pigs. *Front. Genet.* 10, 160. doi:10.3389/fgene.2019.00160

Shi, G., Chen, L., Chen, G., Zou, C., Li, J., Li, M., et al. (2019b). Identification and functional prediction of long intergenic non-coding RNAs related to subcutaneous adipose development in pigs. *Front. Genet.* 10, 160. doi:10.3389/fgene.2019.00160

Shibuya, S., Ozawa, Y., Toda, T., Watanabe, K., Tometsuka, C., Ogura, T., et al. (2014). Collagen peptide and vitamin C additively attenuate age-related skin atrophy in Sod1-deficient mice. *Biosci. Biotechnol. Biochem.* 78 (7), 1212–1220. doi:10.1080/09168451.2014.915728

Söderhäll, C., Marenholz, I., Kerscher, T., Rüschendorf, F., Esparza-Gordillo, J., Worm, M., et al. (2007). Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biol.* 5 (9), e242. doi:10.1371/journal.pbio.0050242

Song, C., Huang, Y., Yang, Z., Ma, Y., Chaogetu, B., Zhuoma, Z., et al. (2019). RNA-seq analysis identifies differentially expressed genes insubcutaneous adipose tissuein qaidamford cattle, cattle-yak, and angus cattle. *Animals* 9 (12), 1077. doi:10.3390/ani9121077

Szalus, K., Zysk, W., Gleń, J., Zabłotna, M., Nowicki, R. J., and Trzeciak, M. (2023). The associations of single nucleotide polymorphisms of the COL3A1, COL6A5, and COL8A1 genes with atopic dermatitis. *J. Personalized Med.* 13 (4), 661. doi:10.3390/jpm13040661

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

Wang, C., Li, H., Chen, K., Wu, B., and Liu, H. (2017). Association of polymorphisms rs1800012 in COL1A1 with sports-related tendon and ligament injuries: a meta-analysis. *Oncotarget* 8 (16), 27627–27634. doi:10.18632/oncotarget.15271

Wang, C., Li, H., Guo, Y., Huang, J., Sun, Y., Min, J., et al. (2020). Donkey genomes provide new insights into domestication and selection for coat color. *Nat. Commun.* 11 (1), 6014. doi:10.1038/s41467-020-19813-7

Wang, J., and Gu, H. (2021). Identification of biological processes and potential inhibitors for aging skin. *Res. Square* 24. doi:10.21203/rs.3.rs-1104916/v1

Wang, J., Sui, J., Mao, C., Li, X., Chen, X., Liang, C., et al. (2021). Identification of key pathways and genes related to the development of hair follicle cycle in cashmere goats. *Genes* 12 (2), 180. doi:10.3390/genes12020180

Wang, M., Li, H., Zhang, X., Yang, L., Liu, Y., Liu, S., et al. (2022). An analysis of skin thickness in the Dezhou donkey population and identification of candidate genes by RNA-seq. *Anim. Genet.* 53 (3), 368–379. doi:10.1111/age.13196

Wickramasinghe, S., Cánovas, A., Rincón, G., and Medrano, J. F. (2014). RNA-sequencing: a tool to explore new frontiers in animal genetics. *Livest. Sci.* 166, 206–216. doi:10.1016/j.livsci.2014.06.015

Yan, S. U., Li, Y. H., Zhao, C. H., Jun, T. E., Wang, Y. H., Wang, T. Q., et al. (2023). Genome-wide association study for numbers of vertebrae in Dezhou donkey population reveals new candidate genes. *J. Integr. Agric.* 22 (10), 3159–3169. doi:10.1016/j.jia.2023.04.038

Yang, Y. L., Rong, Z., and Kui, L. (2017). Future livestock breeding: precision breeding based on multi-omics information and population personalization. *J. Integr. Agric.* 16 (12), 2784–2791. doi:10.1016/s2095-3119(17)61780-5

Zhu, Z., Ma, Y., Li, Y., Li, P., Cheng, Z., Li, H., et al. (2020). The comprehensive detection of miRNA, lncRNA, and circRNA in regulation of mouse melanocyte and skin development. *Biol. Res.* 53 (1), 4–14. doi:10.1186/s40659-020-0272-1

Zou, C., Li, S., Deng, L., Guan, Y., Chen, D., Yuan, X., et al. (2017). Transcriptome analysis reveals long intergenic noncoding RNAs contributed to growth and meat quality differences between yorkshire and wannanhua pig. *Genes* 8 (8), 203. doi:10.3390/genes8080203

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 February 2024
DOI 10.3389/fgene.2024.1351429



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Paolo Zambonelli,
University of Bologna, Italy
Krishnamoorthy Srikanth,
Cornell University, United States

*CORRESPONDENCE
Ziwei Xu,
zjsnkyxzw@163.com
Wangjun Wu,
wuwangjun2012@njau.edu.cn

RECEIVED 06 December 2023
ACCEPTED 26 January 2024
PUBLISHED 13 February 2024

CITATION
Wang B, Hou L, Yang W, Men X, Qi K, Xu Z and Wu W (2024), Construction of a co-expression network affecting intramuscular fat content and meat color redness based on transcriptome analysis.
*Front. Genet.* 15:1351429.
doi: 10.3389/fgene.2024.1351429



# Construction of a co-expression network affecting intramuscular fat content and meat color redness based on transcriptome analysis

Binbin Wang[1], Liming Hou[2], Wen Yang[2], Xiaoming Men[1], Keke Qi[1], Ziwei Xu[1]* and Wangjun Wu[2]*

[1]Institute of Animal Husbandry and Veterinary, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [2]College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, China

**Introduction:** Intramuscular fat content (IFC) and meat color are vital indicators of pork quality.

**Methods:** A significant positive correlation between IFC and redness of meat color (CIE *a** value) indicates that these two traits are likely to be regulated by shared molecular pathways.To identify candidate genes, hub genes, and signaling pathways that regulate these two traits, we measured the IFC and CIE *a** value in 147 hybrid pigs, and selected individuls with extreme phenotypes for transcriptome analysis.

**Results:** The results revealed 485 and 394 overlapping differentially expressed genes (DEGs), using the DESeq2, limma, and edgeR packages, affecting the IFC and CIE *a** value, respectively. Weighted gene co-expression network analysis (WGCNA) identified four modules significantly correlated with the IFC and CIE *a** value. Moreover, we integrated functional enrichment analysis results based on DEGs, GSEA, and WGCNA conditions to identify candidate genes, and identified 47 and 53 candidate genes affecting the IFC and CIE *a** value, respectively. The protein protein interaction (PPI) network analysis of candidate genes showed that 5 and 13 hub genes affect the IFC and CIE *a** value, respectively. These genes mainly participate in various pathways related to lipid metabolism and redox reactions. Notably, four crucial hub genes (*MYC, SOX9, CEBPB, and PPAGRC1A*) were shared for these two traits.

**Discussion and conclusion:** After functional annotation of these four hub genes, we hypothesized that the *SOX9/CEBPB/PPARGC1A* axis could co-regulate lipid metabolism and the myoglobin redox response. Further research on these hub genes, especially the *SOX9/CEBPB/PPARGC1A* axis, will help to understand the molecular mechanism of the co-regulation of the IFC and CIE *a** value, which will provide a theoretical basis for improving pork quality.

KEYWORDS

**intramuscular fat content, meat color redness, RNA-seq, functional enrichment analysis, hub gene**

# 1 Introduction

Pork is a significant and extensively utilized animal resource that has emerged as a principal protein source within human diets. In recent years, China's yearly pork production has surpassed 50 million tons. Duroc × (Landrace × Yorkshire) (DLY) pigs account for over 90% of the pork market due to their rapid growth and high lean meat rate (Duan et al., 2023). With improved living standards, high-quality pork has become more popular among consumers. Meat quality is a crucial indicator for assessing pork production and quality. Essential indicators of meat quality include intramuscular fat content (IFC), meat color, tenderness, and drip loss, which can directly impact pork quality and market competitiveness (Moeller et al., 2010). Consumers favor snowflake meat (a reflection of high IFC or marbling), and IFC deposition is the main cause of snowflake meat (Liu et al., 2020). Meat color is also one of the most direct sensory indicators of pork quality for consumers and directly affects their consumption behavior. In the food industry, the most popular numerical colour space system is the L* (lightness), b* (yellowness) and a* (redness), which is also referred as the CIELAB system, originally defined by the CIE (CIE, 1986). The subjective color scores of the meat showed a stronger correlation with the CIE $a^*$ value ($R = 0.80$) in one study (Sun et al., 2016). Hence, the quality of pork color could be directly assessed based on the CIE $a^*$ value. Despite DLY pork effectively meeting the quantitative demand, its muscle quality falls short of eliciting satisfaction. Both the IFC and CIE $a^*$ value are traits with relatively high heritability (Cabling et al., 2015; Wang et al., 2022) and are the most intuitive indicators of high-quality pork. Consequently, increasing the IFC and CIE $a^*$ value through genetic improvement is a major research focal point for pig breeding enterprises.

IFC refers to the amount of fat that accumulates between muscle fibers or within muscle cells, mainly composed of phospholipids and triglycerides (Shi-Zheng and Su-Mei, 2009). It is widely accepted that changes in meat color in muscles are due to changes in myoglobin levels. This may be due to higher myoglobin levels in slow/oxidative myofibers (red muscle fibers) than in fast/glycolytic myofibers (white muscle fibers). When there is a high proportion of red muscle fibers in muscle tissue, its muscle color exhibits a more distinct red characteristic (Kim et al., 2010). This phenomenon is closely related to the biochemical markers of meat, such as the oxidation state, cytochrome content, and redox forms. Previous studies have shown a significant correlation ($R = 0.260–0.323$) between IFC and CIE $a^*$ (Mortimer et al., 2014; Zhang et al., 2022). Therefore, we speculated that these two traits might have similar genetic backgrounds, but the underlying genetic basis was largely unknown.

Differences in phenotype are caused by a variety of factors, among which changes in gene expression are crucial. Therefore, the variations in the IFC and CIE $a^*$ value within a population might be driven by differences in the expression levels of critical genes involved in regulating these two traits. With the development of next-generation sequencing technologies, the emergence of transcriptome sequencing (RNA-seq) allowed us to detect the expression levels of all genes across the entire genome. Researchers usually use individuals with extreme phenotypes of the IFC and $a^*$ value to perform RNA-seq, allowing them to obtain many candidate genes and signaling pathways related to the IFC and CIE $a^*$ value (Cardoso et al., 2017; Xing et al., 2021; Fernndez-Barroso et al., 2022).

However, organisms are complex systems with interconnected genes regulating biological activities, forming intricate network systems. Therefore, it is crucial to consider the interrelationships between thousands of genes when studying phenotypic variation. Differential expression analysis may not capture critical biological pathways or gene-gene interactions relevant to target traits, as it focuses on the impact of individual genes rather than the influence of gene networks (Xing et al., 2021). Coexpressed genes often form densely connected subgraphs in networks, representing functionally related gene groups or signaling pathways, and exhibit specific biological functions by developing local substructure modules (Barabasi and Oltvai, 2004). These modules reveal interactions among genes at a systems level, aiding researchers in further understanding the mechanisms underlying gene interactions and identifying regulatory hubs of coexpressed genes (Talukdar et al., 2016). Weighted gene co-expression network analysis (WGCNA) is an efficient and accurate method for describing the correlation among all genes or modules within the whole genome with traits. It is particularly advantageous for simultaneously identifying key genes of multiple complex traits (Zhang and Horvath, 2005), such as fat deposition (Xing et al., 2021), meat quality (Zhao et al., 2020), and reproductive performance (Wu et al., 2022).

Based on transcriptomic data, the present study aimed to gain molecular insights into the hub genes and metabolic pathways that coregulate the variations in the IFC and CIE $a^*$ value. We collected individuals with divergent IFC and CIE $a^*$ values for RNA-seq. Subsequently, we identified the differentially expressed genes (DEGs), and performed gene set enrichment analysis (GSEA), WGCNA, and protein protein interaction (PPI) analysis. We identified the candidate genes and modules significantly related to these two traits. Through systematic integration of the above results, we identified the hub genes and pathways that could co-regulate the changes in the IFC and CIE $a^*$ values. These findings contribute to understanding the genetic mechanisms of co-regulation changes in the IFC and CIE $a^*$ value. Moreover, the identified hub genes may serve as potential biomarkers for the synergistic improvement of IFC and meat color in pigs.

# 2 Materials and methods

## 2.1 Animals, sample collection, and phenotype measurement

A total of 147 commercial DLY pigs, consisting of 70 castrated boars and 77 females, were selected for this study. The experimental pigs were reared under standardized indoor conditions and provided *ad libitum* access to feed and water at Jiangsu Kangle Pig Breeding Farm (Changzhou, China). All experimental protocols involving animals were approved by the Nanjing Agricultural University Animal Care and Use Committee (Certification No.: SYXK (Su) 2022–0031). These pigs were slaughtered in six batches at the same slaughterhouse within a month, with 20–30 pigs slaughtered in each batch, with an average live weight of 122.49 ± 16.54 kg (mean ± standard deviation). Following slaughter, LD muscle from the last third and fourth thoracic vertebrae was collected for each pig. Approximately 0.5 g of LD muscle was placed into a 1.5 mL tube and frozen at −80 °C for RNA extraction. Another portion of LD muscle was trimmed to 1 cm × 1 cm × 2 cm along the fiber direction and fixed in 4% paraformaldehyde solution. The meat color redness

value of the LD muscle was assessed three times at 24 h post-mortem using a CR-410 hand-held colorimeter (Kinica Minolta Sensing Inc., Shanghai, China). The mean of the three measurements was the final CIE $a^*$ value. Approximately 300 g of LD muscle was utilized for determining IFC using the Soxhlet extraction method (Supakankul and Mekchay, 2016).

## 2.2 Sample selection

In order to avoid the influence of sex and carcass weight on the selected samples, a general linear model in SAS software was used to analyze the factors affecting the IFC and redness values in 147 DLY pigs. The results showed that sex and carcass weight did not affect IFC and CIE $a^*$ values. Therefore, based on the extreme values of IFC and CIE $a^*$ values, we selected the high IFC group (H_IFC, n = 6), low IFC group (L_IFC, n = 6), high CIE $a^*$ group (H_$a^*$, n = 6), and low CIE $a^*$ group (L_$a^*$, n = 6), respectively. During the selection process, we found that there were 2 samples overlapping between the H_IFC group and the H_$a^*$ group, and 3 samples overlapping between the L_IFC group and the L_$a^*$ group. So, 19 unique samples were used for transcriptome analysis in this study. The means of the IFC and CIE $a^*$ value in the high and low groups were calculated using the two-tailed Student's t-test. Besides, we also calculated the differences of the samples in the H_IFC and H_$a^*$ value groups (H_group, n = 10) and the samples in the L_IFC and L_$a^*$ groups (L_group, n = 9) using the two-tailed Student's t-test. All analyses were conducted using SPSS (v22.0) software (SPSS Inc., Chicago, IL, United States).

## 2.3 Haematoxylin–eosin staining

Selected LD samples were fixed in 4% paraformaldehyde for 24 h at room temperature. Muscle tissue was dehydrated using ethanol, transparently treated with xylene, embedded in paraffin, and cut into 3–4 μm samples for further haematoxylin–eosin (H&E) staining. Sections were deparaffinized in xylene, rehydrated in ethanol and stained with hematoxylin for 10 min. The sections were then rinsed in tap water and stained with eosin for 1 min, dehydrated, transparently treated with xylene and finally sectioned using neutral gum. The prepared sections were observed under the microscope, in which the nuclei and cytoplasm of the muscle cells appeared blue and light red, respectively, and the adipocytes appeared white.

## 2.4 RNA extraction, library construction, and sequencing

Total RNA was extracted from 100 mg of frozen LD muscle using TRIzol reagent (Invitrogen, Carlsbad, CA, United States). The total RNA was quantified and quality controlled using Qubit 2.0 and Agilent 2,100. RNA with an RNA integrity number (RIN) of >7 and RNA quality rating of "A" was used for RNA library construction. RNA libraries were constructed using the VAHTS® universal V8 RNA-seq Library Prep Kit for Illumina (Vazyme, China) according to the manufacturer's instructions. The Illumina NovaSeq 6,000 platform (Illumina, San Diego, CA, United States) was used for transcriptome sequencing based on the high-quality RNA library, and the sequencing read length was paired-end 150 bp. The obtained raw data were filtered to clean data with FastQC (v0.11.5) and Trimmomatic (v0.38) software (Bolger et al., 2014) by removing reads containing adapters, low-quality reads, and reads with an N content of >5%. The sequencing depth of transcriptome data in this study exceeded 40 million reads per sample. The average sequencing depth of the clean reads used for subsequent analysis was 42.91 million reads. The alignment analysis results showed that the average unique mapping rate was 87.53%. The clustering heatmaps between samples showed significant stratification between high and low groups (Supplementary Figure S1). Overall, the sequencing data exhibited high quality, rendering it suitable for subsequent analyses.

## 2.5 Identification of DEGs

The obtained clean reads were mapped to the *Sus scrofa* 11.1 genome from Ensembl 101 using STAR (v2.7.2) software (Dobin et al., 2013) with settings (–sjdbOverhang 135). Finally, a transcriptome gene expression count file was converted using featureCounts (v2.0.0) software (Liao et al., 2014). The DESeq2 (v1.25.9) (Love et al., 2014), limma (Ritchie et al., 2015), and edgeR packages in R (v4.1) (Robinson et al., 2010) software were used to identify DEGs between the groups. DEGs were defined as those with a false discovery rate (FDR) of <0.05 and $|\log2FoldChange| \geq 1$. Furthermore, overlapping DEGs detected by the DESeq2, limma, and edgeR packages were considered true DEGs, and used for subsequent functional enrichment analysis.

## 2.6 Functional annotation and enrichment analysis

To better understand the functions of overlapping DEGs, the R package BioMart (Haider et al., 2009) was used to annotate genes using the reference genome *Sus scrofa* 11.1. The Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of overlapping DEGs were subjected to functional enrichment analysis using the R package clusterProfiler (v4.6.2) (Wu et al., 2021) with the following default parameters: ont = "ALL", nPerm = 1,000, pAdjustMethod = "BH", minGSSize = 10, maxGSSize = 500. In addition, we removed redundancy from the GO terms using the 'simplify' function in the clusterProfiler package, with the following default parameters: cutoff = 0.7, by = "p.adjust", select_fun = min. The overlapping DEGs were visualized as a heatmap plot using the R function heatmap. Additionally, considering that GSEA does not require an arbitrary cutoff for differential gene expression and has a more extensive functional range, we also used GSEA on our datasets based on whole genes of the IFC and CIE $a^*$ groups, using the clusterProfiler package (v4.6.2) (Wu et al., 2021) with the above default parameters. The threshold of significantly enriched GO terms and KEGG pathways was a $q$ value of <0.10.

## 2.7 WGCNA

To construct a co-expression network, we used WGCNA, a package from R (1.72.1) (Langfelder and Horvath, 2008), with RNA-

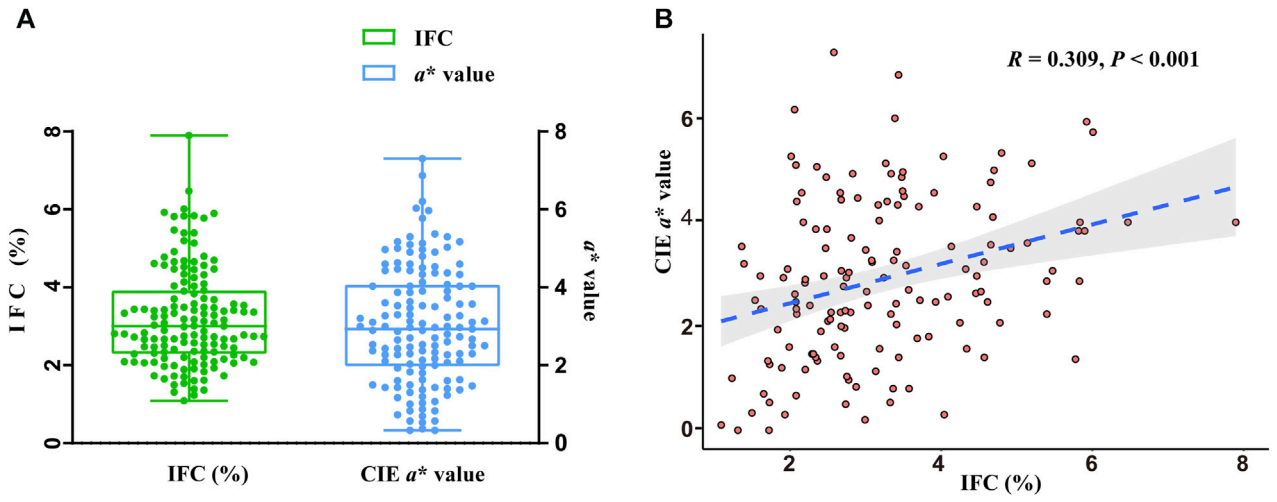



FIGURE 1
Statistical analysis of intramuscular fat content (IFC) and meat color redness (CIE *a**) values. **(A)** Phenotypic values of the IFC and CIE *a** value. **(B)** Correlation analysis of the IFC and CIE *a** value.

seq data (n = 19), with their counts normalized by transcript per million (TPM). After the expression matrix input, genes with TPM values of >1 in more than 10 individuals were selected for a coexpression network setting. The clean expression matrix underwent hierarchical clustering using the group average method to identify outliers, which were samples deviating significantly from the others. There were no outliers in this study, and the final expression matrix contained 10,512 genes and 19 individuals for establishing an unsigned coexpression network based on the step-by-step method.

This study selected a power value of 18 based on the scale-free topology criterion, resulting in a scale-free topology index ($R^2$) of 0.90. The hybrid dynamic tree-cutting approach employs a minimum module size of 30 as the default and commonly used value. To characterize the module expression, module eigengenes (MEs) were calculated as the first principal component of the expression matrix. The WGCNA approach facilitates the identification of biologically significant modules and potential critical modules for further analysis by defining the module trait relationships (MTRs) and gene significance (GS) of each module. The mean value of GS for the genes within a module represented the module significance (MS). To select candidate modules for functional enrichment analysis, modules with MTRs greater than 0.35 and MS exceeding 0.25 were considered based on the criteria reported in previous studies. The GO and KEGG pathway terms of all genes within the critical module were subjected to functional enrichment analysis using the clusterProfiler package (v4.6.2) (Wu et al., 2021) with the above default parameters.

## 2.8 Identification of candidate and hub genes related to the IFC and CIE *a** value

To further identify candidate genes affecting the IFC and CIE *a** value, we performed overlap analysis of significantly enriched GO terms and KEGG pathways in Omicshare platform (https://www.omicshare.com/) derived from overlapping DEGs, GESA, and WGCNA, respectively. The results of the overlap analysis are presented in the Venn network diagram. The selected GO terms and KEGG pathways had a *q* value of <0.1 in all three methods and less than 0.05 in at least two methods. DEGs located in the overlapping GO terms and KEGG pathways were considered candidate genes and used for subsequent PPI analysis.

The construction of a PPI network was employed to analyze the interactions between genes encoding proteins in candidate genes based on the Search Tool for the Retrieval of Interacting Genes (STRING) database (v11.5) (Szklarczyk et al., 2015). Cytoscape software (v3.8.0) (Shannon et al., 2003) was employed to visualize the entire PPI network. This analysis allowed the connection patterns between genes in PPI networks to be explored and visualized. Highly connected genes, also known as hub genes, may play an essential role in influencing the target traits of these candidate genes. The criterion for selecting the hub gene was that the degree of connectivity was greater than 10.

# 3 Results

## 3.1 Phenotypes and sequencing data

The phenotypes of the IFC and CIE *a** value in 147 DLY pigs are shown in Figure 1A. The mean and standard error of the IFC and CIE *a** value were 3.20% ± 0.10% and 2.86% ± 0.13%, respectively. The IFC and CIE *a** value showed a significant positive correlation in 147 DLY pigs ($R$ = 0.309, $p$ < 0.001) (Figure 1B).

Based on the IFC and CIE *a** value, the LD muscle samples were divided into the high IFC (H_IFC, n = 6), low IFC (L_IFC, n = 6), high CIE *a** value (H_*a**, n = 6), and low CIE *a** value (L_*a**, n = 6) groups. The phenotypic values of selected individuals are shown in Figure 2 and Supplementary Table S1. The mean IFCs of the high and low groups were 5.92% and 1.45%, respectively. The mean CIE *a** values of the high and low groups were 4.30 and 1.72, respectively.

The IFC and CIE *a** value in the high groups (H_IFC and H_*a**) were significantly higher than in the low groups (L_IFC and L_*a**) (Figures 2C, D). Moreover, the phenotypic information of the samples in the H_IFC and H_*a** groups (H_group, n = 10) and the samples in the L_IFC and L_a* groups (L_group, n = 9) was counted, and the results showed that the IFC and CIE a* values in the H_group were 5.30% ± 0.91% and 4.85 ± 1.27, respectively, and were 1.75% ± 0.61% and 1.37 ± 1.37, respectively, in the L_ group. The IFC and CIE a* values were significantly higher in the H_group than in the L-group (Figure 2E). In addition, the results of general linear model analysis indicated that sex and carcass weight had no significant impact on the IFC and CIE *a** values (Table 1).

Concerning the RNA-Seq data, 37.48–50.63 million raw reads per sample were generated. After filtering approximately 1.39% of the raw reads, an average of 42.91 million clean reads were used for the following analysis. The mean Q30 and GC percentage values of these clean data were 95.19% and 52.53%, respectively. After alignment using STAR software, 87.53% of the clean reads were uniquely mapped to the *Sus scrofa* 11.1 genome (Supplementary Table S1). Before DEG detection, low expression levels or non-expressed genes were removed based on gene expression counts. The remaining 16,453 genes for IFC and 16,249 for CIE *a** were analyzed in the differential expression analysis.

## 3.2 DEGs

The present study identified 723, 569, and 608 DEGs between the H_IFC and L_IFC groups using DESeq2, limma, and edgeR, respectively (Figure 3A). A total of 485 overlapping DEGs were detected, including 190 upregulated and 295 downregulated DEGs in the H_IFC group, respectively. For the CIE *a** value, 590, 481, and 455 DEGs were identified using DESeq2, limma, and edgeR, respectively (Figure 3C). Three hundred and ninety-four DEGs were shared among the three methods, including 153 upregulated and 241 downregulated DEGs in the H_CIE *a** group. Figures 3B,D exhibit the heatmap of these overlapping DEGs, from which it can be seen that the expression patterns of overlapping DEGs were consistent within groups and different between groups. Moreover, 201 DEGs were shared between these two traits.

## 3.3 Functional enrichment analysis

There were 106 significantly enriched GO (GO_DEGs) terms (Supplementary Table S4; Figure 4A) and 20 significantly enriched KEGG (KEGG_DEGs) pathways (Supplementary Table S5; Figure 4B) based on overlapping DEGs between the H_IFC and L_IFC groups. Among these 106 enriched GO_DEGs terms, most belonged to the biological process (BP) category, and only 1 and 6 terms belonged to the cellular component (CC) and molecular function (MF) categories, respectively. In terms of KEGG_DEGs pathways, more than half of the 20 significantly enriched pathways were closely associated with lipid metabolism and lipolysis, such as the adipocytokine signaling pathway (ssc04920), MAPK signaling pathway (ssc04010), PI3K-Akt signaling pathway (ssc04151) and regulation of lipolysis in adipocytes (ssc04923). For the CIE *a** value, 138 significantly enriched GO_DEGs terms (Supplementary Table S6; Figure 4C) and 22 significantly enriched KEGG_DEGs pathways (Supplementary Table S7; Figure 4D) were detected. Similarly, most of these enriched GO_DEGs terms belonged to the BP category. KEGG_DEGs enrichment analysis revealed that 9 of 12 significant pathways were strongly associated with redox and antioxidant responses, such as the insulin signaling pathway (ssc04910), AMPK signaling pathway (ssc04152), FoxO signaling pathway (ssc04068), adipocytokine signaling pathway (ssc04920), and MAPK signaling pathway (ssc04010). Furthermore, 12 of these 22 significantly enriched pathways were shared with the significantly enriched pathways found in the IFC group. This suggests that there was some similarity in the genetic background between the IFC and CIE *a** value.

To further understand the mechanisms of genetic differences between the high and low groups, GSEA was used. The results showed that 168 significantly enriched GO_GSEA terms (Supplementary Table S8) and 61 significantly enriched KEGG_ GSEA pathways (Supplementary Table S9) were identified between the H_IFC and L_IFC groups. Among these enriched GO_GSEA terms, the top five were related to mitochondrial metabolism and organismal oxidoreductase activity. In terms of KEGG_GSEA, several significant pathways associated with lipid and fatty acid metabolism were enriched, such as oxidative phosphorylation (ssc00190), fatty acid metabolism (ssc01212), the adipocytokine signaling pathway (ssc04920), and ether lipid metabolism (ssc00565). For the CIE *a** value, 390 significantly enriched GO_ GSEA terms Supplementary Table S10) and 76 significantly enriched KEGG_GSEA pathways (Supplementary Table S11) were identified between the H_*a** and L_*a** groups. Redox reactions are an essential factor influencing the CIE *a** value; the top five significantly enriched GO_GSEA terms were mainly related to the cellular response to an organic substance, oxidoreductase activity, and positive regulation of the developmental process. KEGG_GSEA results showed that more than 60% of the significantly enriched pathways in the H_*a** and L_*a** groups were consistent with those significantly enriched in the high and low IFC groups. These overlapping pathways included the above-mentioned lipid metabolic pathways, such as ssc00190, ssc01212, and ssc00565. These results suggested that lipid and fatty acid metabolism are essential factors influencing changes in the CIE *a** value.

## 3.4 Co-expressed gene modules associated with the IFC and CIE *a** value

The expression matrix containing 10,512 genes from 19 individuals was used for WGCNA. Hierarchical cluster analysis revealed no outliers among the19 samples (Supplementary Figure S2A). To build a scale-free network, we chose a soft threshold of = 18, with a scale-free topology fitting index $R^2$ of >0.90 (Supplementary Figure S2B). In this study, nine gene coexpression modules were identified (Figure 5A). The module with the minimum number of genes among these modules was the dark orange module, containing 82 genes, while the maximum number of genes was in the dark red module, including 4,367 genes (Figure 5B). Correlation analysis between module eigengene and the IFC or CIE *a** value was performed, and four modules, including purple, dark

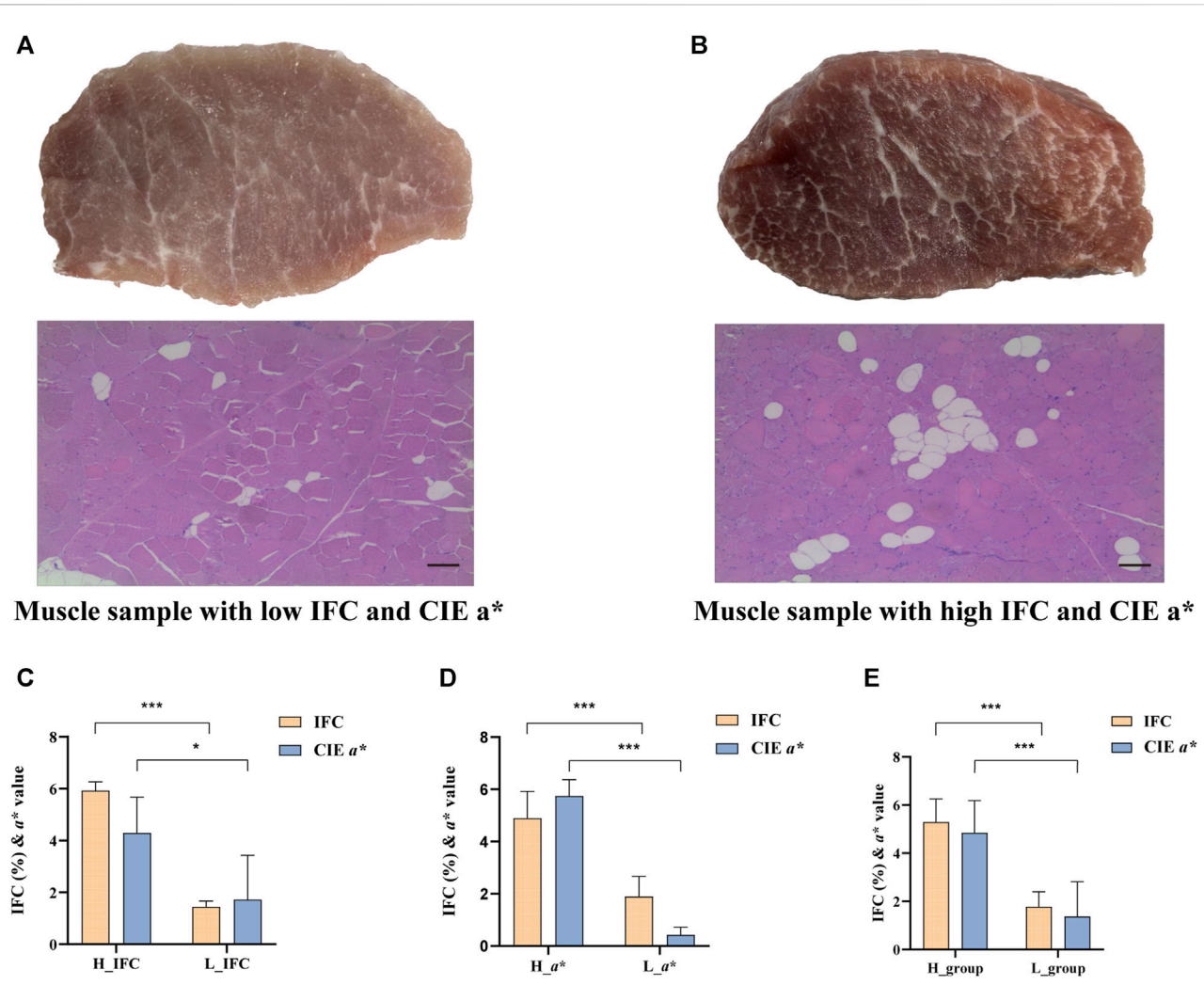



FIGURE 2
IFC and CIE CIE *a** value comparison between the high and low groups. **(A)** Representative plots of latissimus dorsi (LD) tissue and H&E staining of shared samples from the low IFC group and low CIE *a** group, scale bar = 100 μm. **(B)** Representative plots of LD tissue and H&E staining of shared samples from the high IFC group and high CIE *a** group, scale bar = 100 μm. **(C–E)** Comparasion of the IFC and CIE *a** value in different groups. The H_ group represents the sample combination of H_IFC and H_*a**, n = 10; The L_group represents the sample combination of the L_IFC and L_*a** value group, n = 9. Error bars represent the standard deviation (SD), where yellow bars represent the IFC and blue bars represent the CIE *a** value. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, two-tailed Student's t-test.

**TABLE 1 Influencing factors of intramuscular fat content (IFC) and redness value in 147 DLY pigs.**

| Trait | IFC | CIE *a** value |
|---|---|---|
| Sex | NS | NS |
| Carcass weight | NS | NS |

NS, not significant.

grey, dark red, and black, were significantly correlated with the IFC and CIE *a** value (Figure 5C; Supplementary Figure S3; Supplementary Figure S4). Among these four significant modules, the purple module positively correlated with both the IFC and CIE *a** value. In contrast, the dark grey, dark red, and black modules exhibited negative correlations with the IFC and CIE *a** value. These four modules contained a total of 6,045 genes encoding proteins. Subsequently, we focused on 6,045 genes for subsequent functional enrichment analysis. Details of the 6,045 genes are shown in Supplementary Table S12.

## 3.5 Functional enrichment analysis for the four key modules

The significant GO_WGCNA terms and KEGG_WGCNA pathways are presented in Supplementary Tables S13 and S14. The GO_WGCNA results showed that genes in the black red module were significantly enriched in 35 GO terms, which were mainly related to IFC and CIE *a**, such as regulation of the catabolic process (GO:0009894), RNA binding (GO:0003723), negative regulation of lipid localization (GO:1905953), and oxidoreduction-driven active transmembrane transporter activity (GO:0015453). From the KEGG_WGCNA analysis results, 156 pathways were significantly enriched, and most of the

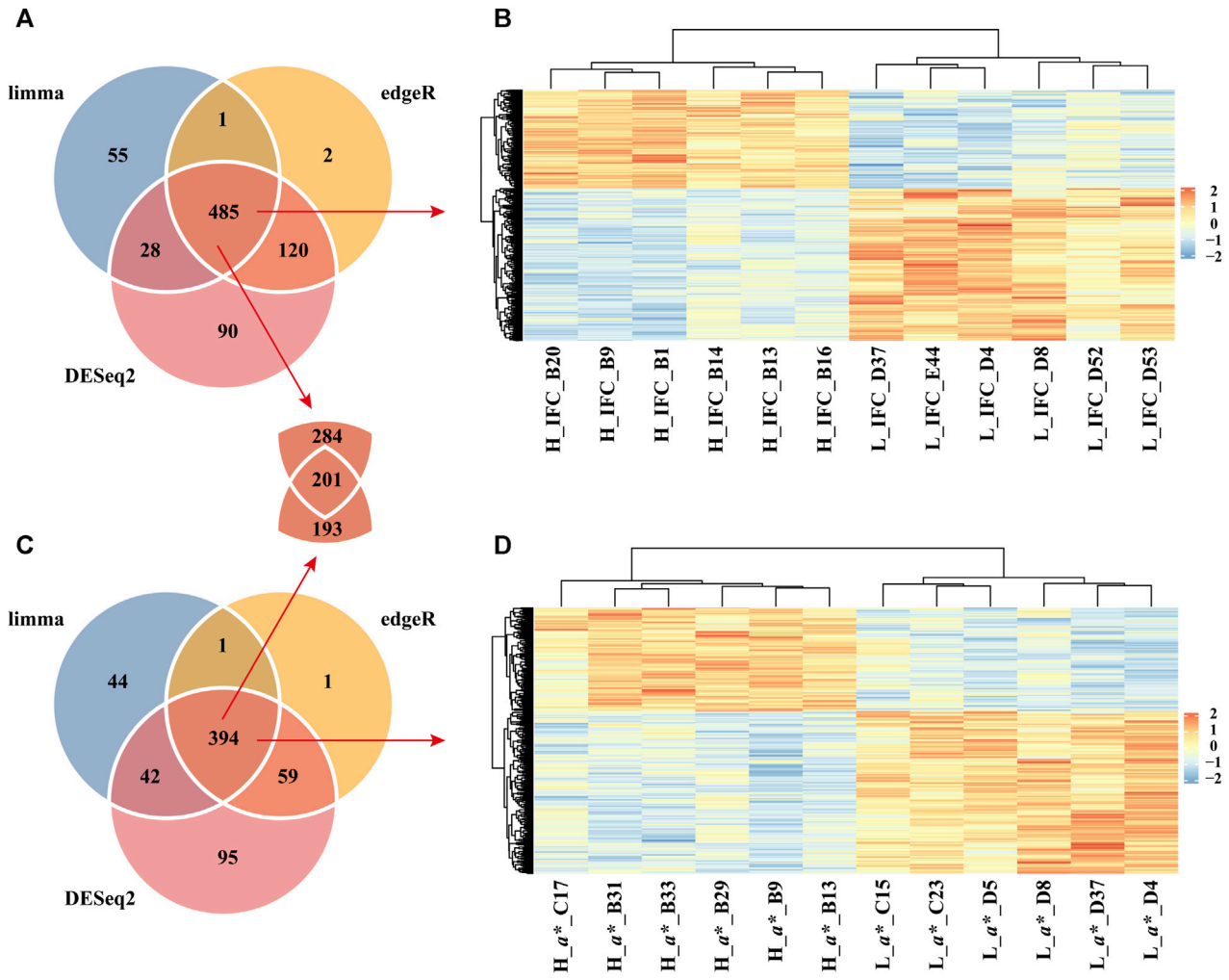



FIGURE 3
Identification of differentially expressed genes (DEGs). **(A)** Venn diagram of DEGs identified using the DESeq2, limma, and edgeR packages for the IFC. **(B)** Heatmap of overlapping DEGs between the H_IFC and L_IFC groups. **(C)** Venn diagram of DEGs identified using the DESeq2, limma, and edgeR packages for the CIE *a** value. **(D)** Heatmap of overlapping DEGs between the H_*a** and L_*a** groups.

significant pathways were related to lipid deposition, decomposition, and oxidation-reduction reactions. These pathways are critical in the regulation of both IFC and CIE *a**, such as the adipocytokine signaling pathway (ssc04920), FoxO signaling pathway (ssc04068), MAPK signaling pathway (ssc04010), and oxidative phosphorylation (ssc00190).

In the black module, the functional enrichment results showed that 41 GO_WGCNA terms and 86 pathways were significantly enriched. These significant GO_WGCNA terms were mainly involved in phosphorylation (GO:0016310), response to oxygen-containing compounds (GO:1901700), the actin cytoskeleton (GO:0015629), and calcium ion binding (GO:0005509). The significant pathways related to lipid metabolism and oxidative reactions mainly included regulation of lipolysis in adipocytes (ssc04923), glycerolipid metabolism (ssc00561), the PI3K-Akt signaling pathway (ssc04151), the MAPK signaling pathway (ssc04010), and the Wnt signaling pathway (ssc04310).

Genes in the purple module were significantly enriched with 6 GO_WGCNA terms and 14 KEGG_WGCNA pathways. These GO terms were mainly involved in extracellular matrix organization (GO:0030198) and collagen binding (GO:0005518). Among the significant KEGG_WGCNA pathways, four were associated with IFC, such as fatty acid metabolism (ssc01212), insulin resistance (ssc04931), calcium signaling pathway (ssc04020), and fatty acid degradation (ssc00071). Genes in the black grey modules were not significantly enriched in GO terms and KEGG pathways, which might have been due to the limited number of genes in this module.

## 3.6 Identification of candidate genes related to the IFC and CIE *a** value

To determine the candidate genes affecting the IFC and CIE *a** value, we first screened the overlapping GO terms and KEGG pathways for each trait based on the functional enrichment analysis results of overlapping DEGs, GSEA, and WGCNA. The DEGs in the overlapping GO terms and KEGG pathways were selected as candidate genes. Finally, hub genes with a connectivity

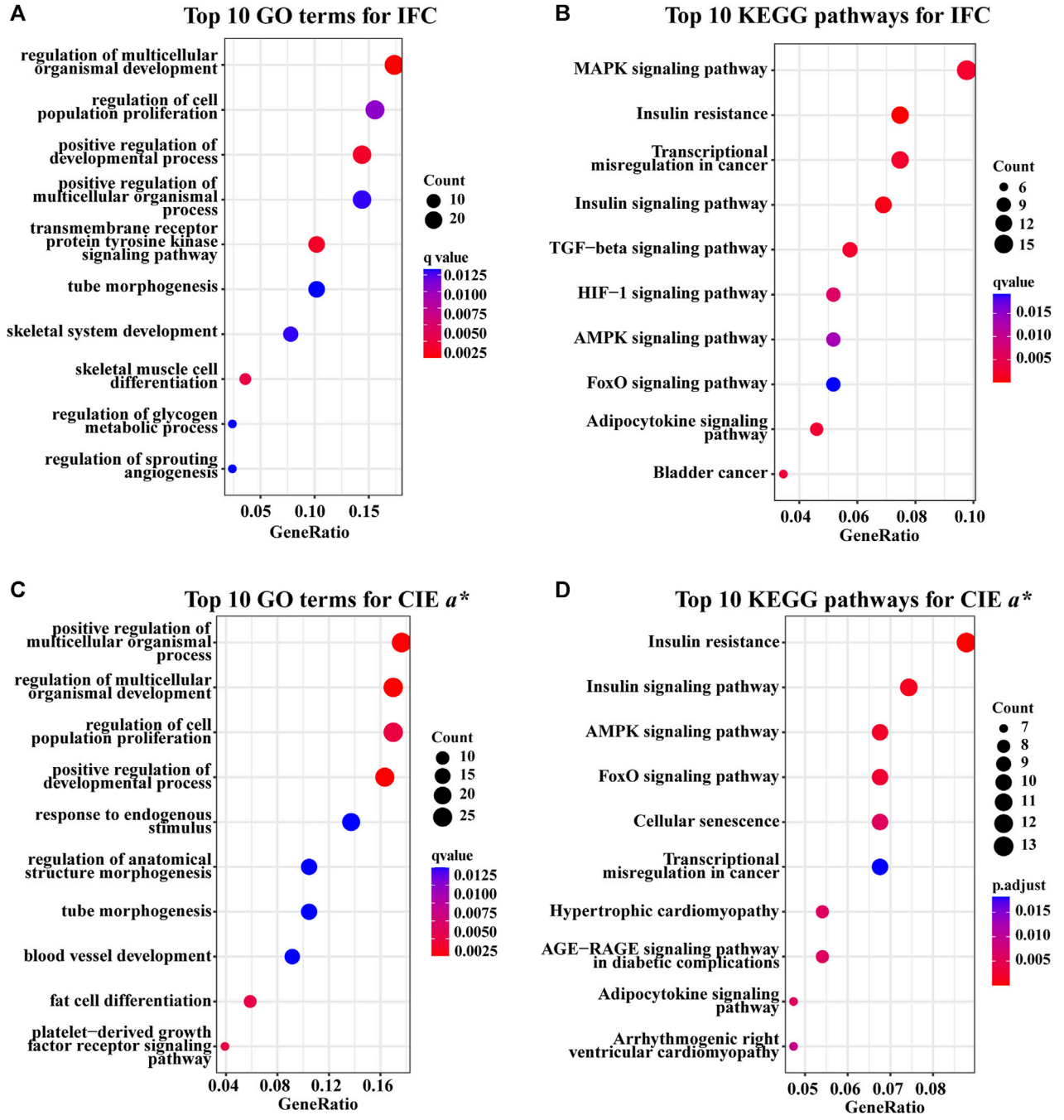



FIGURE 4
GO and KEGG enrichment analysis of overlapping DEGs. **(A)** Top five GO terms of overlapping DEGs for the IFC in the BP, CC, and MF categories. **(B)** Significantly enriched KEGG pathways of overlapping DEGs for IFC. **(C)** Top five GO terms of overlapping DEGs for CIE *a** value in the BP, CC, and MF categories. **(D)** Significantly enriched KEGG pathways of overlapping DEGs for the CIE *a** value. The size of the dot represents the number of overlapping DEGs enriched to this GO term or pathway. The colour of the dot represents the significance of the enrichment, where a redder dot indicates greater significance.

value exceeding ten were obtained by constructinga PPI network of candidate genes. For IFC, 2 overlapping GO terms and 11 overlapping pathways were identified (Figures 6A, B). Most of these overlapping GO terms and pathways were involved in lipid metabolism, such as response to oxygen-containing compounds (GO:1901700), DNA-binding transcription factor activity (GO: 0003700), insulin resistance (ssc04931), the MAPK signaling pathway (ssc04010), adipocytokine signaling pathway (ssc04920), the HIF-1 signaling pathway (ssc04066), and the FoxO signaling pathway (ssc04068). For the CIE *a** value, 6 overlapping GO terms, and 10 overlapping pathways were identified (Figures 6C, D). Most of these overlapping GO terms, and pathways were involved in oxidative phosphorylation, system development, and lipid metabolism, such as response to oxygen-containing compounds

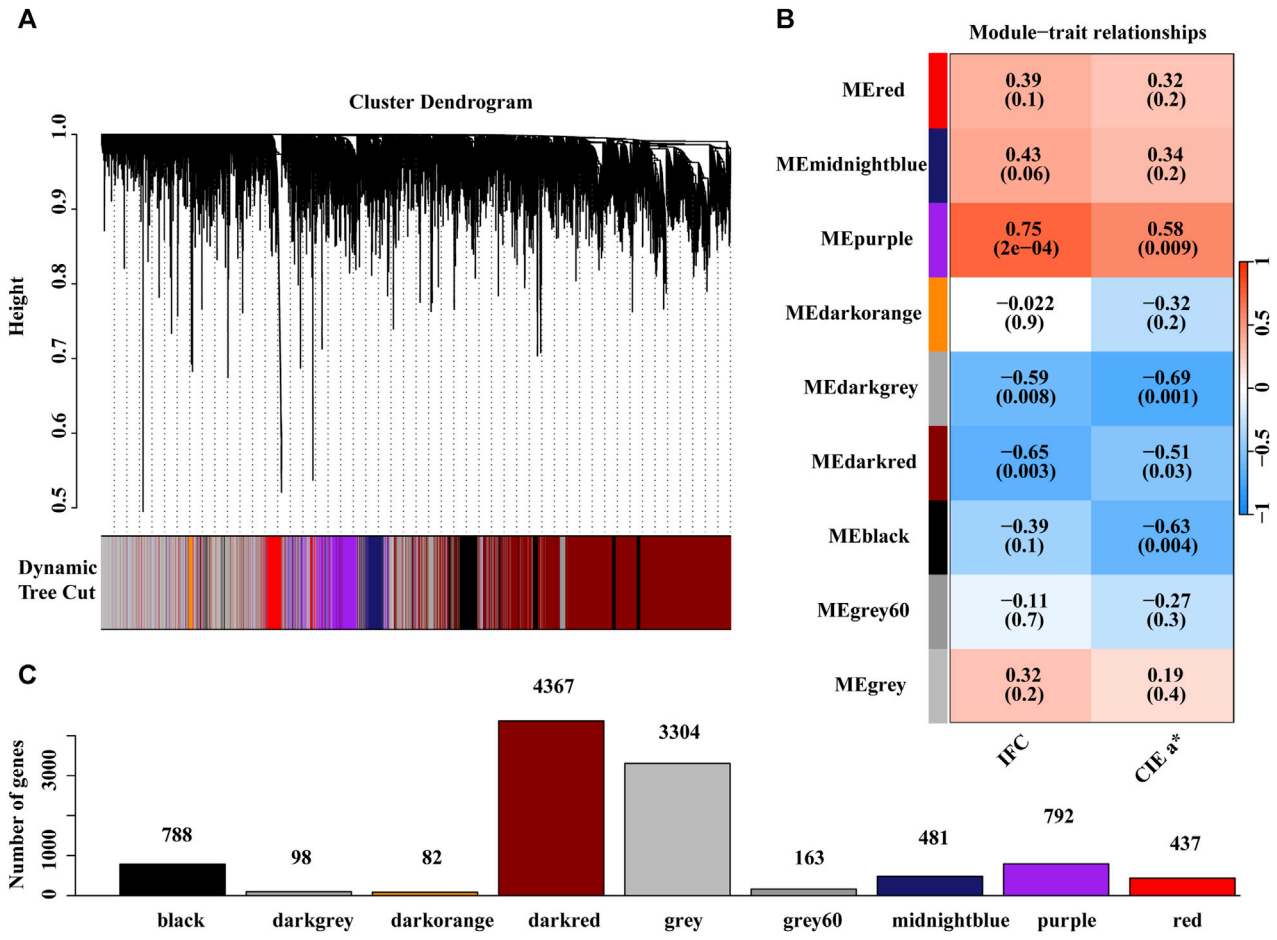



FIGURE 5
Weighted gene co-expression network analysis (WGCNA). **(A)** The gene dendrogram was obtained by clustering the dissimilarity based on consensus Topological Overlap with the corresponding module colors indicated by the color row. **(B)** Matrix with module trait relationships (MTRs) and corresponding *p* values between the detected modules on the *y*-axis and traits (IFC and CIE *a** value) on the *x*-axis, where blue represents a negative correlation, red represents a positive correlation, and white represents no correlation. **(C)** The number of genes contained in each module.

(GO:1901700), negative regulation of signaling (GO:0023057), response to wounding (GO:0009611), circulatory system development (GO:0072359), the FoxO signaling pathway (ssc04068), the adipocytokine signaling pathway (ssc04920), the MAPK signaling pathway (ssc04010), and the HIF-1 signaling pathway (ssc04066).

We selected terms and pathways associated with lipid metabolism and redox in overlapping GO terms and KEGG pathways, and DEGs located in these terms and pathways were considered candidate genes. The selected GO terms and KEGG pathways for the IFC and CIE *a** value are shown in Table 2 and Table 3. The results showed that 47 and 53 genes can be considered candidate genes for the IFC and CIE *a** value, respectively. These candidate genes were used for subsequent PPI network construction. It was worth noting that among these two traits, there was one GO term (response to oxygenated compounds) and three KEGG pathways (adipocyte cytokine signaling pathway, MAPK signaling pathway, and HIF-1 signaling pathway) that were consistent, and these two traits shared 18 candidate genes (Supplementary Table S15).

## 3.7 Hub genes

The interaction relationships of candidate genes affecting the IFC and CIE *a** value were obtained by constructing PPI networks (Figure 7). According to the degree of connectivity, five hub genes (*ATF3*, *SOX9*, *PPARGC1A*, *CEBPB*, and *MYC*) with a connectivity value greater than ten were identified as hub genes for IFC trait. Functional enrichment analysis showed that *CEBPB*, *SOX9*, and *PPARGC1A* were mainly involved in the transcriptional regulation of white adipocyte differentiation and the regulation of fatty acid oxidation. For the CIE *a** value, 13 hub genes (*IL6*, *MYC*, *EGR1*, *CEBPB*, *JUNB*, *THBS1*, *SERPINE1*, *SOCS3*, *DUSP1*, *SOX9*, *PPARGC1A*, *CCL2*, and *FOXO1*) were identified as hub genes. Functional enrichment analysis showed that *SOCS3*, *IL6*, *FOXO1*, *CEBPB*, *SOX9*, and *PPARGC1A* were mainly involved in the adipocytokine signaling pathway, insulin resistance, FoxO signaling pathway, AMPK signaling pathway, and PI3K-Akt signaling pathway. Notably, *MYC*, *CEBPB*, *SOX9*, and *PPARGC1A* were considered hub genes (transcription factors) affecting both traits, and their expression levels were significantly higher in the low group than in the high group.

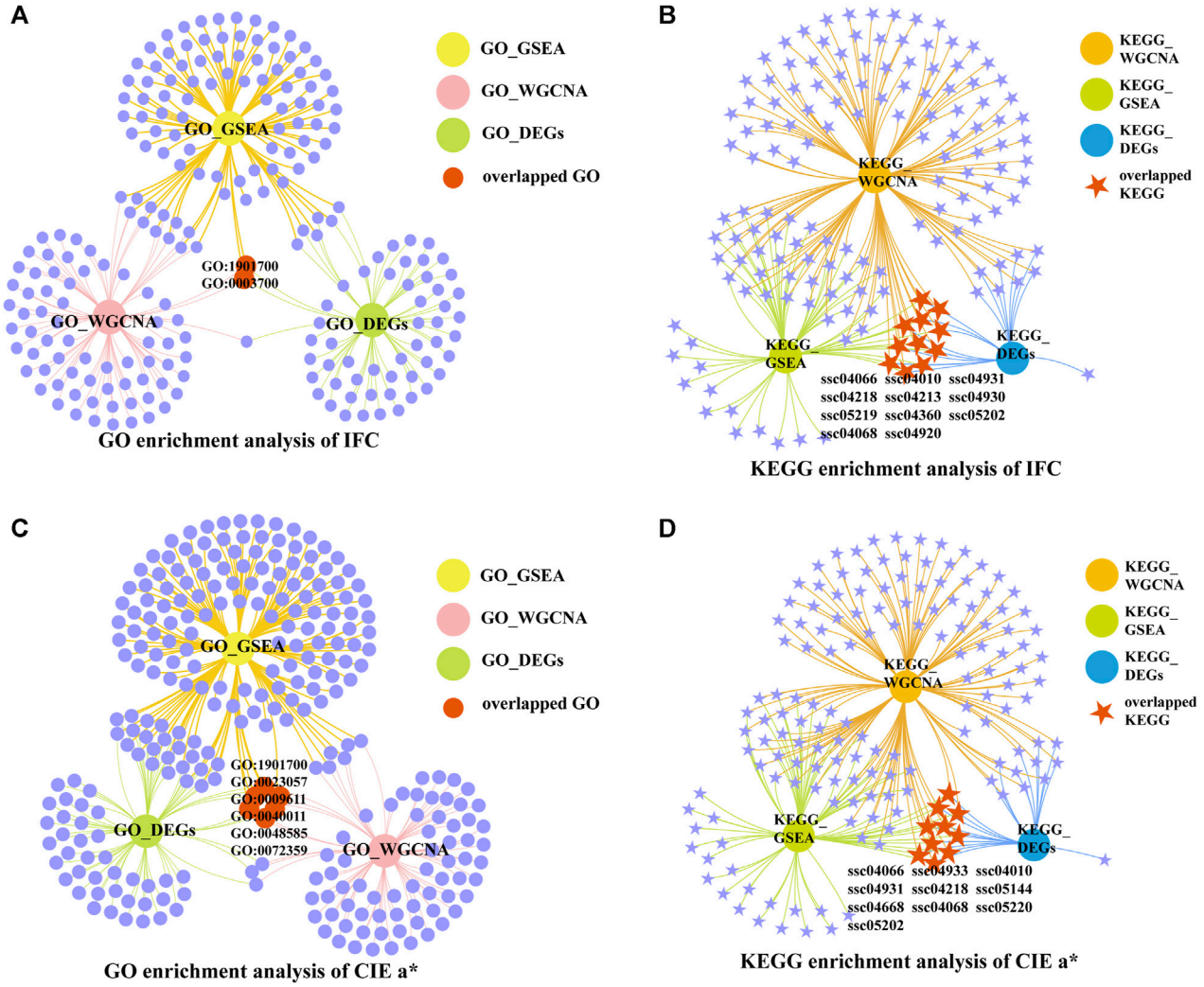



FIGURE 6
Venn network diagrams of enrichment analysis. **(A)** Venn network diagram of significantly enriched GO terms under three conditions for IFC. **(B)** Venn network diagram of significantly enriched KEGG pathways under three conditions for IFC. **(C)** Venn network diagram of significantly enriched GO terms under three conditions for the CIE *a** value. **(D)** Venn network diagram of significantly enriched KEGG pathways under three conditions for CIE *a** value.

## 4 Discussion

DLY pork is dominant in the pork industry; however, its IFC is low, and the meat has a paler color, resulting in limited competitiveness within the premium pork market segment (Chen et al., 2018; Wang et al., 2020). As a result, breeders are eager to undertake genetic improvements in both IFC and redness (CIE *a**) meat color concurrently to cater to consumer market demands. In this study, a highly significant positive correlation ($R = 0.309$, $p < 0.001$) between the IFC and the CIE *a** value was observed, similar to previous reports by Mortimer et al. (Mortimer et al., 2014) and Zhang et al. (Zhang et al., 2022), in which they discovered the correlation coefficients of the IFC and CIE *a** value was 0.260 and 0.323, respectively. The interaction between IFC and meat color is intricate. Several studies have shown that muscles with a higher percentage of red muscle fibers (higher CIE a* values) tend to have a higher IFC (Karlsson et al., 1999; Guo et al., 2011). On the one hand, this is because red muscle fibers contain more neutral fat. On the other hand, the red muscle fiber contains more mitochondria, which are the prominent organelles for fatty acid β-oxidation. Therefore, more lipids may accumulate around the red muscle fibers (internally and externally) to ensure β-oxidation and provide energy to the body. However, the relationship between IFC and muscle redness has not been fully demonstrated. Numerous studies have found significant correlations between IFC and CIE *a**, suggesting that there might be similarities in the genetic background regulating changes in both the IFC and redness value. Consequently, transcriptome analysis was conducted using individuls with extreme IFC and CIE *a** values to identify hub genes and metabolic pathways co-regulating IFC and the redness of pork.

Conducting transcriptomic analysis based on extreme phenotypes is a commonly employed method to identify key genes influencing target traits. For instance, Wang et al. (2023) in the Anqing Six-end-white pigs, employed RNA-seq on high and low

TABLE 2 Overlapping significantly enriched GO terms based on GO enrichment of DEGs, GSEA, and WGCNA.

| Trait | GO ID | Description | GO_DEGs *q* value | GO_GSEA *q* value | GO_WGCNA *q* value | Overlapping DEGs |
|---|---|---|---|---|---|---|
| IFC | GO: 1901700 | response to oxygen-containing compound | 0.069 | 0.047 | 0.005 | *APOD, INHBB, CEBPB, NR4A3, SOX9, MYOD1, CYP26B1, BGLAP, PANX1, THBS1, PCK1* |
| | GO: 0003700 | DNA-binding transcription factor activity | 0.078 | 0.001 | 0.050 | *TGIF1, RUNX1, FOSL2, KLF10, MAFK, SMAD1, MAFF, CSRNP1, CEBPB, NR4A3, SIM1, ATF3, SOX9, MYOD1, CREM, ZSCAN20, KLF5, FOSL1* |
| CIE *a** | GO: 1901700 | response to oxygen-containing compound | 0.023 | <0.001 | 0.005 | *THBS1, INHBB, FOXO1, EGR1, PLSCR4, PLK3, CEBPB, SOCS1, SOX9, GJA1, SLC25A33, SLC11A1, NOCT, CCL2, SLC1A1, APOD* |
| | GO: 0023057 | negative regulation of signaling | 0.034 | 0.004 | 0.012 | *ADRB2, THBS1, SLC25A5, ADM, SIAH2, SPRY1, SOCS3, INHBB, EGR1, ARRDC3, DUSP5, SOCS1, SOX9, GJA1, APOD* |
| | GO: 0009611 | response to wounding | 0.043 | 0.001 | 0.035 | *CCN1, PPL, SERPINE1, THBS1, F3, INHBB, SLC1A1, ITGA5, APOD* |
| | GO: 0072359 | circulatory system development | 0.043 | <0.001 | <0.001 | *CCN1, SERPINE1, JUNB, THBS1, ADM, VEGFA, TIPARP, ITGA5, ANGPTL4, F3, EGR2, SOX9, GJA1, SLC1A1* |

TABLE 3 Overlapping significantly enriched KEGG pathways based on KEGG enrichment of DEGs, GSEA, and WGCNA.

| Trait | KEGG ID | Description | KEGG_DEGs *q* value | KEGG_GSEA *q* value | KEGG_WGCNA *q* value | Overlapping DEGs |
|---|---|---|---|---|---|---|
| IFC | ssc04931 | Insulin resistance | <0.001 | 0.07 | <0.001 | *INSR, PPARGC1A, PTPN1, TRIB3, IRS2, PRKAG2, SOCS3, GFPT2* |
| | ssc04010 | MAPK signaling pathway | 0.002 | 0.052 | <0.001 | *MAP2K3, FLNC, VEGFA, GADD45A, INSR, HSPB1, MAP3K8, GADD45G, MYC, IL1RAP* |
| | ssc04920 | Adipocytokine signaling pathway | 0.002 | 0.071 | <0.001 | *PPARGC1A, IRS2, PRKAG2, SOCS3* |
| | ssc04066 | HIF-1 signaling pathway | 0.006 | 0.021 | <0.001 | *VEGFA, INSR, IL6R, SERPINE1, HK3, TIMP1* |
| | ssc04068 | FoxO signaling pathway | 0.02 | 0.029 | <0.001 | *GADD45A, INSR, GABARAPL1, IRS2, PRKAG2, GADD45G* |
| CIE *a** | ssc04068 | FoxO signaling pathway | 0.002 | <0.001 | <0.001 | *GADD45A, GABARAPL1, IRS2, PRKAG2, GADD45B, FOXO1, IL6, PLK3, FBXO32* |
| | ssc04920 | Adipocytokine signaling pathway | 0.005 | 0.097 | <0.001 | *PPARGC1A, IRS2, PRKAG2, CPT1A, SOCS3* |
| | ssc04010 | MAPK signaling pathway | 0.043 | 0.013 | <0.001 | *GADD45A, FLNC, GADD45B, VEGFA, FGF6, DUSP1, IL1RAP, DUSP4, MYC, DUSP5, DUSP2* |
| | ssc04066 | HIF-1 signaling pathway | 0.074 | 0.021 | <0.001 | *SERPINE1, VEGFA, TIMP1, IL6, HK2* |

IFC groups to discern critical genes affecting intramuscular fat deposition. Ninety-seven DEGs obtained in their study overlapped with those identified in our high and low IFC groups, including *MYC*, *ATF3*, and *LEP*, which have been reported as candidate genes related to lipid metabolism. Furthermore, Fernández-Barroso et al. (2022) conducted RNA-seq in the LD muscle of Iberian pigs based on extreme phenotypes of myoglobin (CIE *a** value). Among the 57 DEGs they obtained, three genes, such as *CCL2*, *VSTM1*, and *ACKR2*, were consistent with our results, and these genes might participate in metabolic pathways linked to redox reactions. Thus, we can conclude that conducting RNA-seq based on extreme phenotypes is an effective strategy.

In this study, WGCNA was used to detect the vital genes and modules associated with the IFC and CIE *a** values using transcriptome data from 19 samples. The results of the WGCNA showed that the purple module demonstrated a positive correlation

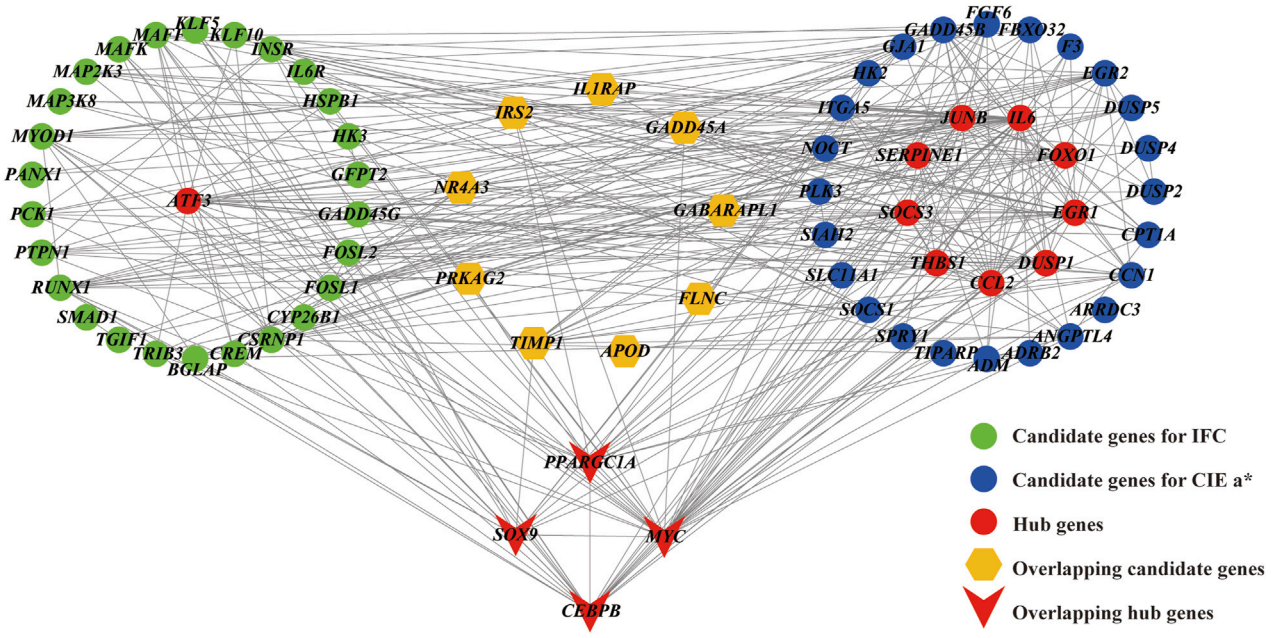



FIGURE 7
Protein protein interaction (PPI) network for the candidate genes affecting the IFC and CIE *a** value. Edges (gray lines) between nodes indicate the interaction of genes in the network. Green circles represent candidate DEGs for IFC, and blue circles represent candidate DEGs for CIE *a**. Brown hexagos represent overlapping candidate DEGs for IFC and CIE *a**. Red circles and V-shapes represent DEGs with connectivity greater than 10 and are considered hub genes. Hub genes shared by the IFC and CIE *a** value are represented by V-shapes.

with both th1e IFC and CIE *a** value. In contrast, the dark grey, dark red, and black modules exhibited negative correlations with the IFC and CIE *a** value. These four modules contained a total of 6,045 genes encoding proteins. Based on the overlap analysis between the DEGs (DEGs of the IFC and DEGs of the CIE *a** value) and the WGCNA results, more than 70% of the DEGs could be detected by WGCNA, indicating the similarity between these two analysis methods and further proving the reliability of the results of this study. However, some genes associated with the IFC and CIE *a** value identified by WGCNA did not exhibit differential expression in the high and low groups. This observation suggests that WGCNA recognized additional information by establishing interconnected networks between genes, aligning well with the foundational principles of WGCNA. This was consistent with the findings of Xing et al. (2021).

The IFC and CIE *a** groups shared four significantly enriched pathways: the FoxO signaling pathway (ssc04068), adipocytokine signaling pathway (ssc04920), MAPK signaling pathway (ssc04010), and HIF-1 signaling pathway (ssc04066) (Table 3). The FoxO signaling pathway governs glucose and lipid metabolism by controlling genes associated with gluconeogenesis, glycogenolysis, and lipid metabolism (Lee and Dong, 2017). It also impacts fatty acid oxidation and storage across diverse tissues (Chen et al., 2023a). Although the direct connection between the FoxO pathway and myoglobin oxidation has not been extensively documented, it is conceivable that this pathway may indirectly influence oxidative processes by regulating energy metabolism and responses to oxidative stress (Egan and Zierath, 2013). The adipocytokine signaling pathway is linked with adipocyte-related functions and metabolism. It modulates insulin sensitivity, glucose uptake, and lipid metabolism, affecting the release of adipokines that influence lipid homeostasis and inflammation (Gu et al., 2019). This pathway likely indirectly affects myoglobin oxidation by influencing factors connected to metabolism and inflammation, thus potentially impacting oxidative processes in muscle tissues (Jorge et al., 2011). The MAPK signaling pathway is integral to various cellular processes, encompassing cell growth, differentiation, and metabolism. It can impact lipid metabolism by regulating genes related to lipogenesis, lipolysis, and fatty acid oxidation (Chen et al., 2023b; Wang et al., 2023). This pathway may contribute to muscle oxidative processes by mediating cellular reactions to stress, lipid peroxidation, and growth cues, thereby influencing myoglobin oxidation under specific conditions (Xu et al., 2018). Activated in response to low oxygen levels, the HIF-1 signaling pathway orchestrates adaptive responses to hypoxia. It influences glycolysis, lipid, and energy metabolism when oxygen levels are low (Zhang et al., 2023). The HIF-1 pathway can affect myoglobin oxidation by regulating the response to hypoxia, potentially influencing oxidative metabolism and the role of myoglobin in oxygen transport and storage (Elkholi et al., 2022). In summary, these pathways may play pivotal roles in both fatty acid metabolism and myoglobin oxidation.

The DEGs in Table 2 and Table 3 were considered candidate genes influencing the IFC and CIE *a** values, and the PPI network was constructed based on them (Figure 7). Based on the degree of connectivity, 5 hub genes (*ATF3*, *SOX9*, *PPARGC1A*, *CEBPB*, and *MYC*) with a connectivity value exceeding ten were regarded as hub genes potentially influencing IFC. Similarly, 13 hub genes impacting the CIE *a** value were identified, including *IL6*, *MYC*, *EGR1*, *CEBPB*, *JUNB*, *THBS1*, *SERPINE1*, *SOCS3*, *DUSP1*, *SOX9*, *PPARGC1A*,

*CCL2*, and *FOXO1*. *ATF3* (activating transcription factor 3), a member of the CREB family of basic leucine zipper transcription factors (TFs). It has been found that the deletion of *ATF3* results in increased lipid body accumulation, and *ATF3* directly regulates transcription of the gene encoding cholesterol 25-hydroxylase (Gold et al., 2012).

*IL6* (interleukin-6) is a pivotal regulatory factor for lipolysis and beta-oxidation. Numerous *in vitro* studies have substantiated that treatment with *IL6* enhances lipolysis and beta-oxidation in both myotubes and adipocytes (Bae et al., 2023; Jackson et al., 2023). *EGR1* (Early growth response 1) is a transcription factor. Mohtar et al. found that insulin/mTORC1-inducible *EGR1* binds to the leptin promoter and activates leptin expression in 3T3-L1 adipocytes, regulating lipid metabolism (Mohtar et al., 2019). The results of Yan et al. suggested that inhibition of *JUNB* might be a key indicator of the regulation of the APOA2-associated PPARα pathway (Yan et al., 2020). *APOA2* is a well-known member of the apolipoprotein family (Ballester et al., 2016), and the PPARα pathway is also a key pathway in regulating lipid metabolism (Cao et al., 2023). *THBS1* (thrombospondin-1) is a prototypical matricellular protein. *THBS1*-null mice exhibited elevated free fatty acids and triglycerides compared to wild-type mice, suggesting impaired fatty acid uptake (Kong et al., 2013). *SERPINE1* (Serpin Family E Member 1), also known as plasminogen activator inhibitor type 1 (PAI-1), is a member of the serine proteinase inhibitor (serpin) superfamily. Several findings have shown that PAI-1 might promote the differentiation of mesenchymal stem cells toward adipogenesis, and PAI-1 deficiency attenuates changes in the levels of adipogenic genes such as PPARγ and aP2 (Tamura et al., 2013; Hu et al., 2019). *SOCS3* (suppressor of cytokine signaling 3) plays an important role in regulating energy metabolism processes. In recent years, researchers have found that *SOCS3* is involved in the AMPK signaling pathway, insulin resistance, adipocytokine signaling pathway, and JAK/STAT pathway, is activated/triggered by leptin signals, and plays important roles in lipid metabolism processes (Liu et al., 2014; Fang et al., 2020; Yang et al., 2020). DUSPs (dual-specificity phosphatases) are the key phosphatases in the MAPK pathway. Recently, *DUSP1* was suggested to play a critical role in the switch from oxidative to glycolytic myofibers (Flach et al., 2011), and can regulate fatty acid oxidation (Roth et al., 2009). *CCL2* (chemokine ligand 2) is a member of the C–C motif family of chemokines. Kang et al. found that after *CCL2* binds to its receptor *CCR2*, it can reduce lipid peroxidation by inhibiting *CCR2*, indicating its important regulatory role in lipid oxidation metabolism (Roth et al., 2009). Current studies suggest that the transcription factor *FOXO1* (forkhead box protein O1) is involved in lipid metabolism and lipolysis in adipocytes (Chakrabarti and Kandror, 2009; Chakrabarti et al., 2011). Song et al. found that interfering with *FOXO1* negatively regulated the expression of adipogenic differentiation marker genes and lipid anabolism marker genes, thus reducing triglyceride content and inhibiting the generation of lipid droplets in bovine adipocytes (Song et al., 2023).

It is worth noting that these two traits share four hub genes: *MYC*, *CEBPB*, *SOX9*, and *PPARGC1A*. *MYC* is a transcription factor that regulates cell proliferation and differentiation in healthy cellular processes. Hall et al. revealed that the activation of *MYC* led to the accumulation of cholesteryl esters stored in lipid droplets (Hall et al., 2020). A previous study found that *MYC* is involved in the MAPK signaling pathway, promoting the glycolysis process in fish T cells (Wei et al., 2020). In addition, *MYC* is involved in the WNT signaling pathway and serves as a target gene/transcriptome factor for WNT, regulating myogenesis (Karczewska-Kupczewska et al., 2016). *CEBPB* (CCAAT/enhancer binding protein β) is a member of the transcription factor family of CEBP. Several studies have reported that *PPARGC1A* (PPAR coactivator-1α, also known as *PGC1α*), a transcriptional co-activator of PPARγ, can bind to *CEBPB* and form a transcription complex. This complex may promote the transcription of *CPT1A* (carnitine palmitoyl transferase 1 A) and activate fatty acid β-oxidation (Du et al., 2019; Wu et al., 2020). *SOX9* (Sex-determining region Y-type box-9) is a member of the Sox supergene family and has been proven to be an essential transcription factor in cartilage formation during chondrocyte proliferation (Akiyama, 2008). Wang et al. confirmed that *SOX9* can directly bind to the promoters of *CEBPB* and *CEBPD*, inhibit their promoter activity, and prevent adipocyte differentiation (Wang and Sul, 2009). This evidence indicated that the *SOX9*/*CEBPB*/*PPARGC1A* axis might play an essential regulatory role in fatty acid β-oxidation. Myoglobin is an oxygen-binding hemeprotein generally localized to oxidative muscle and functions as an oxygen store and reactive oxygen species scavenger (Gödecke, 2010). Schlater et al. confirmed that an increase in lipids could stimulated an increase in myoglobin content in muscle cells of C2C12 mice, which was closely related to fatty acid beta oxidation (Schlater et al., 2014). In summary, we speculated that the *SOX9*/*CEBPB*/*PPARGC1A* axis plays a vital role in the co-regulation of IFC deposition and changes in the redness of meat color. The expression levels of the upstream gene *STAT3* (signal transducer and activator of transcription 3) and downstream *CPT1A* genes (log2FC = 1.17) in the *SOX9*/*CEBPB*/*PPARGC1A* axis were also significantly different in the high and low groups in this study, further supporting the importance of this pathway in the synergistic regulation of lipid and myoglobin metabolism. Thererore, it will be particularly interesting to investigate the co-regulatory mechanism of the *SOX9*/*CEBPB*/*PPARGC1A* axis in IFC and CIE *a** value traits in further studies.

## 5 Conclusion

In this study, we identified 5 hub genes influencing the IFC and 13 hub genes affecting the CIE *a** value through integrating differential gene expression analysis, WGCNA, functional enrichment under various conditions, and PPI network analysis. These genes maninly participate in multiple lipid and myoglobin metabolism pathways. Moreover, we discovered that the *SOX9*/*CEBPB*/*PPARGC1A* axis is the potential pathway co-regulating lipid deposition and the myoglobin redox reaction. These hub genes and the *SOX9*/*CEBPB*/*PPARGC1A* axis may be critical for the IFC and CIE *a** value; however, the functions and regulatory mechanism of these hub genes, particularly the *SOX9*/*CEBPB*/*PPARGC1A* axis, still need to be further elucidated.

## Data availability statement

The original contributions presented in the study are publicly available. This data can be found in the NCBI database, under BioProject PRJNA1052206, https://www.ncbi.nlm.nih.gov/bioproject/1052206.

## Ethics statement

The animal study was approved by the Nanjing Agricultural University Animal Care and Use Committee (Certification No. SYXK (Su) 2022-0031). The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

BW: Data curation, Software, Visualization, Writing–original draft, Writing–review and editing. LH: Writing–review and editing. WY: Writing–review and editing. XM: Writing–review and editing. KQ: Writing–review and editing. ZX: Conceptualization, Methodology, Writing–review and editing. WW: Conceptualization, Methodology, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was financially supported by the National Natural Science Foundation of China (32302695), the Project of Seed Industry Revitalization in Jiangsu Province (JBGS2021024), the Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding (2021C02068), the Key R & D projects of Zhejiang Province (2021C02007), and the China Agriculture Research System of MOF and MARA (CARS36).

## 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.2024.1351429/full#supplementary-material

## References

Akiyama, H. (2008). Control of chondrogenesis by the transcription factor Sox9. *Mod. Rheumatol.* 18, 213–219. doi:10.1007/s10165-008-0048-x

Bae, H. R., Shin, S.-K., Yoo, J.-H., Kim, S., Young, H. A., and Kwon, E.-Y. (2023). Chronic inflammation in high-fat diet-fed mice: unveiling the early pathogenic connection between liver and adipose tissue. *J. Autoimmun.* 139, 103091. doi:10.1016/j.jaut.2023.103091

Ballester, M., Revilla, M., Puig-Oliveras, A., Marchesi, J. A. P., Castelló, A., Corominas, J., et al. (2016). Analysis of the porcine APOA2 gene expression in liver, polymorphism identification and association with fatty acid composition traits. *Anim. Genet.* 47, 552–559. doi:10.1111/age.12462

Barabasi, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi:10.1038/nrg1272

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

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-Australasian J. Anim. Sci.* 28, 1061–1065. doi:10.5713/ajas.14.0783

Cao, D., Khan, Z., Li, X., Saito, S., Bernstein, E. A., Victor, A. R., et al. (2023). Macrophage angiotensin-converting enzyme reduces atherosclerosis by increasing peroxisome proliferator-activated receptor α and fundamentally changing lipid metabolism. *Cardiovasc. Res.* 119, 1825–1841. doi:10.1093/cvr/cvad082

Cardoso, T. F., Cánovas, A., Canela-Xandri, O., González-Prendes, R., Amills, M., and Quintanilla, R. (2017). RNA-seq based detection of differentially expressed genes in the skeletal muscle of Duroc pigs with distinct lipid profiles. *Sci. Rep.* 7, 40005. doi:10.1038/srep40005

Chakrabarti, P., English, T., Karki, S., Qiang, L., Tao, R., Kim, J., et al. (2011). SIRT1 controls lipolysis in adipocytes via FOXO1-mediated expression of ATGL. *J. Lipid Res.* 52, 1693–1701. doi:10.1194/jlr.M014647

Chakrabarti, P., and Kandror, K. V. (2009). FoxO1 controls insulin-dependent adipose triglyceride lipase (ATGL) expression and lipolysis in adipocytes. *J. Biol. Chem.* 284, 13296–13300. doi:10.1074/jbc.C800241200

Chen, S., Lu, Z., Jia, H., Yang, B., Liu, C., Yang, Y., et al. (2023a). Hepatocyte-specific Mas activation enhances lipophagy and fatty acid oxidation to protect against acetaminophen-induced hepatotoxicity in mice. *J. Hepatol.* 78, 543–557. doi:10.1016/j.jhep.2022.10.028

Chen, S., Wu, Z., He, Y., Zhu, L., Wang, J., Lin, H., et al. (2023b). Cyclic di-adenosine monophosphate regulates the osteogenic and adipogenic differentiation of hPDLSCs via MAPK and NF-κB signaling: c-di-AMP regulates the osteogenic and adipogenic differentiation of hPDLSCs. *Acta Biochim. Biophys. Sin. (Shanghai).* 55, 426–437. doi:10.3724/abbs.2023018

Chen, W., Zeng, Q., Xu, H., Fang, G., Wang, S., Li, C., et al. (2018). Comparison and relationship between meat colour and antioxidant capacity of different pig breeds. *Anim. Prod. Sci.* 58, 2152–2157. doi:10.1071/AN16184

CIE (1986). *Colorimetry*. 2nd ed. Wien, Austria: CIE Central Bureau Kegelgasse. Publication No 15.2.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Du, Q., Tan, Z., Shi, F., Tang, M., Xie, L., Zhao, L., et al. (2019). PGC1α/CEBPB/CPT1A axis promotes radiation resistance of nasopharyngeal carcinoma through activating fatty acid oxidation. *Cancer Sci.* 110, 2050–2062. doi:10.1111/cas.14011

Duan, S., Tang, X., Li, W., and Huang, X. (2023). Analysis of the differences in volatile organic compounds in different muscles of pork by GC-IMS. *Molecules* 28, 1726. doi:10.3390/molecules28041726

Egan, B., and Zierath, J. R. (2013). Exercise metabolism and the molecular regulation of skeletal muscle adaptation. *Cell Metab.* 17, 162–184. doi:10.1016/j.cmet.2012.12.012

Elkholi, I. E., Elsherbiny, M. E., and Emara, M. (2022). Myoglobin: from physiological roles to potential implications in cancer. *Biochim. Biophys. Acta - Rev. Cancer* 1877, 188706. doi:10.1016/j.bbcan.2022.188706

Fang, S., Feng, J., Zhang, H., Li, P., Zhang, Y., Zeng, Y., et al. (2020). MiR-455 targeting SOCS3 improve liver lipid disorders in diabetic mice. *Adipocyte* 9, 179–188. doi:10.1080/21623945.2020.1749495

Fernández-Barroso, M. Á., García-Casco, J. M., Núñez, Y., Ramírez-Hidalgo, L., Matos, G., and Muñoz, M. (2022). Understanding the role of myoglobin content in Iberian pigs fattened in an extensive system through analysis of the transcriptome profile. *Anim. Genet.* 53, 352–367. doi:10.1111/age.13195

Flach, R. J. R., Qin, H., Zhang, L., and Bennett, A. M. (2011). Loss of mitogen-activated protein kinase phosphatase-1 protects from hepatic steatosis by repression of cell death-inducing DNA fragmentation factor A (DFFA)-like effector C (CIDEC)/Fat-specific protein 27. *J. Biol. Chem.* 286, 22195–22202. doi:10.1074/jbc.M110.210237

Gödecke, A. (2010). Myoglobin: safeguard of myocardial oxygen supply during systolic compression? *Cardiovasc. Res.* 87, 4–5. doi:10.1093/cvr/cvq126

Gold, E. S., Ramsey, S. A., Sartain, M. J., Selinummi, J., Podolsky, I., Rodriguez, D. J., et al. (2012). ATF3 protects against atherosclerosis by suppressing 25-hydroxycholesterol–induced lipid body formation. *J. Exp. Med.* 209, 807–817. doi:10.1084/jem.20111202

Gu, H., Li, J., Ying, F., Zuo, B., and Xu, Z. (2019). Analysis of differential gene expression of the transgenic pig with overexpression of PGC1α in muscle. *Mol. Biol. Rep.* 46, 3427–3435. doi:10.1007/s11033-019-04805-8

Guo, J., Shan, T., Wu, T., Zhu, L. N., Ren, Y., An, S., et al. (2011). Comparisons of different muscle metabolic enzymes and muscle fiber types in Jinhua and Landrace pigs. *J. Anim. Sci.* 89, 185–191. doi:10.2527/jas.2010-2983

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). BioMart Central Portal—unified access to biological data. *Nucleic Acids Res.* 37, 23–27. doi:10.1093/nar/gkp265

Hall, Z., Wilson, C. H., Burkhart, D. L., Ashmore, T., Evan, G. I., and Griffin, J. L. (2020). Myc linked to dysregulation of cholesterol transport and storage in nonsmall cell lung cancer. *J. Lipid Res.* 61, 1390–1399. doi:10.1194/jlr.RA120000899

Hu, Q., Peng, J., Chen, X., Li, H., Song, M., Cheng, B., et al. (2019). Obesity and genes related to lipid metabolism predict poor survival in oral squamous cell carcinoma. *Oral Oncol.* 89, 14–22. doi:10.1016/j.oraloncology.2018.12.006

Jackson, H. C., Pheiffer, C., Jack, B., and Africander, D. (2023). Time-and glucose-dependent differentiation of 3T3-L1 adipocytes mimics dysfunctional adiposity. *Biochem. Biophys. Res. Commun.* 671, 286–291. doi:10.1016/j.bbrc.2023.06.026

Jorge, M. L. M. P., de Oliveira, V. N., Resende, N. M., Paraiso, L. F., Calixto, A., Diniz, A. L. D., et al. (2011). The effects of aerobic, resistance, and combined exercise on metabolic control, inflammatory markers, adipocytokines, and muscle insulin signaling in patients with type 2 diabetes mellitus. *Metabolism* 60, 1244–1252. doi:10.1016/j.metabol.2011.01.006

Karczewska-Kupczewska, M., Stefanowicz, M., Matulewicz, N., Nikołajuk, A., and Strączkowski, M. (2016). Wnt signaling genes in adipose tissue and skeletal muscle of humans with different degrees of insulin sensitivity. *J. Clin. Endocrinol. Metab.* 101, 3079–3087. doi:10.1210/jc.2016-1594

Karlsson, A. H., Klont, R. E., and Fernandez, X. (1999). Skeletal muscle fibres as factors for pork quality. *Livest. Prod. Sci.* 60, 255–269. doi:10.1016/S0301-6226(99)00098-6

Kim, G.-D., Jeong, J.-Y., Hur, S.-J., Yang, H.-S., Jeon, J.-T., 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. *Food Sci. Anim. Resour.* 30, 626–633. doi:10.5851/kosfa.2010.30.4.626

Kong, P., Gonzalez-Quesada, C., Li, N., Cavalera, M., Lee, D.-W., and Frangogiannis, N. G. (2013). Thrombospondin-1 regulates adiposity and metabolic dysfunction in diet-induced obesity enhancing adipose inflammation and stimulating adipocyte proliferation. *Am. J. Physiol. Metab.* 305, 439–450. doi:10.1152/ajpendo.00006.2013

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Lee, S., and Dong, H. H. (2017). FoxO integration of insulin signaling with glucose and lipid metabolism. *J. Endocrinol.* 233, R67–R79. doi:10.1530/JOE-17-0002

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Liu, S., Huang, J., Wang, X., and Ma, Y. (2020). Transcription factors regulate adipocyte differentiation in beef cattle. *Anim. Genet.* 51, 351–357. doi:10.1111/age.12931

Liu, Z., Gan, L., Yang, X., Zhang, Z., and Sun, C. (2014). Hydrodynamic tail vein injection of SOCS3 eukaryotic expression vector *in vivo* promoted liver lipid metabolism and hepatocyte apoptosis in mouse. *Biochem. Cell Biol.* 92, 119–125. doi:10.1139/bcb-2013-0117

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome. Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Moeller, S. J., Miller, R. K., Edwards, K. K., Zerby, H. N., Logan, K. E., Aldredge, T. L., et al. (2010). Consumer perceptions of pork eating quality as affected by pork quality attributes and end-point cooked temperature. *Meat Sci.* 84, 14–22. doi:10.1016/j.meatsci.2009.06.023

Mohtar, O., Ozdemir, C., Roy, D., Shantaram, D., Emili, A., and Kandror, K. V. (2019). Egr1 mediates the effect of insulin on leptin transcription in adipocytes. *J. Biol. Chem.* 294, 5784–5789. doi:10.1074/jbc.AC119.007855

Mortimer, S. I., van der Werf, J. H. J., Jacob, R. H., Hopkins, D. L., Pannier, L., Pearce, K. L., et al. (2014). Genetic parameters for meat quality traits of Australian lamb meat. *Meat Sci.* 96, 1016–1024. doi:10.1016/j.meatsci.2013.09.007

Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, 4. doi:10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Roth, R. J., Le, A. M., Zhang, L., Kahn, M., Samuel, V. T., Shulman, G. I., et al. (2009). MAPK phosphatase-1 facilitates the loss of oxidative myofibers associated with obesity in mice. *J. Clin. Invest.* 119, 3817–3829. doi:10.1172/JCI39054

Schlater, A. E., De Miranda, M. A., Jr, Frye, M. A., Trumble, S. J., and Kanatous, S. B. (2014). Changing the paradigm for myoglobin: a novel link between lipids and myoglobin. *J. Appl. Physiol.* 117, 307–315. doi:10.1152/japplphysiol.00973.2013

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-Zheng, G., and Su-Mei, Z. (2009). Physiology, affecting factors and strategies for control of pig meat intramuscular fat. *Recent Pat. Food. Nutr. Agric.* 1, 59–74. doi:10.2174/1876142910901010059

Song, Y., Zhang, J., Jiang, C., Song, X., Wu, H., Zhang, J., et al. (2023). FOXO1 regulates the formation of bovine fat by targeting CD36 and STEAP4. *Int. J. Biol. Macromol.* 248, 126025. doi:10.1016/j.ijbiomac.2023.126025

Sun, X., Young, J., Liu, J. H., Bachmeier, L., Somers, R. M., Chen, K. J., et al. (2016). Prediction of pork color attributes using computer vision system. *Meat Sci.* 113, 62–64. doi:10.1016/j.meatsci.2015.11.009

Supakankul, P., and Mekchay, S. (2016). Association of NLK polymorphisms with intramuscular fat content and fatty acid composition traits in pigs. *Meat Sci.* 118, 61–65. doi:10.1016/j.meatsci.2016.03.025

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, 447–452. doi:10.1093/nar/gku1003

Talukdar, H. A., Asl, H. F., Jain, R. K., Ermel, R., Ruusalepp, A., Franzén, O., et al. (2016). Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst.* 2, 196–208. doi:10.1016/j.cels.2016.02.002

Tamura, Y., Kawao, N., Okada, K., Yano, M., Okumoto, K., Matsuo, O., et al. (2013). Plasminogen activator inhibitor-1 is involved in streptozotocin-induced bone loss in female mice. *Diabetes* 62, 3170–3179. doi:10.2337/db12-1552

Wang, B., Li, P., Hou, L., Zhou, W., Tao, W., Liu, C., et al. (2022). Genome-wide association study and genomic prediction for intramuscular fat content in Suhuai pigs using imputed whole-genome sequencing data. *Evol. Appl.* 15, 2054–2066. doi:10.1111/eva.13496

Wang, H., Wang, J., Yang, D., Liu, Z., Zeng, Y., and Chen, W. (2020). Expression of lipid metabolism genes provides new insights into intramuscular fat deposition in Laiwu pigs. *Asian-Australasian J. Anim. Sci.* 33, 390–397. doi:10.5713/ajas.18.0225

Wang, L., Wei, S., Feng, S., Zhao, J., He, X., Fu, S., et al. (2023a). Effects of dietary oat supplementation on carcass traits, muscle metabolites, amino acid profiles, and its association with meat quality of Small-tail Han sheep. *Food Chem.* 411, 135456. doi:10.1016/j.foodchem.2023.135456

Wang, Y., and Sul, H. S. (2009). Pref-1 regulates mesenchymal cell commitment and differentiation through Sox9. *Cell Metab.* 9, 287–302. doi:10.1016/j.cmet.2009.01.013

Wang, Y. L., Hou, Y. H., Ling, Z. J., Zhao, H. L., Zheng, X. R., Zhang, X. D., et al. (2023b). RNA sequencing analysis of the longissimus dorsi to identify candidate genes underlying the intramuscular fat content in Anqing Six-end-white pigs. *Anim. Genet.* 54, 315–327. doi:10.1111/age.13308

Wei, X., Zhang, Y., Li, C., Ai, K., Li, K., Li, H., et al. (2020). The evolutionarily conserved MAPK/Erk signaling promotes ancestral T-cell immunity in fish via c-Myc–mediated glycolysis. *J. Biol. Chem.* 295, 3000–3016. doi:10.1074/jbc.RA119.012231

Wu, H., Liu, B., Chen, Z., Li, G., and Zhang, Z. (2020). MSC-induced lncRNA HCP5 drove fatty acid oxidation through miR-3619-5p/AMPK/PGC1α/CEBPB axis to promote stemness and chemo-resistance of gastric cancer. *Cell Death Dis.* 11, 233. doi:10.1038/s41419-020-2426-z

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov* 2, 100141. doi:10.1016/j.xinn.2021.100141

Wu, Z., Gao, Z., Liang, H., Fang, T., Wang, Y., Du, Z., et al. (2022). Network analysis reveals different hub genes and molecular pathways for pig *in vitro* fertilized early embryos and parthenogenotes. *Reprod. Domest. Anim.* 57, 1544–1553. doi:10.1111/rda.14231

Xing, K., Liu, H., Zhang, F., Liu, Y., Shi, Y., Ding, X., et al. (2021). Identification of key genes affecting porcine fat deposition based on co-expression network analysis of weighted genes. *J. Anim. Sci. Biotechnol.* 12, 100–116. doi:10.1186/s40104-021-00616-9

Xu, L., Zhang, H., Yue, H., Wu, S., Yang, H., Wang, Z., et al. (2018). Gas stunning with CO2 affected meat color, lipid peroxidation, oxidative stress, and gene expression of mitogen-

activated protein kinases, glutathione S-transferases, and Cu/Zn-superoxide dismutase in the skeletal muscles of broilers. *J. Anim. Sci. Biotechnol.* 9, 37. doi:10.1186/s40104-018-0252-2

Yan, P., Zhou, B., Ma, Y., Wang, A., Hu, X., Luo, Y., et al. (2020). Tracking the important role of JUNB in hepatocellular carcinoma by single-cell sequencing analysis. *Oncol. Lett.* 19, 1478–1486. doi:10.3892/ol.2019.11235

Yang, X., Jia, J., Yu, Z., Duanmu, Z., He, H., Chen, S., et al. (2020). Inhibition of JAK2/STAT3/SOCS3 signaling attenuates atherosclerosis in rabbit. *BMC Cardiovasc. Disord.* 20, 133. doi:10.1186/s12872-020-01391-7

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, 17–45. doi:10.2202/1544-6115.1128

Zhang, W., Han, B., Zhang, H., Fu, R., Lu, Y., and Zhang, G. (2023). Integrated transcriptomic and metabolomic analysis of cortical neurons reveals dysregulated lipid metabolism, enhanced glycolysis and activated HIF-1 signaling pathways in acute hypoxia. *Heliyon* 9, e14949. doi:10.1016/j.heliyon.2023.e14949

Zhang, X., Liu, C., Kong, Y., Li, F., and Yue, X. (2022). Effects of intramuscular fat on meat quality and its regulation mechanism in Tan sheep. *Front. Nutr.* 9, 908355–908413. doi:10.3389/fnut.2022.908355

Zhao, X., Wang, C., Wang, Y., Zhou, L., Hu, H., Bai, L., et al. (2020). Weighted gene co-expression network analysis reveals potential candidate genes affecting drip loss in pork. *Anim. Genet.* 51, 855–865. doi:10.1111/age.13006

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 29 April 2024
DOI 10.3389/fgene.2024.1341555



OPEN ACCESS

EDITED BY
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Shamik Polley,
West Bengal University of Animal and Fishery Sciences, India
Mir Asif Iquebal,
Indian Council of Agricultural Research, India

*CORRESPONDENCE
Xu Wang,
xzw0070@auburn.edu

†These authors have contributed equally to this work

RECEIVED 20 November 2023
ACCEPTED 27 March 2024
PUBLISHED 29 April 2024

CITATION
Wang H, Su B, Zhang Y, Shang M, Wang J, Johnson A, Dilawar H, Bruce TJ, Dunham RA and Wang X (2024), Transcriptome analysis revealed potential mechanisms of channel catfish growth advantage over blue catfish in a tank culture environment.
*Front. Genet.* 15:1341555.
doi: 10.3389/fgene.2024.1341555



# Transcriptome analysis revealed potential mechanisms of channel catfish growth advantage over blue catfish in a tank culture environment

Haolong Wang[1,2†], Baofeng Su[2,3†], Ying Zhang[1,2†], Mei Shang[3], Jinhai Wang[3], Andrew Johnson[3], Hamza Dilawar[3], Timothy J. Bruce[3], Rex A. Dunham[2,3] and Xu Wang[1,2,4,5*]

[1]Department of Pathobiology, College of Veterinary Medicine, Auburn University, Auburn, AL, United States, [2]Auburn University Center for Advanced Science, Innovation, and Commerce, Alabama Agricultural Experiment Station, Auburn, AL, United States, [3]School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL, United States, [4]Scott-Ritchey Research Center, College of Veterinary Medicine, Auburn University, Auburn, AL, United States, [5]HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States

Channel catfish (*Ictalurus punctatus*) and blue catfish (*Ictalurus furcatus*) are two economically important freshwater aquaculture species in the United States, with channel catfish contributing to nearly half of the country's aquaculture production. While differences in economic traits such as growth rate and disease resistance have been noted, the extent of transcriptomic variance across various tissues between these species remains largely unexplored. The hybridization of female channel catfish with male blue catfish has led to the development of superior hybrid catfish breeds that exhibit enhanced growth rates and improved disease resistance, which dominate more than half of the total US catfish production. While hybrid catfish have significant growth advantages in earthen ponds, channel catfish were reported to grow faster in tank culture environments. In this study, we confirmed channel fish's superiority in growth over blue catfish in 60-L tanks at 10.8 months of age (30.3 g and 11.6 g in this study, respectively; $p < 0.001$). In addition, we conducted RNA sequencing experiments and established transcriptomic resources for the heart, liver, intestine, mucus, and muscle of both species. The number of expressed genes varied across tissues, ranging from 5,036 in the muscle to over 20,000 in the mucus. Gene Ontology analysis has revealed the functional specificity of differentially expressed genes within their respective tissues, with significant pathway enrichment in metabolic pathways, immune activity, and stress responses. Noteworthy tissue-specific marker genes, including *lrrc10*, *fabp2*, *myog*, *pth1a*, *hspa9*, *cyp21a2*, *agt*, and *ngtb*, have been identified. This transcriptome resource is poised to support future investigations into the molecular mechanisms underlying environment-dependent heterosis and advance genetic breeding efforts of hybrid catfish.

KEYWORDS

**aquaculture, catfish growth phenotype, tissue, organ, transcriptome, expression marker gene**

# Introduction

Channel catfish (*Ictalurus punctatus*) and blue catfish (*Ictalurus furcatus*) are two native North American catfish species. Blue catfish are the largest catfish in the US and reach sexual maturity at older ages than channel catfish (Graham, 1999). Blue catfish grow slower than channel catfish during the first 2 years. As a result, channel catfish reach market size before blue catfish. Channel catfish is traditionally considered the most important and popular species for catfish farmers and producers in the US. However, channel catfish is not as resistant as blue catfish to enteric septicemia of catfish (ESC), which causes an annual economic loss of 50 million dollars (Wolters and Johnson, 1994; Yeh et al., 2005).

To combat pathogenic infections in channel catfish, genetic enhancement of catfish was achieved through interspecific hybridization. The hybrid between female channel catfish and male blue catfish (*I. punctatus* × *I. furcatus*) constitutes more than 50% of the total US catfish production (Torrans and Ott, 2018), and the hybrid catfish grow 20%–100% faster than commonly cultured strains of channel catfish, depending upon the environment (Dunham et al., 1990; Dunham et al., 2008). The hybrid catfish combines many of the best traits of their parental species and is a highly desirable fish in commercial pond culture. It offers improved feed conversion efficiency (Dunham et al., 2008; Brown et al., 2011), increased carcass yield (Bosworth, 2012), better low oxygen tolerance (Dunham et al., 1983), disease resistance (Arias et al., 2012), and enhanced harvestability (Dunham and Masser, 2012). However, heterosis in growth is environment-dependent. Catfish fry are typically reared in indoor tanks, and juvenile catfish will be transferred from indoor tanks to earthen ponds for aquaculture. Previous studies on the growth trait have shown that hybrid catfish are not superior in the tank culture environment, whereas channel catfish had a growth advantage instead (Dunham et al., 1990). The molecular mechanism of growth advantage in channel catfish is still poorly understood. Gene expression research sheds light on the physiology of a set of cells or tissues at a certain period, including cellular adaptations to different environments (Singh et al., 2018). Although organism cells carry out similar processes for key biological functions in their own tissue environment, they display unique functions that support the definition of their phenotype (Sonawane et al., 2017). Characterizing gene expression differences between the two catfish species across various tissues will provide valuable insights into understanding the phenotypic variations in growth and disease resistance and serve as the first pass to investigate molecular mechanisms underlying environment-dependent heterosis.

RNA sequencing is a transcriptome-wide approach used to characterize gene expression profiles in various catfish species. Several studies have recently been conducted to analyze differentially expressed genes (DEGs) and functional pathways in blue catfish, channel catfish, and their hybrids. For instance, in the liver transcriptome, a group of genes associated with fatty acid metabolism was discovered to be significantly upregulated in channel catfish compared to blue catfish and hybrids (Wang et al., 2022a). In another study, a set of DEGs involved in the formation of the swim bladder were identified between channel catfish and other catfish. These genes were enriched in the Wnt signaling pathway and the hedgehog signaling pathway (Yang et al., 2018). Taking advantage of RNA sequencing, these findings shed light on the distinctive genetic characteristics and potential functional differences between these catfish species and their hybrids. However, a broader organ selection is still needed to better understand the transcriptomes in channel catfish and blue catfish.

In this study, five organs (heart, liver, muscle, mucus, and intestine) were selected from channel catfish and blue catfish transcriptome characterization at the 10.8-month juvenile fish stage. Muscle development and muscle growth are directly relevant to the overall quality of fish meat (Fuentes et al., 2013; Xu et al., 2019). The heart plays a crucial role in pumping oxygen and nutrients and dealing with environmental stress (Saetan et al., 2021). The liver is involved in various metabolic processes that support fish growth, including but not limited to metabolism, nutrient storage, energy production, and detoxification (Zhang et al., 2021). Mucus serves as a protective layer against pathogens, parasites, and environmental toxins, as well as a barrier to fight infection through its immune-related functions (Lange et al., 2018). The intestine also has an immune function to defend against harmful microbes and interact with the gut microbiota. In addition, it is the major organ for digestion and nutrient assimilation, which promotes growth. The comparative transcriptomic analyses provide insights into gene function differences between the two species and the molecular basis of the channel catfish's growth advantage in the tank culture environment.

# Materials and methods

## Fish maintenance and tissue sample collection

The experimental animal protocols regarding animal care and tissue collections were approved by the Auburn University Institutional Animal Care and Use Committee (AU-IACUC) with the approval number PRN-2019-3520. Blue catfish (BB) and channel catfish (CC) were cultured at the Auburn University Fish Genetics Research Unit in Auburn, Alabama, United States. Both catfish species were maintained in the indoor recirculatory aquaculture system with separate 60-L rectangular tanks (60 cm × 23 cm × 43.5 cm) at an initial density of 1,000 fry per tank. The fry were fed with Purina® AquaMax® Fry Starter 100 for the first 3 months. At 2 months old, fry density was adjusted to 100 fry per tank and then to 50 fish per tank at 4 months old. Starting at 4 months, the fry were fed with Purina® AquaMax® Fry Starter 200 for 3 months and then with Purina® AquaMax® Fry Starter 300 three times a day. Dissolved oxygen was maintained above 5 mg/L, with pH levels between 7.0 and 7.5. At 10.8 months of age, three randomly selected fish from each species were euthanized with 300 ppm tricaine methanesulfonate (MS-222, Syndel Inc., Ferndale, WA, United States). Muscle, liver, intestine, mucus, and heart tissues were dissected immediately after euthanasia. All tissue samples were flash-frozen in liquid nitrogen and stored in a −80°C freezer until RNA extraction.

## RNA extraction, library preparation, and sequencing

Three replicates were performed for each tissue of the two catfish species at 10.8 months of age. The total RNA was extracted using the

Quick RNA Microprep Kit (Zymo Research, Irvine, CA, United States) following the manufacturer's protocol. RNA concentrations were measured using a NanoDrop One$^{C}$ Microvolume UV-Vis Spectrophotometer (Thermo Scientific, Waltham, MA, United States). The library for each tissue sample was constructed using the NEBNext® Poly(A) mRNA Magnetic Isolation Module and NEBNext® Ultra™ II RNA Library Prep Kit (New England BioLabs, Ipswich, MA, United States) with 1 μg of total RNA input. The library PCR amplifications were conducted using 18 cycles. The concentration of sequencing libraries was quantified using a Qubit 3.0 Fluorometer (Thermo Scientific, Waltham, MA, United States), and the average size of cDNA libraries was evaluated with the D1000 ScreenTape assay using the TapeStation 4,200 System (Agilent Technologies, Santa Clara, CA, United States). The libraries were sequenced on an Illumina NovoSeq6000 sequencer to generate 2 × 150-bp paired-end reads at Novogene (Novogene Corporation Inc., Sacramento, CA, United States).

## RNA sequencing analysis

The quality of the raw reads was assessed by FastQC (version 0.11.6) (Andrews, 2010). Low-quality bases and adapter sequences were trimmed using Trimmomatic (version 0.39), and sequencing reads shorter than 36 bp in length were excluded from subsequent analysis (Bolger et al., 2014). RNA-seq reads were aligned to the blue catfish (*I. furcatus*) reference genome (Wang et al., 2022b) using STAR aligner version 2.7.5c (Dobin et al., 2013). The gene read counts of each sample were quantified and summarized using HTseq version 1.0 (Anders et al., 2015). Genes with extremely low expression values in all tissues were excluded, and genes with counts >1 in at least three samples were retained for subsequent analysis.

## Identification of differentially expressed genes

To determine gene expression levels for each sample, read counts were normalized using the edgeR package in R (version 3.6.4) (Robinson et al., 2010). The differentially expressed genes (DEGs) between CC and BB for each tissue sample were identified using the cutoff of |log2 (fold change)| > 1.5 and a false discovery rate (FDR) < 0.05. The Benjamini–Hochberg method was used to determine the adjusted *p*-values.

## Gene ontology and functional enrichment analysis

For functional enrichment analysis, blue catfish genes were mapped to the zebrafish (*Danio rerio*) assembly GRCz11 (Howe et al., 2013) using DIAMOND version 2.0.0 (Buchfink et al., 2021) to determine the gene names. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa and Goto, 2000) analyses were performed using Metascape (Zhou et al., 2019) with default parameters. The DEG gene symbols were used as the input gene list. The optional parameters of "input as species" and "analysis as species" were selected as "any species" and "zebrafish," respectively. The GO terms analysis was conducted for biological processes, cellular components, and molecular functions.

## Analysis of tissue-specific gene expression

The τ index was used to determine the tissue specificity (Yanai et al., 2005) in gene expression, which ranges from 0 (non-specific, expressed equally in all tissues) to 1 (highly specific, only expressed in one tissue). For each gene, τ is computed according to the formula $\tau \text{ index} = \frac{\sum_i^N (1-x_i)}{N-1}, i = 1$ , where $N$ is the number of tissues, and $x_i$ is the normalized expression value. A cutoff of $\tau > 0.9$ was used to detect tissue-specific genes (Yanai et al., 2005).

## Quantitative reverse transcription PCR validation of tissue-specific genes

Quantitative reverse transcription PCR (qRT-PCR) experiments were performed to validate the tissue-specific genes identified from RNA-seq data. One candidate gene from each group (channel catfish only, blue catfish only, and both species) was selected, including *fabp2*, *cyp21a2*, and *pth1a*. A housekeeping gene, *gapdh*, was included as a reference. The relative gene expression levels of these genes were quantified in the heart, intestine, liver, mucus, and muscle of channel catfish and blue catfish, with three replicates for each tissue. Primer sequences used for qRT-PCR validation are listed in Supplementary Table S4. The first-strand cDNA synthesis was conducted using the LunaScript® RT SuperMix Kit (New England BioLabs, Ipswich, MA, United States) with 1 μg of total RNA, following the manufacturer's protocol. The same total RNA samples for the RNA-seq experiments were used for validation. Quantitative reverse transcription PCR was performed on a Bio-Rad C1000 Touch Thermal Cycler with CFX96 Real-Time PCR Detection Systems (Bio-Rad Laboratories, Hercules, CA, United States) in a 20 μL final reaction volume. The reaction mixture included 10 μL of Luna universal qPCR Master Mix, 0.5 μL of each primer, 6 μL of nuclease-free water, and 3 μL of cDNA template. The standard amplification protocol was 95°C for 60 s, followed by 40 cycles at 95°C for 15 s and 60°C for 30 s with two technical replicates. The relative gene expression value was computed using the $2^{-\Delta\Delta CT}$ method.

# Results

## Channel catfish exhibit superior growth during the early life stages of tank culture

Although heterosis has been reported in pond culture, channel catfish (CC, channel catfish × channel catfish) exhibit superior growth compared to blue catfish (BB, blue catfish × blue catfish) and their reciprocal hybrids (CB, channel catfish female × blue catfish male, and BC, blue catfish female × channel catfish male) in tanks and other smaller culturing units (Figure 1A) (Dunham et al., 1987; Dunham et al., 1990; Argue et al., 2014; Wang et al., 2022a). In this study, we measured the body weight at 3 weeks and 10.8 months of age for N = 20 fish of each of the four genetic types (CC, BB, CB,

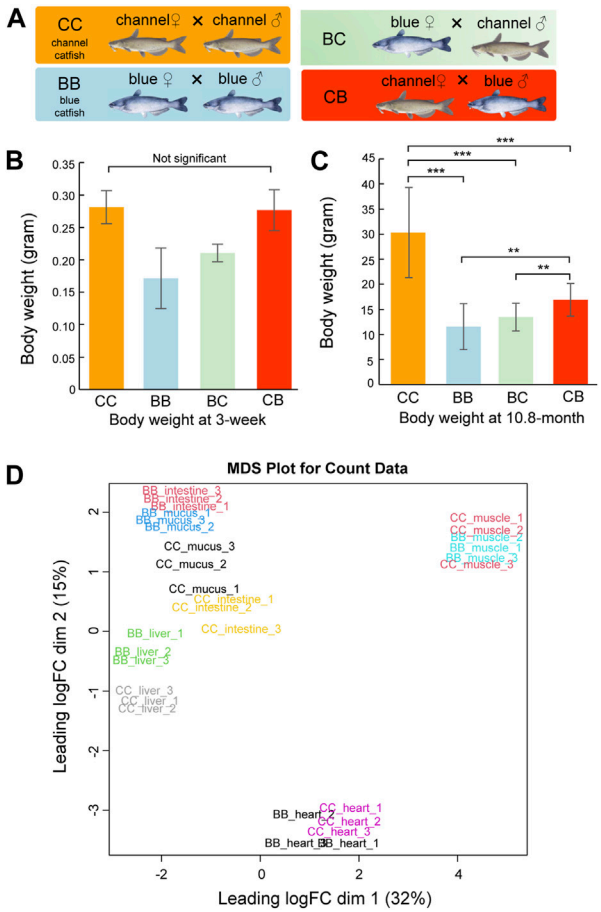



FIGURE 1
Body weight measurements and multi-dimensional scaling (MDS) plot of transcriptome quantification from five tissues of channel catfish and blue catfish. **(A)** Four genetic types of channel catfish *Ictalurus punctatus* (CC), blue catfish *Ictalurus furcatus* (BB), and their reciprocal hybrids (BC and CB). **(B)** Barplot of body weight for four genetic types at 3 weeks of age. **(C)** Barplot of body weight for four genetic types at 10.8 months of age. Statistical significance was assessed by the nonparametric Mann–Whitney *U* test (**, $p < 0.01$; ***, $p < 0.001$). **(D)** MDS plot of five tissues in blue catfish (BB) and channel catfish (CC). The expressed gene counts were used as input. The *x*-axis and *y*-axis represent the first two dimensions.To graphics: This image should be moved to figure 1

and BC). No significant difference was observed between channel catfish and blue catfish at 3 weeks ($p > 0.05$; Figure 1B), but channel catfish were significantly heavier at 10.8 months than blue catfish (30.3 g vs 11.6g, $p < 0.001$; Figure 1C and Data S1). The difference in body weight suggests that the growth rate of channel catfish is approximately three times higher than that of blue catfish, which is consistent with previous studies (Wang et al., 2022c).

## Transcriptome-wide expression profiling revealed significant differences in gene expression in the liver, mucus, and intestine between channel catfish and blue catfish

To investigate gene expression differences in important organs between channel catfish and blue catfish, the heart, liver, intestine, mucus, and muscle were selected for transcriptome analysis at 10.8 months of age (Figure 1D and Supplementary Table S1). In total, 697 million 150-bp reads (209.2 Gbp of sequences) were generated (Supplementary Table S2). On average, 73% of RNA-seq reads were uniquely mapped to the blue catfish reference genome (Wang et al., 2022b). To investigate the tissue-specific gene expression profiles in both species, a multi-dimensional scaling (MDS) plot was generated using normalized gene counts and the transcriptomic profiles clustered together by tissue in the first two dimensions (Figure 1E). Overall, tissues exhibit a higher degree of resemblance than species, reflecting functional similarities among individual tissues. Dimension 1 separated skeletal muscle and heart from the remaining tissues, which is consistent with the fact that they are derived from mesoderm (Figure 1E). The skeletal muscle and heart tissues from BB and CC are intermingled, whereas liver, intestine, and skin mucus samples from the two species are well separated (Figure 1E). The results suggest that skeletal muscle and heart muscle are more functionally conserved than other tissues. Notably, skin mucus and intestine transcriptomic profiles are more similar within species. Mucus-secreting cells are also present in the intestine, and both organs are in contact with the microbiota (gut and skin microbiota).

10.3389/fgene.2024.1341555

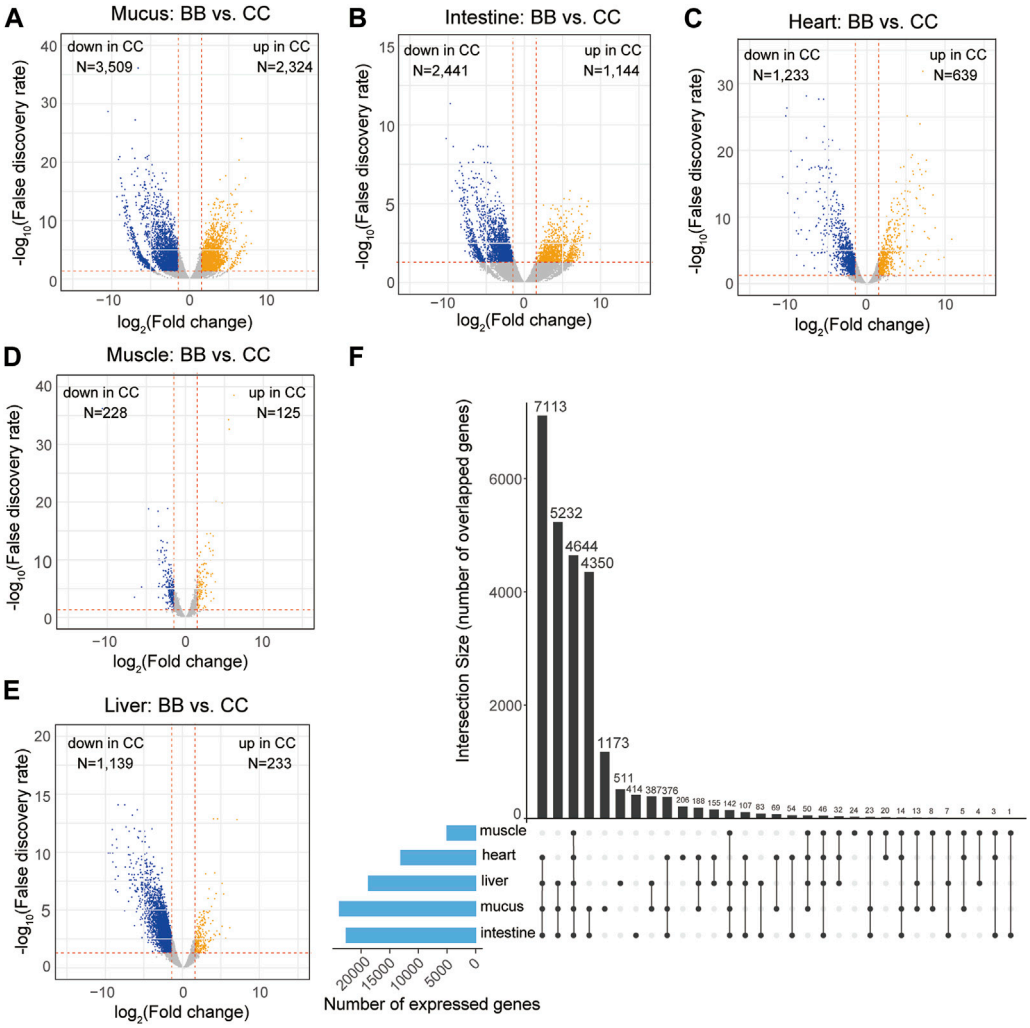



FIGURE 2
Transcriptome-wide differentially expressed genes (DEGs) in five organs between channel catfish and blue catfish. Volcano plots of pairwise comparisons in the mucus **(A)**, intestine **(B)**, heart **(C)**, muscle **(D)**, and liver **(E)**. DEGs with a false discovery rate (FDR) < 0.05 are highlighted. The *x*-axis stands for log2 (fold change), and the *y*-axis represents $-\log_{10}$(FDR). The vertical lines indicate |log2FoldChange| = 1.5. **(F)** Upset plot showing the intersection of expressed gene sets across five organs between channel catfish and blue catfish.

## Detection of differentially expressed genes between channel catfish and blue catfish in five organs

The expression levels of each gene were determined using the reads per kilobase of transcript per million mapped reads (RPKM). To identify DEGs, a pairwise analysis of differential gene expression was conducted between channel catfish and blue catfish (Data S2). Among these five organs, the mucus exhibited the largest number of DEGs, with 2,324 upregulated and 3,509 downregulated DEGs in channel catfish compared to blue catfish (Figure 2A). The intestine and heart also had more than 2,000 DEGs between channel catfish and blue catfish (Figures 2B,C). In contrast, muscle displayed the lowest number of DEGs, with only 125 upregulated DEGs and 228 downregulated DEGs (Figure 2D), suggesting functional conservation between species. There were ~1,400 DEGs identified in the liver samples (Figure 2E).

To compare the overall expression profiles among five organs, the numbers of expressed genes (RPKM >2) in each organ were investigated (Figure 2F). The mucus transcriptome had the largest number of expressed genes (>20,000 genes), whereas the muscle had only 5,036 genes detected, indicating considerable variation among organs. A total of 4,644 genes were found to be shared among all five organs (Figure 2F), accounting for ~90% of the muscle transcriptome or ~25% of the mucus transcriptome. The liver, heart, intestine, and mucus exhibited the largest overlapping gene set, with 7,113 expressed genes common to these organs. More than 4,000 genes were expressed exclusively in the mucus and intestine, which is consistent with the similar gene expression profiles depicted in the MDS plot (Figure 1E). Regarding genes that were only expressed within an individual organ, mucus had the largest number of organ-specific genes (1,173), while muscle tissue had the lowest number, with only 24 such genes identified (Figure 2F).

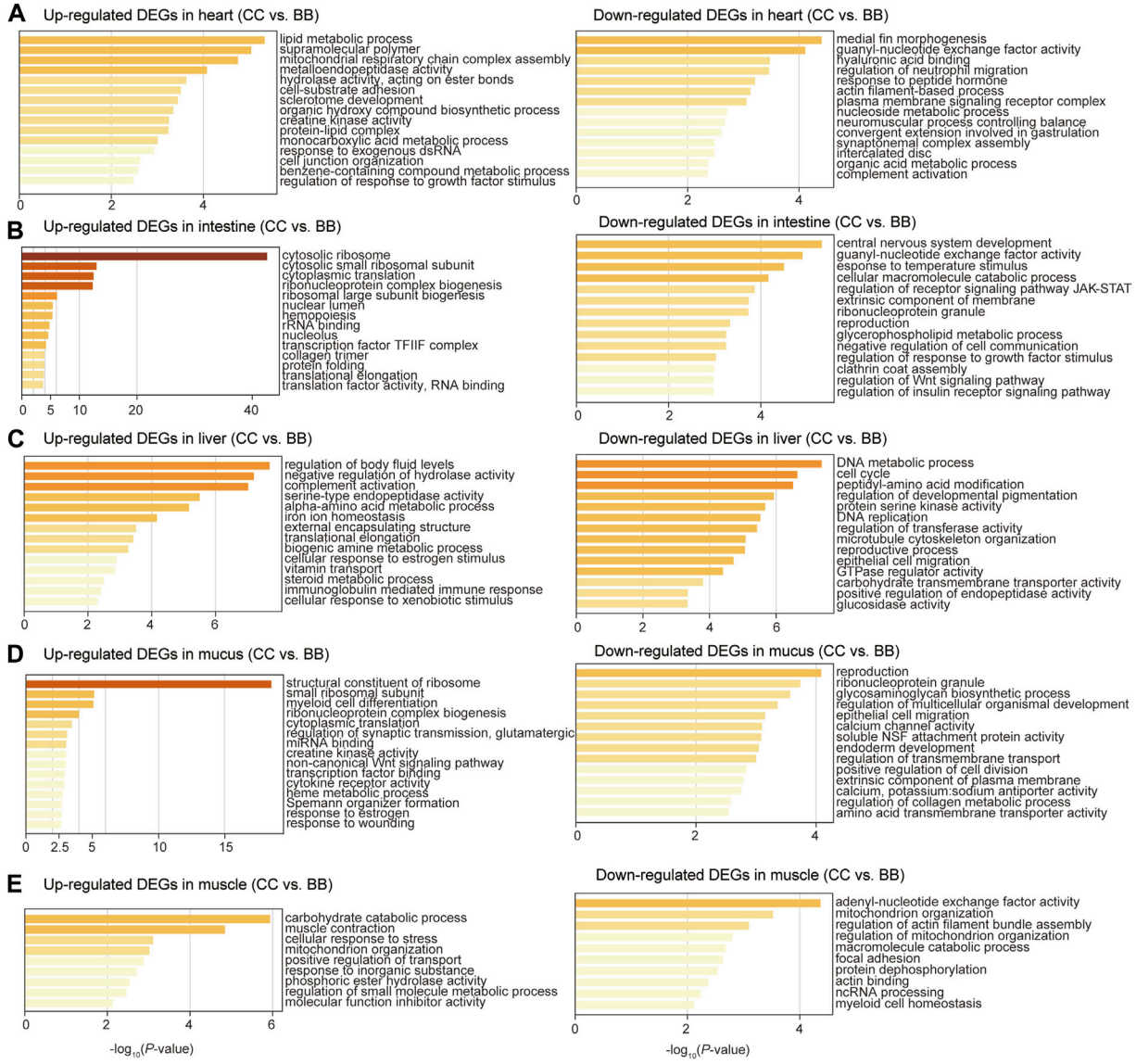



FIGURE 3
Gene ontology enrichment analysis of differentially expressed genes between channel catfish and blue catfish. Enrichment scores (*x*-axis) were determined by $-\log_{10}$(*p*-value) for significantly enriched terms of upregulated and downregulated genes in channel catfish in the heart **(A)**, intestine **(B)**, liver **(C)**, mucus **(D)**, and muscle **(E)**.

## Stress response, immune activity, and metabolic pathways are enriched among DEGs

To identify the biological function of DEGs in each tissue, gene ontology (GO) enrichment analyses were performed comparing blue catfish and channel catfish. The upregulated DEGs were defined as genes with higher expression levels in channel catfish than those in blue catfish. In the heart, upregulated DEGs were significantly enriched in the lipid metabolic process ($p < 10^{-5}$), while downregulated DEGs clustered in immunity-related terms such as regulation of neutrophil migration and complement activation (Figure 3A). In the intestine, the most significant function term was "cytosolic ribosome" in the upregulated DEGs ($p < 10^{-40}$; Figure 3B). In contrast, the downregulated DEGs were mainly enriched in response to temperature stimulus ($p < 10^{-4}$), which is associated with the stress response process (Figure 3B). In the liver, upregulated DEGs were primarily enriched in the cellular lipid metabolic process ($p < 0.001$; Figure 3C). Interestingly, downregulated DEGs were also clustered in metabolic terms such as carboxylic acid metabolic process, NADP metabolic process, and cellular lipid metabolic process (Figure 3C). In mucus, the enrichment analysis revealed that upregulated DEGs were significantly associated with the structural constituents of ribosomes ($p < 10^{-17}$), creatine kinase activity, cytokine receptor activity, and response to wounding (Figure 3D). For muscle tissues, the top three most enriched terms among upregulated DEGs were carbohydrate catabolic process, muscle contraction, and cellular response to stress (Figure 3E).

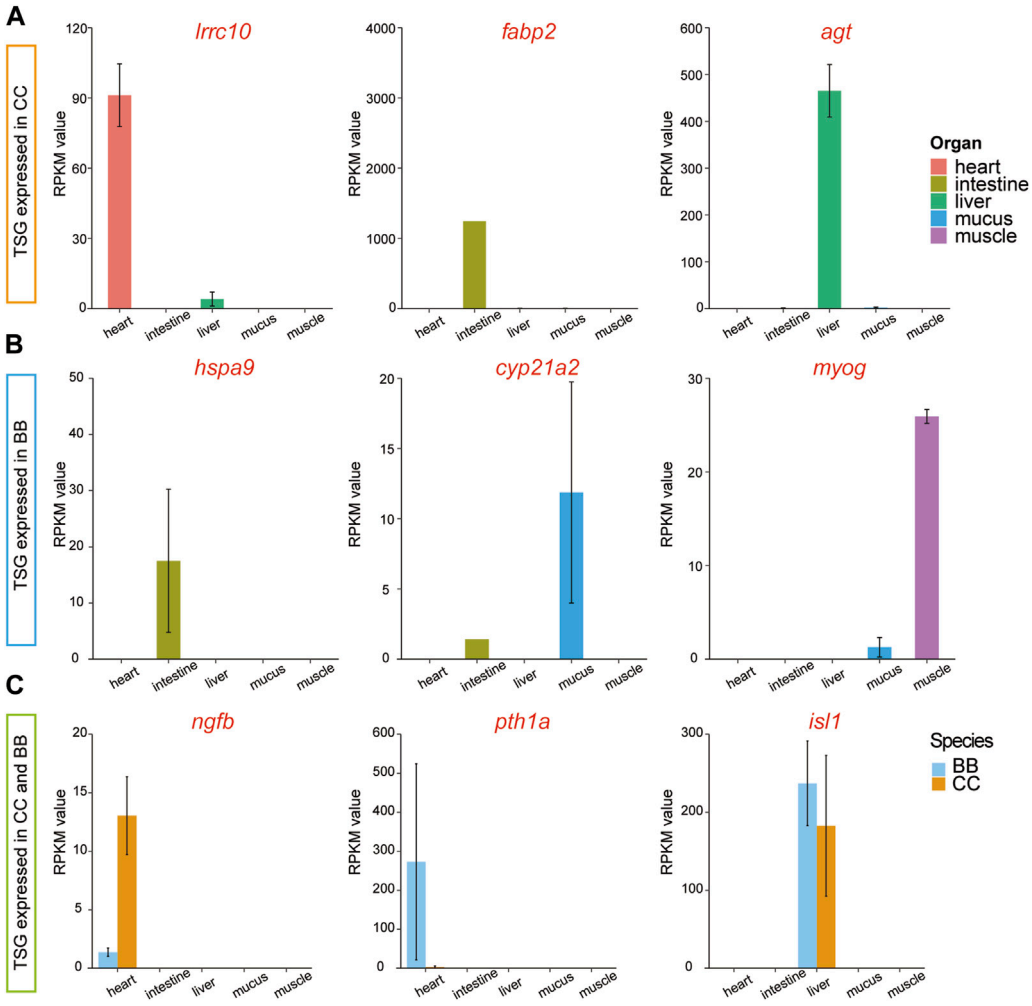



FIGURE 4
Tissue-specific and species-specific gene candidates in channel catfish and blue catfish at 10.8 months of age. Barplots of RNA-seq reads per kilobase of transcript per million mapped reads (RPKM) values. **(A)** Tissue-specific genes (*lrrc10*, *fabp2*, and *agt*) found only in channel catfish. **(B)** Tissue-specific genes (*hspa9*, *cyp21a2*, and *myog*) found only in blue catfish. **(C)** Tissue-specific genes (*ngfb*, *pth1a*, and *isl1*) found in channel catfish and blue catfish.To graphics: This image should be moved to figure 4

## Tissue-specific genes in channel catfish and blue catfish

To examine the diversity of expression patterns among tissues, we utilized a tissue specificity index (τ value) to quantify the specificity of the gene profile (Data S3). In this study, a gene with a τ value greater than 0.9 was classified as a tissue-specific gene (TSG). Thus, TSGs were identified across the five tissues in two species. Among the different tissues in channel catfish, the number of TSGs ranged from 33 in muscle tissue to 1,872 in mucus tissue (Supplementary Table S3 and Supplementary Data S4). The distribution of TSG across tissues in blue catfish followed a similar pattern. Among the tissue-specific genes, three highly expressed genes, *irrc10*, *fabp2*, and *agt*, were exclusively detected in channel catfish (Figure 4A). Three different genes, *hspa9*, *cyp21a2*, and *myog*, were identified only in blue catfish. It is worth noting that some tissue-specific genes were expressed in both species, including *ngfb*, *pth1a*, and *isl1* (Figure 4C).

To confirm tissue-specific expression identified from RNA-seq data using an independent approach, qRT-PCR was performed to determine the relative expression levels of selected genes (see Materials and Methods). The relative expression value of *fabp2* was significantly higher in the intestine of channel catfish than in other tissues ($p < 0.05$; Supplementary Figure S1). No *fabp2* expression was detected in blue catfish, indicating that *fabp2* is intestine-biased in channel catfish only. *Cyp21a2* was highly expressed in both the intestine and mucus of blue catfish, while little to no expression was detected in channel catfish (Supplementary Figure S2). *Pth1a* was exclusively expressed in the heart in both species, but its expression level in the blue catfish was over 100 times higher than in the channel catfish (Supplementary Figure S3). The relative expression values of all three genes are consistent with the expression pattern identified from RNA-seq data.

# Discussion

## Tissue-specific transcriptomes in blue catfish and channel catfish

In this study, transcriptomes from five major tissues/organs provide a genomic resource for investigating the transcriptomic differences between two economically important catfish species. Given the variations observed in growth performance and disease resistance between blue catfish and channel catfish, there is a strong interest in understanding local adaptation, genome evolution, and the genetic basis underlying these traits. Over the past decade, expressed sequence tag (EST) sequencing (Li et al., 2007), single nucleotide polymorphisms (SNPs) information (Liu et al., 2011), and full-length cDNAs identification (Chen et al., 2010) have been characterized in blue catfish and channel catfish. Although swimbladder RNA-seq data were reported to investigate the differences in chamber formation between blue catfish and channel catfish (Yang et al., 2018), the transcriptomic divergence study across multiple tissues between these two species is still limited.

Transcriptomes are most commonly used in fish to characterize molecular physiology and identify genes that respond to or ameliorate environmental stresses (Basu et al., 2002; Cossins and Crawford, 2005). Gene expression regulation shows considerable variation among different organs, individuals, and species (Hsieh et al., 2003). All tissue-specific transcriptomes used in this study were from peripheral tissues, including the heart, intestine, liver, mucus, and muscle. Overall, the peripheral tissue transcriptomes separated into distinct clusters on the MDS plot, suggesting a divergence of gene expression patterns. The utilization of peripheral tissue transcriptomes in catfish research can contribute to the understanding of catfish biology and create opportunities for further investigations in various fields, including metabolism, immune responses, development, stress response, and physiology.

## The growth advantage in channel catfish may be associated with metabolic regulation and tissue development

Cardiac tissue plays a vital role in fish physiology as it is responsible for pumping oxygenated blood throughout the fish's body and delivering essential nutrients and hormones to various organs and tissues. Cardiac transcriptome analyses have been very effective in discovering candidate genes in studies of cardiac toxicity, response to hypoxia, and cardiac disease in various fish species (Shih et al., 2015; Saetan et al., 2021; Xiao et al., 2021). Although no treatment was administered in the present study, two heart tissue-specific genes were identified through a comparison of gene profiles among tissues, including *lrrc10* and *pth1a* (Figure 3). *Irrc10*, a highly conserved gene unique to the heart, is implicated in embryonic development and tissue differentiation processes. *Lrrc10* was reported as a cardiac-specific factor (*Serdin1*) in mice that is essential for heart development (Adameyko et al., 2005; Manuylov et al., 2008). In zebrafish, the *Lrrc10* morphants exhibited cardiac functional defects, as evidenced by a decrease in ejection fraction and cardiac output (Kim et al., 2007). The *pth1a* gene encodes a protein called parathyroid hormone receptor 1 (PTH1 receptor), which plays a crucial role in the regulation of calcium and phosphate homeostasis. This is consistent with the downregulated DEGs enriched in response to peptide hormones (Figure 3A). It was also shown to play a role in bone remodeling, which involves the continuous breakdown and formation of bone tissue, in zebrafish (Aceto et al., 2015).

Fish intestine serves several important functions related to digestion, nutrient absorption, body fluid balance, and immune defense, which are critical for growth and disease-resistance phenotypes. A dramatically upregulated intestine-specific gene in channel catfish, *fabp2* (fatty acid-binding protein 2; Figure 4A), encodes an intestinal fatty acid-binding protein (I-FABP). Fatty acid-binding proteins (FABPs) play a crucial role in the transcriptional regulation of genes associated with lipid metabolism, which can significantly impact fat deposition in animals (Venkatachalam et al., 2018). *Fabp2* was initially identified in mammals and is expressed exclusively in the intestine (Gajda and Storch, 2015). It has also been reported in fish species with variable expression patterns (Sharma et al., 2004; Venkatachalam et al., 2017). Numerous studies have provided evidence that the expression of FABP in fish is regulated by nutritional factors (Venold et al., 2013; Xu et al., 2017). Starvation stress affects the expression level of FABP in various fish species, typically leading to downregulation in response to prolonged periods of starvation (Kaitetzidou et al., 2015; Ölmez et al., 2015). Therefore, alterations in FABP levels and gene expression can serve as indicators of lipid accumulation. During the dissection process in the present study, a substantial presence of white adipose tissues was notably observed in channel catfish, indicating that channel catfish tend to accumulate a greater amount of energy sources than blue catfish at the 10.8-month developmental stage.

Skeletal muscle constitutes the major portion of the fish trunk, comprising approximately 40%–60% of the total body weight (Xu et al., 2019). Fish muscle performs a variety of physiological functions associated with locomotion, movement, and metabolism. The regulation of muscle fiber development and growth is maintained by myogenic regulatory factor (MRF) genes, including *myod*, *myf5*, *myog*, and *mrf4*. Among these genes, *myog* is a crucial member of the myogenic regulator family, responsible for governing the differentiation of mesodermal cells into myoblasts, which subsequently form the muscle fibers. Notably, *myog* is the only gene among the MRFs expressed in all skeletal muscle cell lines (Hasty et al., 1993). It was also reported that the MRF genes are regulated by the GH-insulin-like growth factor (IGF) axis (Fuentes et al., 2013). Gene silencing and knockout of the myostatin gene have been found to promote somatic growth in many fish species, such as zebrafish (Gao et al., 2016), medaka (Chiang et al., 2016), and channel catfish (Khalil et al., 2017). In the present study, the highly expressed *myog* gene in the muscle of channel catfish may contribute to the difference in growth performance between blue catfish and channel catfish in a tank environment (Figure 4B).

## Difference in stress response and immune activity between blue catfish and channel catfish

Mucus provides the mucosal barrier as the first line of defense against pathogens. The secretion patterns of mucus not only

influence the rates of bacterial shedding but also play important roles in the production of enzymes, antimicrobial peptides, and secreted immunoglobulins (Xu et al., 2013). Comparing the mucus transcriptome between blue catfish and channel catfish revealed a significant number of DEGs, indicating potential differences in immune response mechanisms. Specifically, the upregulated DEGs in mucus were enriched with gene ontology terms related to cytokine receptor activity (GO: 0004896) and response to wounding (GO: 0009611), suggesting unique immune patterns in channel catfish (Figure 3D). Blue catfish also have their own set of GO terms for stress response. For example, a mucus-specific gene, *cyp21a2* was highly expressed in blue catfish compared to channel catfish (Figure 4B). It encodes the cytochrome P450 enzyme 21-hydroxylase, which plays a crucial role in catalyzing a key step in the biosynthesis of glucocorticoids (such as cortisol) and mineralocorticoids (Miller and Auchus, 2011). In fish, cortisol serves as the primary circulating glucocorticoid, with effects mediated through the glucocorticoid receptor (GR) (Faught and Vijayan, 2016). The primary function of *cyp21a2* is to facilitate the production of cortisol, which is involved in regulating metabolism, immune responses, stress response, and maintaining homeostasis in the body. In zebrafish, *cyp21a2* knockout induced a reduction in cortisol levels (Eachus et al., 2017). As a common stress indicator, the higher cortisol levels in blue catfish may indicate that this species is more sensitive to environmental and psychological stress. Further study needs to be conducted at the physiological level to confirm this. In addition to glucocorticoids, *cyp21a2* is also responsible for the production of mineralocorticoids, primarily aldosterone. Aldosterone is involved in regulating sodium and potassium balance, blood pressure, and fluid balance in the body.

The liver plays an important role in the metabolism, detoxification, nutrient storage, synthesis of blood proteins, and immune function of fish. The *agt* gene, also known as angiotensinogen, encodes the angiotensinogen protein. Angiotensinogen is a precursor protein that plays a significant role in the renin-angiotensin system (RAS) (Nishimura, 2004), a hormonal cascade involved in regulating blood pressure, fluid balance, and sodium homeostasis. The *agt* gene is also likely to be involved in the immune response, as a previous study indicated that the expression level of *agt* significantly increased after bacterial infection in ayu (Chen et al., 2008).

Heat tolerance is a critical trait in aquaculture species (Tan et al., 2019). Heat shock proteins (HSPs) belong to a superfamily of proteins that are triggered by various stressors, including physical, chemical, and biological factors, such as high temperature, hypoxia, infection, and toxins (Kregel, 2002). HSP70 is a widely recognized stress protein in aquatic organisms, playing a crucial role in stress responses, including thermotolerance (Bertotto et al., 2011), and also participating in the regulation of the immune system (Tsan and Gao, 2009). In the present study, *hspa9*, as a member of the heat shock protein 70 (HSP70) family, was found to be highly expressed in blue catfish intestine tissue compared to channel catfish (Figure 4B), indicating potential stress susceptibility in blue catfish. In Japanese flounder, the *hspa9* was identified with a high level of expression in the transcriptome after infection with *Edwardsiella tarda*.

The sensory and neural systems enable fish species to perceive the world around them and respond appropriately to the environment (Borghezan et al., 2021). The *ngfb* gene encodes the nerve growth factor beta (NGFβ) protein, which plays a vital role in the development and survival of nerve cells, particularly sensory neurons responsible for transmitting pain, temperature, and touch sensations. A significant difference in *ngfb* expression levels was observed in this study between channel catfish and blue catfish (Figure 4C), which may impact their function. It was found that *ngfb* was downregulated in the hippocampal neurons induced by lipopolysaccharides (LPS) (Fang et al., 2020). Collectively, divergence in multiple organs may contribute to the differences in stress response and immune activity between channel catfish and blue catfish.

## Data availability statement

The raw RNA-seq datasets and gene read counts for this study can be found in the NCBI GEO (Gene Expression Omnibus) databases under the accession number GSE247918.

## Ethics statement

The animal study was approved by Auburn University Institutional Animal Care and Use Committee. The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

HW: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing–original draft, and writing–review and editing. BS: conceptualization, investigation, methodology, supervision, writing–original draft, and writing–review and editing. YZ: data curation, formal analysis, investigation, methodology, visualization, writing–original draft, and writing–review and editing. MS: investigation, methodology, and writing–review and editing. JW: investigation, methodology, and writing–review and editing. AJ: investigation, methodology, and writing–review and editing. HD: investigation, methodology, and writing–review and editing. TB: conceptualization, resources, supervision, and writing–review and editing. RD: conceptualization, project administration, resources, supervision, writing–original draft, and writing–review and editing. XW: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, visualization, writing–original draft, and writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study is supported by a laboratory start-up fund from Auburn University College of Veterinary Medicine and a USDA-NIFA award 2023-67016-39455. XW is also supported by a National Science Foundation award 1928770, a Scott-Fund seed project award

from the Scott-Ritchey Research Center, the USDA National Institute of Food and Agriculture (NIFA) Hatch project USDA-NIFA-ALA05-2-18041, and an Alabama Agriculture Experiment Station (AAES) Agriculture Research Enhancement, Exploration, and Development (AgR-SEED) award. HaW is supported by the Auburn University Presidential Graduate Research Fellowship and the College of Veterinary Medicine Dean's Fellowship.

## Acknowledgments

The authors thank Auburn University Easley Cluster for the computational support of this work.

## 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## 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.2024.1341555/full#supplementary-material

## References

Aceto, J., Nourizadeh-Lillabadi, R., Marée, R., Dardenne, N., Jeanray, N., Wehenkel, L., et al. (2015). Zebrafish bone and general physiology are differently affected by hormones or changes in gravity. *PloS one* 10, e0126928. doi:10.1371/journal.pone.0126928

Adameyko, I. I., Mudry, R. E., Houston-Cummings, N. R. M., Veselov, A. P., Gregorio, C. C., and Tevosian, S. G. (2005). Expression and regulation of mouse SERDIN1, a highly conserved cardiac-specific leucine-rich repeat protein. *Dev. Dyn. official Publ. Am. Assoc. Anatomists* 233, 540–552. doi:10.1002/dvdy.20368

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Andrews, S. (2010). "FastQC: a quality control tool for high throughput sequence data," in *Babraham bioinformatics* (Cambridge, United Kingdom: Babraham Institute).

Argue, B. J., Kuhlers, D. L., Liu, Z., and Dunham, R. A. (2014). Growth of channel catfish (*Ictalurus punctatus*), blue catfish (*I. furcatus*), and their F1, F2, F3, and F1 reciprocal backcross hybrids in earthen ponds. *J. Animal Sci.* 92, 4297–4305. doi:10.2527/jas.2013-7549

Arias, C. R., Cai, W., Peatman, E., and Bullard, S. A. (2012). Catfish hybrid *Ictalurus punctatus* × *I. furcatus* exhibits higher resistance to columnaris disease than the parental species. *Dis. aquatic Org.* 100, 77–81. doi:10.3354/dao02492

Basu, N., Todgham, A., Ackerman, P., Bibeau, M., Nakano, K., Schulte, P., et al. (2002). Heat shock protein genes and their functional significance in fish. *Gene* 295, 173–183. doi:10.1016/s0378-1119(02)00687-x

Bertotto, D., Poltronieri, C., Negrato, E., Richard, J., Pascoli, F., Simontacchi, C., et al. (2011). Whole body cortisol and expression of HSP70, IGF-I and MSTN in early development of sea bass subjected to heat shock. *General Comp. Endocrinol.* 174, 44–50. doi:10.1016/j.ygcen.2011.08.003

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

Borghezan, E. A., Pires, THDS, Ikeda, T., Zuanon, J., and Kohshima, S. (2021). A review on fish sensory systems and amazon water types with implications to biodiversity. *Front. Ecol. Evol.* 8, 589760. doi:10.3389/fevo.2020.589760

Bosworth, B. G. (2012). Effects of winter feeding on growth, body composition, and processing traits of co-cultured Blue Catfish, Channel Catfish, and Channel Catfish× Blue Catfish hybrids. *North Am. J. Aquac.* 74, 553–559. doi:10.1080/15222055.2012.686958

Brown, T. W., Chappell, J. A., and Boyd, C. E. (2011). A commercial-scale, in-pond raceway system for Ictalurid catfish production. *Aquac. Eng.* 44, 72–79. doi:10.1016/j.aquaeng.2011.03.003

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. methods* 18, 366–368. doi:10.1038/s41592-021-01101-x

Chen, F., Lee, Y., Jiang, Y., Wang, S., Peatman, E., Abernathy, J., et al. (2010). Identification and characterization of full-length cDNAs in channel catfish (*Ictalurus punctatus*) and blue catfish (*Ictalurus furcatus*). *PLoS One* 5, e11546. doi:10.1371/journal.pone.0011546

Chen, J., Shi, Y. H., Li, M. Y., Ding, W. C., and Niu, H. (2008). Molecular cloning of liver angiotensinogen gene in ayu (Plecoglossus altivelis) and mRNA expression changes upon Aeromonas hydrophila infection. *Fish Shellfish Immunol.* 24, 659–662. doi:10.1016/j.fsi.2008.01.015

Chiang, Y.-A., Kinoshita, M., Maekawa, S., Kulkarni, A., Lo, C.-F., Yoshiura, Y., et al. (2016). TALENs-mediated gene disruption of myostatin produces a larger phenotype of medaka with an apparently compromised immune system. *Fish Shellfish Immunol.* 48, 212–220. doi:10.1016/j.fsi.2015.11.016

Cossins, A. R., and Crawford, D. L. (2005). Fish as models for environmental genomics. *Nat. Rev. Genet.* 6, 324–333. doi:10.1038/nrg1590

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Dunham, R. A., Brummett, R. E., Ella, M. O., and Smitherman, R. O. (1990). Genotype-environment interactions for growth of blue, channel and hybrid catfish in ponds and cages at varying densities. *Aquaculture* 85, 143–151. doi:10.1016/0044-8486(90)90013-d

Dunham, R. A., and Masser, M. P.: Production of hybrid catfish. Southern Regional Aquaculture Center Stoneville, Mississippi; 2012.

Dunham, R. A., Smitherman, R. O., and Goodman, R. K. (1987). Comparison of mass selection, crossbreeding, and hybridization for improving growth of channel catfish. *Progressive Fish-Culturist* 49, 293–296. doi:10.1577/1548-8640(1987)49<293:comsca>2.0.co;2

Dunham, R. A., Smitherman, R. O., and Webber, C. (1983). Relative tolerance of channel x blue hybrid and channel catfish to low oxygen concentrations. *Progressive Fish-Culturist* 45, 55–57. doi:10.1577/1548-8659(1983)45[55:rtocxb]2.0.co;2

Dunham, R. A., Umali, G. M., Beam, R., Kristanto, A. H., and Trask, M. (2008). Comparison of production traits of NWAC103 channel catfish, NWAC103 channel catfish× blue catfish hybrids, Kansas Select 21 channel catfish, and blue catfish grown at commercial densities and exposed to natural bacterial epizootics. *North Am. J. Aquac.* 70, 98–106. doi:10.1577/a07-006.1

Eachus, H., Zaucker, A., Oakes, J. A., Griffin, A., Weger, M., Güran, T., et al. (2017). Genetic disruption of 21-hydroxylase in zebrafish causes interrenal hyperplasia. *Endocrinology* 158, 4165–4173. doi:10.1210/en.2017-00549

Fang, Y., Shi, B., Liu, X., Luo, J., Rao, Z., Liu, R., et al. (2020). Xiaoyao pills attenuate inflammation and nerve injury induced by lipopolysaccharide in hippocampal neurons *in vitro*. *Neural Plast.* 2020, 8841332. doi:10.1155/2020/8841332

Faught, E., and Vijayan, M. M. (2016). Mechanisms of cortisol action in fish hepatocytes. *Comp. Biochem. Physiology Part B Biochem. Mol. Biol.* 199, 136–145. doi:10.1016/j.cbpb.2016.06.012

Fuentes, E. N., Valdés, J. A., Molina, A., and Björnsson, B. T. (2013). Regulation of skeletal muscle growth in fish by the growth hormone–insulin-like growth factor system. *General Comp. Endocrinol.* 192, 136–148. doi:10.1016/j.ygcen.2013.06.009

Gajda, A. M., and Storch, J. (2015). Enterocyte fatty acid-binding proteins (FABPs): different functions of liver and intestinal FABPs in the intestine. *Prostagl. Leukot. Essent. Fat. Acids* 93, 9–16. doi:10.1016/j.plefa.2014.10.001

Gao, Y., Dai, Z., Shi, C., Zhai, G., Jin, X., He, J., et al. (2016). Depletion of myostatin b promotes somatic growth and lipid metabolism in zebrafish. *Front. Endocrinol.* 7, 88. doi:10.3389/fendo.2016.00088

Graham, K. (1999). "A review of the biology and management of blue catfish," in Catfish 2000: proceedings of the international ictalurid symposium American Fisheries Society, Symposium, 37–49.

Hasty, P., Bradley, A., Morris, J. H., Edmondson, D. G., Venuti, J. M., Olson, E. N., et al. (1993). Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nature* 364, 501–506. doi:10.1038/364501a0

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi:10.1038/nature12111

Hsieh, W.-P., Chu, T.-M., Wolfinger, R. D., and Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165, 747–757. doi:10.1093/genetics/165.2.747

Kaitetzidou, E., Chatzifotis, S., Antonopoulou, E., and Sarropoulou, E. (2015). Identification, phylogeny, and function of fabp2 paralogs in two non-model teleost fish species. *Mar. Biotechnol.* 17, 663–677. doi:10.1007/s10126-015-9648-6

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Khalil, K., Elayat, M., Khalifa, E., Daghash, S., Elaswad, A., Miller, M., et al. (2017). Generation of myostatin gene-edited channel catfish (*Ictalurus punctatus*) via zygote injection of CRISPR/Cas9 system. *Sci. Rep.* 7, 7301. doi:10.1038/s41598-017-07223-7

Kim, K.-H., Antkiewicz, D. S., Yan, L., Eliceiri, K. W., Heideman, W., Peterson, R. E., et al. (2007). Lrrc10 is required for early heart development and function in zebrafish. *Dev. Biol.* 308, 494–506. doi:10.1016/j.ydbio.2007.06.005

Kregel, K. C. (2002). Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. *J. Appl. physiology* 92, 2177–2186. doi:10.1152/japplphysiol.01267.2001

Lange, M. D., Farmer, B. D., and Abernathy, J. (2018). Catfish mucus alters the Flavobacterium columnare transcriptome. *FEMS Microbiol. Lett.* 365, fny244. doi:10.1093/femsle/fny244

Li, P., Peatman, E., Wang, S., Feng, J., He, C., Baoprasertkul, P., et al. (2007). Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs. *BMC genomics* 8, 177–179. doi:10.1186/1471-2164-8-177

Liu, S., Zhou, Z., Lu, J., Sun, F., Wang, S., Liu, H., et al. (2011). Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC genomics* 12, 53–13. doi:10.1186/1471-2164-12-53

Manuylov, N. L., Manuylova, E., Avdoshina, V., and Tevosian, S. (2008). Serdin1/Lrrc10 is dispensable for mouse development. *genesis* 46, 441–446. doi:10.1002/dvg.20422

Miller, W. L., and Auchus, R. J. (2011). The molecular biology, biochemistry, and physiology of human steroidogenesis and its disorders. *Endocr. Rev.* 32, 81–151. doi:10.1210/er.2010-0013

Nishimura, H. (2004). Phylogeny and ontogeny of the renin-angiotensin system. *Angiotensin Vol. I* 31–70. doi:10.1016/j.cbpa.2020.110879

Ölmez, A., Bayir, M., Wang, C., and Bayir, A. (2015). Effects of long-term starvation and refeeding on fatty acid metabolism-relatedgene expressions in the liver of zebrafish, *Danio rerio*. *Turkish J. Veterinary & Animal Sci.* 39, 654–660. doi:10.3906/vet-1507-54

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Saetan, W., Ye, M., Lin, X., Lin, X., Zhang, Y., Huang, Y., et al. (2021). Comparative transcriptome analysis of heart tissue in response to hypoxia in Silver Sillago (Sillago sihama). *J. Ocean Univ. China* 20, 949–958. doi:10.1007/s11802-021-4692-5

Sharma, M. K., Denovan-Wright, E. M., Degrave, A., Thisse, C., Thisse, B., and Wright, J. M. (2004). Sequence, linkage mapping and early developmental expression of the intestinal-type fatty acid-binding protein gene (fabp2) from zebrafish (*Danio rerio*). *Comp. Biochem. Physiology Part B Biochem. Mol. Biol.* 138, 391–398. doi:10.1016/j.cbpc.2004.05.009

Shih, Y.-H., Zhang, Y., Ding, Y., Ross, C. A., Li, H., Olson, T. M., et al. (2015). Cardiac transcriptome and dilated cardiomyopathy genes in zebrafish. *Circ. Cardiovasc. Genet.* 8, 261–269. doi:10.1161/CIRCGENETICS.114.000702

Singh, K. P., Miaskowski, C., Dhruva, A. A., Flowers, E., and Kober, K. M. (2018). Mechanisms and measurement of changes in gene expression. *Biol. Res. Nurs.* 20, 369–382. doi:10.1177/1099800418772161

Sonawane, A. R., Platig, J., Fagny, M., Chen, C. Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi:10.1016/j.celrep.2017.10.001

Tan, S., Wang, W., Tian, C., Niu, D., Zhou, T., Jin, Y., et al. (2019). Heat stress induced alternative splicing in catfish as determined by transcriptome analysis. *Comp. Biochem. Physiology Part D Genomics Proteomics* 29, 166–172. doi:10.1016/j.cbd.2018.11.008

Torrans, L., and Ott, B. (2018). Effect of grading fingerling hybrid catfish (♀ channel catfish × ♂ blue catfish) on growth, production, feed conversion, and Food fish size distribution. *North Am. J. Aquac.* 80, 187–192. doi:10.1002/naaq.10024

Tsan, M.-F., and Gao, B. (2009). Heat shock proteins and immune system. *J. Leucocyte Biol.* 85, 905–910. doi:10.1189/jlb.0109005

Venkatachalam, A. B., Fontenot, Q., Farrara, A., and Wright, J. M. (2018). Fatty acid-binding protein genes of the ancient, air-breathing, ray-finned fish, spotted gar (*Lepisosteus oculatus*). *Comp. Biochem. Physiology Part D Genomics Proteomics* 25, 19–25. doi:10.1016/j.cbd.2017.10.002

Venkatachalam, A. B., Parmar, M. B., and Wright, J. M. (2017). Evolution of the duplicated intracellular lipid-binding protein genes of teleost fishes. *Mol. Genet. Genomics* 292, 699–727. doi:10.1007/s00438-017-1313-5

Venold, F. F., Penn, M. H., Thorsen, J., Gu, J., Kortner, T. M., Krogdahl, Å., et al. (2013). Intestinal fatty acid binding protein (fabp2) in Atlantic salmon (*Salmo salar*): localization and alteration of expression during development of diet induced enteritis. *Comp. Biochem. Physiology Part A Mol. Integr. Physiology* 164, 229–240. doi:10.1016/j.cbpa.2012.09.009

Wang, H., Bruce, T. J., Su, B., Li, S., Dunham, R. A., and Wang, X. (2022a). Environment-dependent heterosis and transgressive gene expression in reciprocal hybrids between the channel catfish *Ictalurus punctatus* and the blue catfish *Ictalurus furcatus*. *Biol. (Basel)* 11, 117. doi:10.3390/biology11010117

Wang, H., Bruce, T. J., Su, B., Li, S., Dunham, R. A., and Wang, X. (2022c). Environment-dependent heterosis and transgressive gene expression in reciprocal hybrids between the channel catfish *Ictalurus punctatus* and the blue catfish *Ictalurus furcatus*. *Biology* 11, 117. doi:10.3390/biology11010117

Wang, H., Su, B., Butts, I. A. E., Dunham, R. A., and Wang, X. (2022b). Chromosome-level assembly and annotation of the blue catfish *Ictalurus furcatus*, an aquaculture species for hybrid catfish reproduction, epigenetics, and heterosis studies. *Gigascience* 11, giac070. doi:10.1093/gigascience/giac070

Wolters, W. R., and Johnson, M. R. (1994). Enteric septicemia resistance in blue catfish and three channel catfish strains. *J. Aquatic Animal Health* 6, 329–334. doi:10.1577/1548-8667(1994)006<0329:esribc>2.3.co;2

Xiao, C., Wang, C., Zhang, Q., Yang, X., Huang, S., Luo, Y., et al. (2021). Transcriptomic analysis of adult zebrafish heart and brain in response to 2, 6-dichloro-1, 4-benzoquinone exposure. *Ecotoxicol. Environ. Saf.* 226, 112835. doi:10.1016/j.ecoenv.2021.112835

Xu, H., Zhang, Y., Wang, C., Wei, Y., Zheng, K., and Liang, M. (2017). Cloning and characterization of fatty acid-binding proteins (fabps) from Japanese seabass (Lateolabrax japonicus) liver, and their gene expressions in response to dietary arachidonic acid (ARA). *Comp. Biochem. Physiology Part B Biochem. Mol. Biol.* 204, 27–34. doi:10.1016/j.cbpb.2016.11.006

Xu, Z., Parra, D., Gómez, D., Salinas, I., Zhang, Y.-A., von Gersdorff Jørgensen, L., et al. (2013). Teleost skin, an ancient mucosal surface that elicits gut-like immune responses. *Proc. Natl. Acad. Sci.* 110, 13097–13102. doi:10.1073/pnas.1304319110

Xu, Y., Tan, Q., Kong, F., Yu, H., Zhu, Y., Yao, J., et al. (2019). Fish growth in response to different feeding regimes and the related molecular mechanism on the changes in skeletal muscle growth in grass carp (*Ctenopharyngodon idellus*). *Aquaculture* 512, 734295. doi:10.1016/j.aquaculture.2019.734295

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659. doi:10.1093/bioinformatics/bti042

Yang, Y., Fu, Q., Liu, Y., Wang, X., Dunham, R., Liu, S., et al. (2018). Comparative transcriptome analysis reveals conserved branching morphogenesis related genes involved in chamber formation of catfish swimbladder. *Physiol. Genomics* 50, 67–76. doi:10.1152/physiolgenomics.00089.2017

Yeh, H. Y., Shoemaker, C. A., and Klesius, P. H. (2005). Evaluation of a loop-mediated isothermal amplification method for rapid detection of channel catfish *Ictalurus punctatus* important bacterial pathogen Edwardsiella ictaluri. *J. Microbiol. Methods* 63, 36–44. doi:10.1016/j.mimet.2005.02.015

Zhang, L., Yu, Y., Dong, L., Gan, J., Mao, T., Liu, T., et al. (2021). Effects of moderate exercise on hepatic amino acid and fatty acid composition, liver transcriptome, and intestinal microbiota in channel catfish (*Ictalurus punctatus*). *Comp. Biochem. Physiology Part D Genomics Proteomics* 40, 100921. doi:10.1016/j.cbd.2021.100921

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 May 2024
DOI 10.3389/fgene.2024.1346903



OPEN ACCESS

EDITED BY
Nuno Carolino,
National Institute for Agricultural and Veterinary Research (INIAV), Portugal

REVIEWED BY
Karine Assis Costa,
Universidade Estadual Paulista—UNESP, Brazil
Lucas Lima Verardo,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil
Larissa Graciano Braga,
São Paulo State University São Paulo, Brazil in collaboration with reviewer LV

*CORRESPONDENCE
Houqiang Xu,
gzdxxhq@163.com

RECEIVED 30 November 2023
ACCEPTED 15 April 2024
PUBLISHED 02 May 2024

CITATION
Xiao M, Ruan Y, Huang J, Dai L, Xu J and Xu H (2024), Association analysis between Acetyl-Coenzyme A Acyltransferase-1 gene polymorphism and growth traits in Xiangsu pigs. *Front. Genet.* 15:1346903. doi: 10.3389/fgene.2024.1346903



# Association analysis between Acetyl-Coenzyme A Acyltransferase-1 gene polymorphism and growth traits in Xiangsu pigs

Meimei Xiao[1,2,3], Yong Ruan[1,2,3], Jiajin Huang[1,2,3], Lingang Dai[1,2,3], Jiali Xu[1,2,3] and Houqiang Xu[1,2,3]*

[1]Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University, Guiyang, China, [2]Guizhou Provincial Key Laboratory of Animal Genetics, Breeding and Reproduction, Guizhou University, Guiyang, China, [3]College of Animal Science, Guizhou University, Guiyang, China

**Introduction:** Acetyl-Coenzyme A Acyltransferase-1 (*ACAA1*) is a peroxisomal acyltransferase involved in fatty acid metabolism. Current evidence does not precisely reveal the effect of the *ACAA1* gene on pig growth performance.

**Methods:** The present study assessed the mRNA expression levels of the *ACAA1* gene in the heart, liver, spleen, lung, kidney of 6-month-old Xiangsu pigs and in the longissimus dorsi muscle at different growth stages (newborn, 6 months and 12 months of age) using RT-qPCR. The relationship between single-nucleotide polymorphisms (SNPs) of *ACAA1* gene and growth traits in 6-month-old and 12-month-old Xiangsu pigs was investigated on 184 healthy Xiangsu pigs using Sanger sequencing.

**Results:** The *ACAA1* gene was expressed in heart, liver, spleen, lung, kidney, and longissimus dorsi muscle of 6-month-old pigs, with the highest level of expression in the liver. *ACAA1* gene expression in the longissimus dorsi muscle decreased with age ($p < 0.01$). In addition, four SNPs were identified in the *ACAA1* gene, including exon g.48810 A>G (rs343060194), intron g.51546 T>C (rs319197012), exon g.55035 T>C (rs333279910), and exon g.55088 C>T (rs322138947). Hardy-Weinberg equilibrium ($p > 0.05$) was found for the four SNPs, and linkage disequilibrium (LD) analysis revealed a strong LD between g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) ($r^2 = 1.000$). Association analysis showed that g.48810 A>G (rs343060194), g.51546 T>C (rs319197012), g.55035 T>C (rs333279910), and g.55088 C>T (rs322138947) varied in body weight, body length, body height, abdominal circumference, leg and hip circumference and living backfat thickness between 6-month-old and 12-month-old Xiangsu pigs.

**Conclusion:** These findings strongly demonstrate that the *ACAA1* gene can be exploited for marker-assisted selection to improve growth-related phenotypes in Xiangsu pigs and present new candidate genes for molecular pig breeding.

KEYWORDS

**Xiangsu pigs, *ACAA1*, single-nucleotide polymorphism, fat deposition, growth traits**

# 1 Introduction

Pork is one of the important sources of animal protein for humans. Improving the growth traits of pigs is an ongoing goal in the field of animal husbandry. Growth traits such as living backfat thickness (LBT), body length (BL), body height (BH), chest circumference (CC), chest depth (CD), and rump circumference (RC) are directly related to the economic efficiency of pigs (Liu et al., 2021; Zhang et al., 2021).Growth traits are quantitative traits that are regulated by a few major genes and a large number of minor genes (Boyle et al., 2017). With the rapid development of molecular breeding and sequencing technologies, many genes that regulate pig growth traits have been identified and confirmed (Shi et al., 2022).

Acetyl-Coenzyme A Acyltransferase-1 (*ACAA1*) cleaves 3-ketoacyl-CoA to acetyl-CoA and acyl-CoA by catalyzing the β-oxidation of fatty acids in peroxisomes, driving the synthesis and secretion of fatty acids (Wanders et al., 2001; Wang et al., 2021). This enzyme is also key in regulating fatty acid oxidation and lipid metabolism (Luo et al., 2018). The *ACAA1* gene is downstream in the peroxisome proliferator–activated receptor (PPAR) signaling pathway. The *PPAR* enzyme critically regulates fatty acid synthesis and transport, catalyzes the synthesis of esterified cholesterol from free cholesterol and long-chain fatty acids, and plays a crucial role in fatty acid metabolism (Li et al., 2017). Recent research on the *ACAA1* gene has primarily focused on human cancer and metabolic diseases. Emerging evidence indicates that *ACAA1* gene expression is downregulated in hepatocellular carcinoma and renal clear cell carcinoma (Li et al., 2017; Liu et al., 2015; Nwosu et al., 2018; Yan et al., 2017; Zhang et al., 2019). The *ACAA1* gene was revealed to be highly expressed in triple-negative breast cancer cells, and inhibiting the *ACAA1* gene decreased the proliferation of triple-negative breast cancer cells (Peng et al., 2023). *ACAA1* is a type 2 diabetes (T2D) biomarker that can predict the metabolic characteristics of pre-diabetes in mouse models (Kumar et al., 2015). Research on the *ACAA1* gene in animal husbandry has linked ACAA1 mutation to milk production traits of buffalo. Analysis of the liver transcriptomes and the microarray dataset of Hereford (beef breed) and Holstein-Friesian (dairy breed) bulls with different genetic backgrounds revealed that Hereford bulls were highly involved in fatty acid biosynthesis and lipid metabolism by up-regulating *ACAA1* gene expression as compared to Holstein-Friesian (Lisowski et al., 2014).

Single Nucleotide Polymorphism (SNP) refers to the DNA sequence polymorphism caused by single nucleotide variation at the chromosome genomic level, and the frequency of this variation is more than 1% in at least one population (Taylor et al., 2001; Vignal et al., 2002). Five SNPs (g.-681 A>T, g.-24348 G>T, g.-806 C>T, g.-1868 C>T and g.-23117 C>T) were identified in the buffalo *ACAA1* gene, among which g.-681 A>T, g.-24348 G>T, and g.-23117 C>T are significantly associated with milk production traits in buffaloes. In addition, the g.-681 A>T mutation in the promoter region significantly changed the transcriptional activity (Deng et al., 2023). A missense variant rs117916664 of the *ACAA1* gene was identified in a Han Chinese early-onset familial Alzheimer's disease (AD) family and found to be associated with early-onset familial AD (Luo et al., 2021). A genetic polymorphism in the *ACAA1* gene alters the association between endotoxin exposure and asthma (Sordillo et al., 2011). However, data on the polymorphism of the *ACAA1* gene in pigs is scarce.

Xiangsu pig is a novel breeding strain that utilizes Sutai pig and Congjiang Xiang pig as parents and repeatedly backcrossed with Congjiang Xiang pig as male parent. Congjiang Xiang pig has early sexual maturity, strong fat deposition capacity, and strong disease resistance but slow growth (Liu et al., 2018; Tang et al., 2018; Xu et al., 2022). The SuTai pig is a breed characterized by its high reproductive rate and strong adaptability (Bao et al., 2012). The Congjiang xiang pig accounts for 87.5% of the genetic lineage within the Xiangsu pig population, allowing for the full inheritance of its genetic traits in subsequent generations (Xu et al., 2022). Therefore, we selected *ACAA1* gene as a candidate gene for the growth traits of the Xiangsu pig and evaluated the relationship between *ACAA1* gene polymorphism and the growth traits of the Xiangsu pig, which is valuable for Xiangsu pig breeding in the future.

# 2 Materials and methods

## 2.1 Experimental animals

The animal experiments fully adhered to the guidelines of the Animal Welfare Committee of Guizhou University (EAE-GZU-2022-E031). The production cycle (farrowing to growing-finishing) of Xiangsu pig is 12 months. A total of 184 healthy Xiangsu pigs under the same feeding level were selected to track and record the growth traits (body weight, body length, body height, chest circumference, abdominal circumference, tube circumference, leg and hip circumference and living backfat thickness) of 6-month-old and 12-month-old Xiangsu pigs. The measurement method of body weight, body length, body height, chest circumference, abdominal circumference, tube circumference, leg and hip circumference was referred to as NY/T2894-2016. The probe of the handheld veterinary ultrasound diagnostic device (KX5200) had been positioned vertically on the 10th and 11th thoracic vertebrae of pigs to measure the living backfat thickness of 6-month-old and 12-month-old pigs.

## 2.2 Primer design

The upstream and downstream primers were designed by Primer Premier 5.0 software using the pig *ACAA1* gene (accession number: NC_010455.5) and mRNA sequence (accession number: XM_003132103.4); *GADPH* gene (accession number: NC_010447.5) and mRNA sequence (accession number: NM_001206359.1) available in NCBI GeneBank. The primers were synthesized by Beijing Qingke Biotechnology Co., Ltd., (Table 1).

## 2.3 Collection of blood and tissue samples

The blood (5 mL) of 184 3-month-old Xiangsu pigs was drawn through the jugular vein using an EDTA anticoagulant tube, labeled with the number and date, and stored in a refrigerator at −20°C for DNA extraction. Three 6-month-old Xiangsu pigs were randomly selected from Xiangsu pigs for slaughter. The heart, liver, spleen,

TABLE 1 Primer information of *ACAA1* gene sequence.

| Primer names | Primer sequences (5′→3′) | Product size/bp | Annealing temperature/°C |
|---|---|---|---|
| ACAA1-Exon7 | F:GACTCTTCAGAGGAAGAGAGAGGAG | 649 | 57 |
| | R:CAGCAGACGATGACTCTGCTGAT | | |
| ACAA1-Exon9 | F:TTGTTAGATGTGTCCTTCACTGTGG | 675 | 63 |
| | R:TCAACTTTCTAGGCCTCCAGAGTT | | |
| ACAA1-Exon15 | F:TGAGGTCTGGCATCTTCTGTGC | 615 | 63 |
| | R:CTCAGAGGTGGAGCAGTACAAAGAG | | |
| ACAA1-qPCR | F:ATGGGGATAACCTCAGAGAACGT | 175 | 55 |
| | R:TCTCATTGCCCTTGTCATCGTAG | | |
| GADPH | F:GGTCGGAGTGAACGGATTT | 247 | 60 |
| | R:CCATTTGATGTTGGCGGGA | | |

Note: F denotes the upstream primer, and R denotes the downstream primer. bp: base pair.

lung, kidney and longissimus dorsi muscle were collected in 2 mL cryopreservation tubes and stored in a refrigerator at −80°C for RNA extraction in order to compare the expression of *ACAA1* gene in different tissues of 6-month-old pigs. Three pigs are selected for slaughter from each age group (newborn and 12 months old) at each stage, and the longissimus dorsi muscle was collected and preserved in 2 mL cryopreservation tubes in a refrigerator at −80°C for RNA extraction in order to compare the expression of *ACAA1* at different ages.

## 2.4 DNA and RNA extraction

DNA was extracted from 184 blood samples of Xiangsu pigs using the whole blood DNA extraction kit (D3392-01, Omega). Total RNA of heart, liver, spleen, lung and kidney of Xiangsu pigs at 6 months of age and total RNA of longissimus dorsi muscle at newborn, 6 months and 12 months of age was extracted using the TRIzol Extraction Kit (15,596,026; Thermo Fisher). The concentration and purity of DNA and RNA were determined using an ultra-micro spectrophotometer (Thermo Mano Drop 2000), followed DNA by storage at −20°C and RNA by storage at −80°C (Supplementary Table S1).

## 2.5 cDNA synthesis

cDNA was synthesized using the RNA Reverse Transcription Kit (A234-10; GenStar). By this kit 1 μg RNA, 1 μL Primer Mix, 10 μL 2× StarScript III Buffer, 1 μL StarScript III Enzyme Mix, and then supplemented with Nuclease-free Water to a final volume of 20 μL. The mixture was incubated at 50°C for 15 min, followed at 85°C for 5 min. The resulting cDNA was stored at −20°C for subsequent experiments.

## 2.6 Amplification

A 30 μL system was used for PCR amplification. The PCR reaction mixture was prepared as follows: 15 μL 2×Taq PCR Starmix, 10.5 μL $ddH_2O$, 1.5 μL DNA template (40 ng/μL), forward and reverse primers, 1.5 μL each. The PCR amplification procedure included a pre-denaturation at 94°C for 3 min; after 35 cycles, 94°C denaturation for 30 s (Tm see Table 1), annealing for 30 s, 72°C extension for 1 min; 72°C final extension for 5 min; infinite hold at 4°C. The PCR products (184) were visualized with 1% agarose gel electrophoresis and sent to Qingke Biological Co., Ltd. for sequencing.

## 2.7 Real-time fluorescent quantitative PCR

The total RNA concentration was standardized (1000 ng/μL), and 1 μg total RNA, 2 μL 5 × gDNA Eraser Buffer gDNA Eraser, and RNase-Free water were added to the enzyme-free PCR tube to obtain a 10 μL reaction volume. The samples were incubated at 37°C for 5 min to remove gDNA. In addition, a 10 μL Master Mix (including 1 μL PrimeScript RT Enzyme Mix I, 1 μL RT Primer Mix, 4 μL 5× PrimeScript Buffer 2, and 4 μL RNase-Free $ddH_2O$) was prepared on ice. Total RNA (without gDNA) and Master Mix were mixed in an enzyme-free PCR tube, incubated at 42°C for 15 min, then at 85°C for 5 min, and the resultant cDNA was stored at −20°C.

A 10 μL real-time fluorescence quantitative PCR reaction mixture was constituted as follows: 5 μL 2× PowerUp SYBP Green Master Mix (A25742; Thermo Fisher), 0.5 μL cDNA template, 0.4 μL (10 μmol/L) forward and reverse primers, and 3.7 μL $ddH_2O$. The quantitative real-time PCR amplification steps were set as follows: 50°C UDG enzyme activation 2 min; pre-denaturation at 95°C for 2 min; 95°C denaturation 15 s, 55°C annealing 30 s, 72°C extension 30 s, 40 cycles; from 72°C to 95°C, a temperature increase step by 1.6°C per second for 15 s, and then a temperature decrease step by 1.6°C per second from 95°C to 60°C. Each sample had 3 replicates.

## 2.8 Statistical analysis

The presence of SNPs in *ACAA1* sequence was determined via peak plotting against the PCR sequencing reads using the

SeqMan software. The genotype, allele frequency, and Hardy-Weinberg equilibrium (HWE) of each mutation site were computed directly, and whether the genotype conformed to HWE was analyzed using the χ2 test and *p*-value. Nei's method was employed to analyze the genetic indexes of the population, including gene heterozygosity (*He*), gene homozygosity (*Ho*), and polymorphism information content (*PIC*) (Nei and Roychoudhury, 1974). The effective number of alleles (*Ae*) is related to the distribution of gene frequency in the population, and the markers were calculated with GenAlEx 6.5 (New Brunswick, NJ, United States) (Peakall and Smouse, 2012). *PIC* refers to a measure used to assess the ability to detect polymorphism among individuals in a population. The range of polymorphic information content is between 0 and 1, with higher values indicating greater information content and polymorphism in genetic markers (Serrote et al., 2020). The SHEsis platform (http://analysis.bio-x.cn) was used for linkage disequilibrium (LD) analysis and haplotype analysis of single-nucleotide polymorphisms (SNPs) in the *ACAA1* gene (Li et al., 2009; Shi and He, 2005). Squared allele-frequency correlations ($r^2$) and Standardized disequilibrium coefficients ($D'$) were used to estimate the level of LD (Du et al., 2007). The $r^2$ value is commonly used to evaluate the degree of linkage disequilibrium. When $r^2 > 0.33$, it is a strong linkage disequilibrium state (Ardlie et al., 2002). $D'$ is the normalized coefficient of linkage disequilibrium (LD) divided by the theoretical maximum difference between the observed and expected allele frequencies (Bozorgmehr et al., 2020). Haplotype and diplotypes analyses were performed based on SNPs.

*ACAA1* genotype association analysis was performed using IBM SPSS 22.0 (IBM, New York, NY, United States). The least square method was applied to the general linear model (GLM) to examine the association between genotypes and growth traits of 184 Xiangsu pigs. A statistical model, $Y_{ij} = \mu + G_i + Sj + e_{ij}$, was developed where $Y_{ij}$ denotes the observed growth trait; $\mu$ denotes the overall population means; $G_i$ represents the fixed effect of the genotype, *Sj* represents the random effect of sire, and *eij* denotes the random error (Naicy et al., 2017).

The relative expression of the *ACAA1* gene was calculated using the $2^{-\Delta\Delta Ct}$ method (Saitou and Nei, 1987), with glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene as the endogenous reference gene, where Δ Ct = Ct (target gene)-Ct (*GADPH*), and $2^{-\Delta\Delta Ct}$ represents the differetial expression multiple relative to *GADPH* expression A one-way analysis of variance (ANOVA) was conducted to compare the differences among the heart, liver, spleen, lung, kidney, and longissimus dorsi muscle of 6-month-old pigs. Additionally, ANOVA was performed to assess the differences in the longissimus dorsi muscle of newborns, 6-month-old, and 12-month-old pigs.

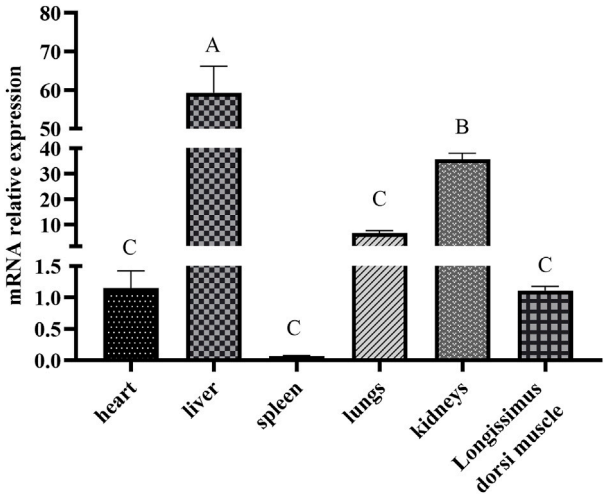



FIGURE 1
Analysis of the differential expression of the *ACAA1* gene in various tissues at 6 months of age. Note: Relative mRNA expression levels were calculated by $2^{-\Delta\Delta Ct}$ method. "A, B, C" indicate extremely significant differences among different tissues of six-month-old Xiangsu pigs ($p < 0.01$); the same uppercase letters indicate no significant difference.

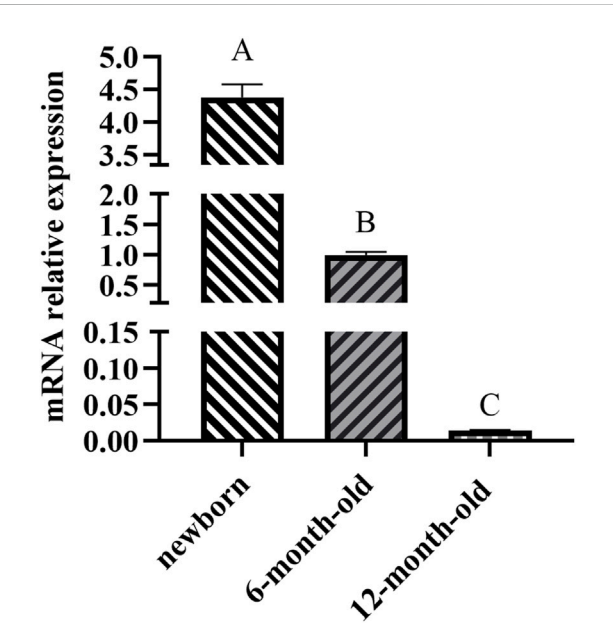



FIGURE 2
Analysis of the differential expression of the *ACAA1* gene in dorsal longest muscle at newborn, 6 months, and 12 months of age. Note: Relative mRNA expression levels were calculated by $2^{-\Delta\Delta Ct}$ method. "A, B, C" denotes extremely significant differences in dorsal longest muscle at birth, 6 months and 12 months of age ($p < 0.01$).

# 3 Results

## 3.1 Expression level of *ACAA1* gene in Xiangsu pig tissues

The *ACAA1* gene was expressed in the heart, liver, spleen, lung, kidney, and longissimus dorsi muscle of 6-month-old Xiangsu pigs, with higher levels in the liver and kidney and lower levels in the spleen and longissimus dorsi muscle ($p < 0.01$) (Figure 1). Newborn piglets exhibited the highest expression of the *ACAA1* gene ($p < 0.01$) in the longissimus dorsi muscle, and it decreased with age (Figure 2).

10.3389/fgene.2024.1346903

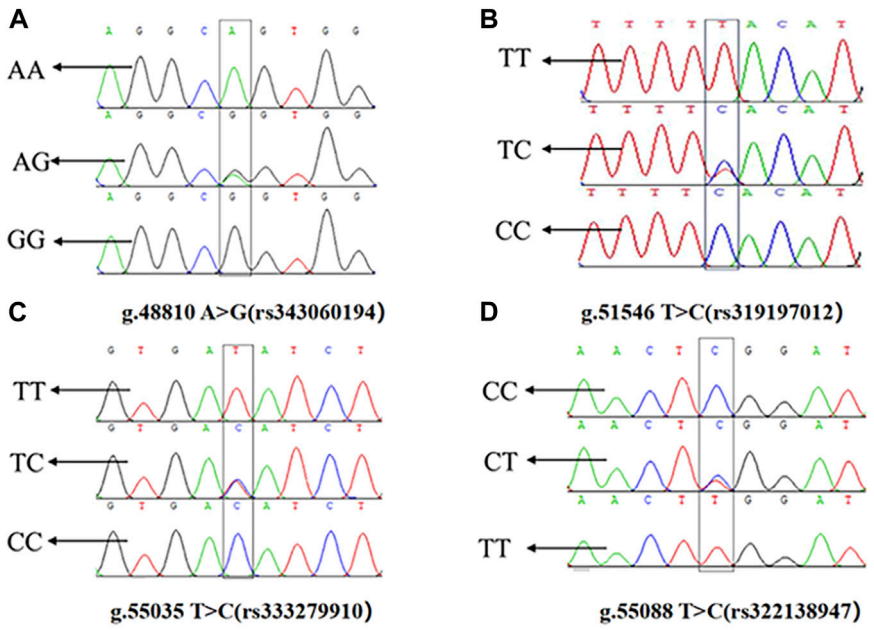



FIGURE 3
Sequencing peaks of four SNPs in the *ACAA1* gene. **(A)** g.48810 A>G (rs343060194), **(B)** g.51546 T>C (rs319197012), **(C)** g.55035 T>C (rs333279910), **(D)** g.55088 C>T (rs322138947).

## 3.2 Analysis of the *ACAA1* gene polymorphism

The presence of SNPs in *ACAA1* gene sequence was determined via peak plotting against the PCR sequencing reads using the SeqMan. Four SNPs were detected in the *ACAA1* gene of Xiangsu pigs, including exon 7 g.48810 A>G (rs343060194), intron 9 g.51546 T>C (rs319197012), exon 15 g.55035 T>C (rs333279910), and exon 15 g.55088 C>T (rs322138947) (Figure 3). Using DNA Star software to compare the sequences of three exon mutation sites g.48810 A>G (rs343060194), g.55035 T>C (rs333279910), g.55088 C>T (rs322138947) with the NCBI amino acid reference sequence of the ACAA1 (XP_003132151.1). Three exon mutation sites indicated no changes in the amino acid sequence, therefore they are synonymous mutations.

## 3.3 Genetic polymorphism analysis of the *ACAA1* gene

Each mutation site had three genotypes. The chi-square test ($\chi 2$) revealed that the four SNPs loci g.48810 A>G (rs343060194), g.51546 T>C (rs319197012), g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) were in HWE ($p > 0.05$) (Table 2).

The homozygosity (*Ho*) of the four SNPs of the *ACAA1* gene was 0.5151–0.5605, and the heterozygosity (*He*) was 0.4395–0.4849 (Table 3). The homozygosity (*Ho*) was higher than the heterozygosity (*He*), demonstrating that the four loci in this population showed a low degree of variation; effective number of alleles (*Ae*) was 1.7841–1.9413. The polymorphism information content ranged from 0.3429 to 0.3673, indicating a moderate polymorphism level ($0.25<PIC<0.5$).

## 3.4 Linkage disequilibrium and haplotype analysis of SNPs in the *ACAA1* gene

Linkage disequilibrium (LD) analysis was performed on the four SNPs in the *ACAA1* gene (Figure 4). The *D'* values of the four SNPs ranged between 0.414 and 1.000, and the $r^2$ values ranged between 0.058 and 1.000. The $r^2$ for g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) was 1.000, indicating a strong LD.

Table 4 displays the findings of the *ACAA1* gene haplotype analysis. The population had six haplotypes with a frequency greater than 5.00%, and less than 5.00% were excluded from statistical analysis. Among the six haplotypes, Hap 1 (-GCCT-) had the highest frequency (25.00%), while Hap 6 (-ATCT-) had the lowest frequency (5.70%). Based on the paired combinations of six haplotypes, five diplotypes combinations with frequencies greater than 5.00% were obtained, including Hap1/3, -GCCT/ATTC-; Hap2/2, -GTTC/GTTC-; Hap2/5, -GTTC/GCTC-; Hap2/4, -GTTC/GTCT-; Hap1/1, -GCCT/GCCT- (Table 5).

## 3.5 Association analysis between *ACAA1* gene and growth performance

The association between 4 SNP loci g.48810 A>G (rs343060194), g.51546 T>C (rs319197012), g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947)) and 8 growth traits at 6 months (Table 6) was examined using the SPSS 22 software. The results showed that in the 6-month-old pigs, there was a significant difference in body weight between the AG genotype and the GG genotype at the exon g.48810 A>G (rs343060194) locus ($p < 0.05$). The leg and hip circumference of the CC genotype at the intron g.51546 T>C (rs319197012) locus

TABLE 2 Genotype frequencies and allele frequencies of *ACAA1* gene in Xiangsu pigs.

| SNPs | rs number | Genotypic frequency | | | Allele frequency | | χ2 | *p* |
|---|---|---|---|---|---|---|---|---|
| g.48810 A>G | rs343060194 | AA | AG | GG | A | G | 3.32 | 0.19 |
| | | 0.14 (25) | 0.38 (70) | 0.48 (89) | 0.33 | 0.67 | | |
| g.51546 T>C | rs319197012 | TT | TC | CC | T | C | 1.87 | 0.39 |
| | | 0.39 (72) | 0.43 (79) | 0.18 (33) | 0.61 | 0.39 | | |
| g.55035 T>C | rs333279910 | TT | TC | CC | T | C | 1.20 | 0.55 |
| | | 0.36 (67) | 0.45 (82) | 0.19 (35) | 0.59 | 0.41 | | |
| g.55088 C>T | rs322138947 | CC | CT | TT | C | T | 1.20 | 0.55 |
| | | 0.36 (67) | 0.45 (82) | 0.19 (35) | 0.59 | 0.41 | | |

Note: $p > 0.05$ indicates that the gene frequency in the population is at Hardy-Weinberg equilibrium. The number of samples is indicated in brackets.

TABLE 3 Genetic information of *ACAA1* gene population.

| SNPs | rs number | Effective allele number (*Ae*) | Homozygosity (*Ho*) | Heterozygosity (*He*) | Polymorphism information content (*PIC*) |
|---|---|---|---|---|---|
| g.48810 A>G | rs343060194 | 1.7841 | 0.5605 | 0.4395 | 0.3429 |
| g.51546 T>C | rs319197012 | 1.8918 | 0.5225 | 0.4775 | 0.3635 |
| g.55035 T>C | rs333279910 | 1.9413 | 0.5151 | 0.4849 | 0.3673 |
| g.55088 C>T | rs322138947 | 1.9413 | 0.5151 | 0.4849 | 0.3673 |

Note: *PIC* < 0.25 is low polymorphism, 0.25<*PIC* < 0.5 is medium polymorphism, and *PIC* > 0.5 is high polymorphism.

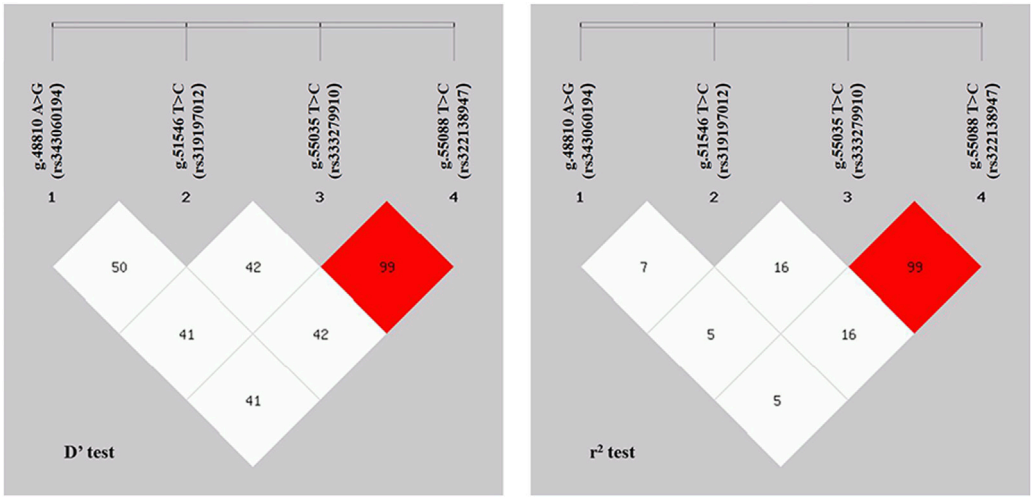



FIGURE 4
$r^2$ and $D'$ values in linkage disequilibrium analysis of *ACAA1* gene SNPs.

were significantly different from that of the TC genotype ($p < 0.05$). Furthermore, the body weight of the TC genotype at the exon g.55035 T>C (rs333279910) locus was significantly different from that of the TT genotype ($p < 0.05$), while the living backfat thickness of the TT genotype was significantly different from that of the CC genotype ($p < 0.01$). In addition, at the g.55088 C>T (rs322138947) locus, the body weight of the CT genotype was significantly different from that of the CC genotype ($p < 0.05$), and the living backfat thickness of the CC genotype was significantly different from that of the TT genotype ($p < 0.01$).

The association between 4 SNP loci g.48810 A>G (rs343060194), g.51546 T>C (rs319197012), g.55035 T>C (rs333279910), and g.55088 C>T (rs322138947)) and 8 growth traits at 12 months (Table 7) was examined using the SPSS 22 software. The results showed that the body length of pigs with the GG genotype at the g.48810 A>G (rs343060194) locus

TABLE 4 Haplotype frequencies of 4 SNPs in Xiangsu pig.

| Haplotype | g.48810 A>G (rs343060194) | g.51546 T>C (rs319197012) | g.55035 T>C (rs333279910) | g.55088 C>T (rs322138947) | Frequency (%) |
|---|---|---|---|---|---|
| Hap1 | G | C | C | T | 25.00 |
| Hap2 | G | T | T | C | 24.20 |
| Hap3 | A | T | T | C | 21.30 |
| Hap4 | G | T | C | T | 9.40 |
| Hap5 | G | C | T | C | 8.80 |
| Hap6 | A | T | C | T | 5.70 |

Note: Haplotypes with frequencies <5.00% were excluded from the analysis.

TABLE 5 Diplotypes and frequency of *ACAA1* gene.

| Diplotypes | g.48810 A>G (rs343060194) | g.51546 T>C (rs319197012) | g.55035 T>C (rs333279910) | g.55088 C>T (rs322138947) | Frequency (%) |
|---|---|---|---|---|---|
| Hap1/3 | GA | CT | CT | TC | 19.02 |
| Hap2/2 | GG | TT | TT | CC | 11.96 |
| Hap2/5 | GG | TC | TT | CC | 7.61 |
| Hap2/4 | GG | TT | TC | CT | 5.98 |
| Hap1/1 | GG | CC | CC | TT | 5.43 |

Note: Diplotypes with frequencies <5.00% were excluded from the analysis.

TABLE 6 Association analysis between *ACAA1* gene and growth traits of 6-month-old Xiangsu pigs.

| SNPS | Genotypes | B W (kg) | B L (cm) | B H (cm) | C C (cm) | A C (cm) | T C (cm) | L H C (cm) | L B T (mm) |
|---|---|---|---|---|---|---|---|---|---|
| g.48810 A>G (rs343060194) | AA | 70.56 ± 1.76$^{ab}$ | 95.56 ± 3.11 | 63.76 ± 3.02 | 95.72 ± 3.73 | 96.88 ± 3.28 | 17.96 ± 0.79 | 62.88 ± 2.01 | 10.88 ± 0.89 |
| | AG | 70.97 ± 1.83$^{a}$ | 94.97 ± 4.49 | 64.84 ± 4.63 | 95.47 ± 2.86 | 97.17 ± 2.86 | 18.00 ± 0.92 | 62.20 ± 2.04 | 10.88 ± 1.01 |
| | GG | 70.29 ± 2.19$^{b}$ | 94.91 ± 3.16 | 63.79 ± 3.28 | 94.93 ± 3.32 | 97.06 ± 3.71 | 17.94 ± 0.94 | 62.20 ± 1.83 | 10.60 ± 1.51 |
| g.51546 T>C (rs319197012) | TT | 70.28 ± 2.04 | 95.15 ± 3.22 | 63.54 ± 3.19 | 95.08 ± 3.12 | 96.88 ± 3.05 | 17.83 ± 0.95 | 62.23 ± 1.80$^{ab}$ | 10.65 ± 1.61 |
| | TC | 70.70 ± 1.97 | 94.65 ± 4.34 | 64.53 ± 4.54 | 95.13 ± 3.17 | 97.05 ± 3.59 | 17.95 ± 0.93 | 62.09 ± 2.02$^{b}$ | 10.89 ± 1.01 |
| | CC | 71.00 ± 2.03 | 95.63 ± 2.97 | 64.76 ± 3.10 | 95.88 ± 3.52 | 97.58 ± 3.35 | 18.15 ± 0.91 | 62.91 ± 1.97$^{a}$ | 10.59 ± 0.93 |
| g.55035 T>C (rs333279910) | TT | 70.13 ± 1.99$^{b}$ | 94.72 ± 2.74 | 63.66 ± 3.05 | 94.90 ± 3.08 | 96.45 ± 3.22 | 17.78 ± 0.95 | 62.03 ± 1.65 | 10.99 ± 0.97$^{A}$ |
| | TC | 70.93 ± 1.98$^{a}$ | 95.04 ± 4.52 | 64.52 ± 4.39 | 95.27 ± 3.10 | 97.41 ± 3.47 | 18.05 ± 0.94 | 62.30 ± 2.12 | 10.75 ± 0.98$^{AB}$ |
| | CC | 70.66 ± 2.02$^{ab}$ | 95.57 ± 3.19 | 64.40 ± 3.78 | 95.86 ± 3.70 | 97.49 ± 3.15 | 18.00 ± 0.87 | 62.77 ± 1.97 | 10.26 ± 2.01$^{B}$ |
| g.55088 C>T (rs322138947) | CC | 70.13 ± 1.99$^{b}$ | 94.72 ± 2.74 | 63.66 ± 3.05 | 94.90 ± 3.08 | 96.45 ± 3.22 | 17.78 ± 0.95 | 62.03 ± 1.65 | 10.99 ± 0.97$^{A}$ |
| | CT | 70.93 ± 1.98$^{a}$ | 95.03 ± 4.52 | 64.52 ± 4.39 | 95.27 ± 3.10 | 97.41 ± 3.47 | 18.05 ± 0.94 | 62.30 ± 2.12 | 10.75 ± 0.98$^{AB}$ |
| | TT | 70.6 ± 2.04$^{ab}$ | 95.57 ± 3.19 | 64.40 ± 3.78 | 95.86 ± 3.70 | 97.49 ± 3.15 | 18.00 ± 0.87 | 62.77 ± 1.97 | 10.26 ± 2.01$^{B}$ |

Note: BW, Body weight/kg; BL, Body length/cm; BH, Body height/cm; CC, Chest circumference/cm, AC, Abdominal circumference/cm; TC, Tube circumference/cm; LHC, Leg and hip circumference/cm; LBT, Living backfat thickness/mm. The data are expressed as mean ± standard deviation. Different lowercase letters represent significant differences at 0.05 level ($p < 0.05$), different uppercase letters indicate significant differences at 0.01 level ($p < 0.01$), and the same letters (case-insensitive) show no significant difference ($p > 0.05$).

was significantly different from that of the AG genotype ($p < 0.01$). The abdominal circumference of pigs with the AG genotype was significantly different from that of the GG genotype ($p < 0.05$). The body weight and abdominal circumference of pigs with the TC genotype at the g.51546 T>C (rs319197012) intron locus were significantly different from that of the TT genotype ($p < 0.05$), and the same was observed for the leg and hip circumference and living backfat thickness with the TC genotype from that of the TT genotype ($p < 0.01$). The body weight of pigs with the TC genotype at the g.55035 T>C

TABLE 7 Association analysis between *ACAA1* gene and growth traits of 12-month-old Xiangsu pigs.

| SNPS | Genotypes | B W (kg) | B L (cm) | B H (cm) | C C (cm) | A C (cm) | T C (cm) | L H C (cm) | L B T (mm) |
|---|---|---|---|---|---|---|---|---|---|
| g.48810 A>G (rs343060194) | AA | 137.72 ± 4.77 | 130.44 ± 1.39$^{AB}$ | 72.20 ± 1.29 | 119.52 ± 1.94 | 126.88 ± 3.77$^{ab}$ | 19.64 ± 0.86 | 87.16 ± 3.08 | 13.88 ± 1.22 |
| | AG | 138.17 ± 4.10 | 130.04 ± 1.26$^{B}$ | 71.57 ± 1.48 | 119.83 ± 1.98 | 128.14 ± 3.52$^{a}$ | 19.76 ± 0.79 | 87.96 ± 2.52 | 14.18 ± 1.13 |
| | GG | 137.12 ± 4.06 | 130.56 ± 1.17$^{A}$ | 71.92 ± 1.60 | 119.91 ± 2.20 | 126.69 ± 3.43$^{b}$ | 19.69 ± 1.35 | 87.19 ± 2.73 | 14.02 ± 1.04 |
| g.51546 T>C (rs319197012) | TT | 136.76 ± 4.38$^{b}$ | 130.49 ± 1.30 | 71.72 ± 1.62 | 119.71 ± 2.11 | 126.71 ± 3.46$^{b}$ | 19.60 ± 0.76 | 86.86 ± 2.88$^{B}$ | 13.71 ± 1.22$^{B}$ |
| | TC | 138.25 ± 3.86$^{a}$ | 130.20 ± 1.19 | 71.89 ± 1.51 | 119.77 ± 1.98 | 127.87 ± 3.46$^{a}$ | 19.82 ± 1.40 | 88.02 ± 2.42$^{A}$ | 14.37 ± 0.90$^{A}$ |
| | CC | 137.88 ± 4.29$^{ab}$ | 130.39 ± 1.27 | 71.91 ± 1.35 | 120.21 ± 2.26 | 127.03 ± 3.86$^{ab}$ | 19.67 ± 0.85 | 87.52 ± 2.80$^{AB}$ | 14.08 ± 1.04$^{AB}$ |
| g.55035 T>C (rs333279910) | TT | 136.66 ± 4.32$^{b}$ | 130.43 ± 1.26 | 71.70 ± 1.56$^{b}$ | 119.94 ± 2.04 | 126.72 ± 3.60 | 19.75 ± 1.50 | 86.78 ± 2.89$^{b}$ | 13.74 ± 1.19$^{B}$ |
| | TC | 138.27 ± 3.97$^{a}$ | 130.21 ± 1.23 | 71.68 ± 1.52$^{b}$ | 119.71 ± 1.98 | 127.62 ± 3.48 | 19.73 ± 0.80 | 87.83 ± 2.50$^{a}$ | 14.30 ± 0.99$^{A}$ |
| | CC | 137.86 ± 4.20$^{ab}$ | 130.51 ± 1.27 | 72.40 ± 1.35$^{a}$ | 119.89 ± 2.39 | 127.49 ± 3.62 | 19.57 ± 0.74 | 88.00 ± 2.66$^{a}$ | 14.11 ± 1.05$^{AB}$ |
| g.55088 C>T (rs322138947) | CC | 136.66 ± 4.32$^{b}$ | 130.43 ± 1.26 | 71.70 ± 1.56$^{b}$ | 119.94 ± 2.04 | 126.72 ± 3.60 | 19.75 ± 1.50 | 86.78 ± 2.89$^{b}$ | 13.74 ± 1.19$^{B}$ |
| | CT | 138.27 ± 3.97$^{a}$ | 130.21 ± 1.23 | 71.68 ± 1.52$^{b}$ | 119.71 ± 1.98 | 127.62 ± 3.48 | 19.73 ± 0.80 | 87.83 ± 2.50$^{a}$ | 14.30 ± 0.99$^{A}$ |
| | TT | 137.86 ± 4.20$^{ab}$ | 130.51 ± 1.27 | 72.40 ± 1.35$^{a}$ | 119.89 ± 2.39 | 127.49 ± 3.62 | 19.57 ± 0.74 | 88.00 ± 2.66$^{a}$ | 14.11 ± 1.05$^{AB}$ |

Note: BW, Body weight/kg; BL, Body Length/cm; BH, Body height/cm; CC, Chest circumference/cm; AC, Abdominal circumference/cm; TC, Tube circumference/cm; LHC, Leg and hip circumference/cm; LBT, Living backfat thickness/mm. The data are expressed as mean ± standard deviation. Different lowercase letters represent significant differences at 0.05 level ($p < 0.05$), different uppercase letters indicate significant differences at 0.01 level ($p < 0.01$), and the same letters (case-insensitive) show no significant difference ($p > 0.05$).

(rs333279910) exon locus was significantly different from that of the TT genotype ($p < 0.05$), and the same was observed for the living backfat thickness with the TC genotype from that of the TT genotype ($p < 0.01$). The body height of pigs with the CC genotype was significantly different from that of the TC and TT genotypes ($p < 0.05$). The leg and hip circumference of pigs with the TC and CC genotypes were significantly different from that of the TT genotype ($p < 0.05$). The body weight of pigs with the CT genotype at the g.55088 C>T (rs322138947) locus was significantly different from that of the CC genotype ($p < 0.05$); and the same was observed for the living backfat thickness with the CT genotype from that of the CC genotype ($p < 0.01$). The body height of pigs with the TT genotype was significantly different from that of the CC and CT genotypes ($p < 0.05$); the leg and hip circumference of pigs with the CT and TT genotypes were significantly different from that of the CC genotype ($p < 0.05$); there were no significant differences in other indicators ($p > 0.05$).

## 3.6 Association analysis between *ACAA1* gene diplotypes and growth performance

Association analysis was performed between five diploid combinations and the growth traits of 6-month-old of Xiangsu pigs. Hap1/3, -GCCT/ATTC- outperformed Hap2/2, -GTTC/GTTC- in terms of body weight. Hap1/1, -GCCT/GCCT-was superior to Hap2/2, -GTTC/GTTC- in chest circumference, Hap2/5, -GTTC/GCTC- outperformed Hap2/2, -GTTC/GTTC- in terms of tube circumference, Hap1/1, -GCCT/GCCT-outperformed Hap1/3, -GCCT/ATTC- Hap2/5, -GTTC/GCTC-, Hap2/4, -GTTC/GTCT-in terms of leg and hip circumference (Table 8). The association analysis between five diplotypes and growth traits of 12-month-old Xiangsu pigs revealed that Hap1/1, -GCCT/GCCT-was superior to Hap2/2, -GTTC/GTTC- in body weight and leg and hip circumference. Hap2/4, -GTTC/GTCT-was superior Hap1/3, -GCCT/ATTC- in body length, Hap2/5, -GTTC/GCTC- outperformed Hap2/2, -GTTC/GTTC in terms of body height and tube circumference (Table 9). In a nutshell, Hap1/1, -GCCT/GCCT-can be employed as an advantageous genotype combination for subsequent breeding.

# 4 Discussion

This study investigated and analyzed the expression levels of the *ACAA1* gene in different tissues (heart, liver, spleen, lung, kidney, and longissimus dorsi muscle) of Xiangsu pigs at 6 months of age The results revealed that the *ACAA1* gene

TABLE 8 Relationship between diploid types and growth traits at 6 months of age in Xinagsu pigs.

| Diplotype | Frequency (%) | B W (kg) | B L (cm) | B H (cm) | C C (cm) | A C (cm) | T C (cm) | L H C (cm) | L B T (mm) |
|---|---|---|---|---|---|---|---|---|---|
| Hap1/3 | 19.0 | 71.20 ± 1.69$^{a}$ | 94.54 ± 5.6 | 65.34 ± 5.66 | 95.23 ± 2.49$^{ab}$ | 97.00 ± 2.73 | 18.03 ± 0.89$^{ab}$ | 62.06 ± 2.10$^{b}$ | 10.82 ± 1.08 |
| Hap2/2 | 12.0 | 69.55 ± 2.20$^{b}$ | 94.22 ± 3.2 | 63.18 ± 3.29 | 94.00 ± 2.64$^{b}$ | 95.77 ± 2.81 | 17.50 ± 0.91$^{b}$ | 62.13 ± 1.46$^{ab}$ | 10.65 ± 1.12 |
| Hap2/5 | 7.6 | 70.79 ± 1.85$^{ab}$ | 95.50 ± 2.9 | 64.71 ± 2.95 | 95.93 ± 3.08$^{ab}$ | 98.21 ± 4.04 | 18.29 ± 1.07$^{a}$ | 61.86 ± 1.17$^{b}$ | 11.27 ± 0.86 |
| Hap2/4 | 6.0 | 70.91 ± 2.12$^{ab}$ | 95.45 ± 3.7 | 64.45 ± 2.66 | 95.09 ± 3.14$^{ab}$ | 97.63 ± 4.06 | 17.91 ± 1.14$^{ab}$ | 61.91 ± 1.81$^{b}$ | 10.59 ± 0.92 |
| Hap1/1 | 5.4 | 70.80 ± 2.25$^{ab}$ | 95.50 ± 3.8 | 65.60 ± 3.17 | 96.30 ± 2.50$^{a}$ | 97.30 ± 3.13 | 17.90 ± 0.88$^{ab}$ | 63.40 ± 1.51$^{a}$ | 10.47 ± 1.07 |

Note: BW, Body weight/kg; B L, Body length/cm; BH: Body height/cm; CC, Chest circumference/cm; AC, Abdominal circumference/cm; TC, Tube circumference/cm; LHC, Leg and hip circumference/cm; LBT, Living backfat thickness/mm. The data are expressed as mean ± standard deviation.Different lowercase letters represent significant differences at 0.05 level ($p < 0.05$), different uppercase letters indicate significant differences at 0.01 level ($p < 0.01$), and the same letters (case-insensitive) show no significant difference ($p > 0.05$).

TABLE 9 Relationship between diploid types and growth traits at 12 months of age in Xinagsu pigs.

| Diplotype | Frequency (%) | B W (kg) | B L (cm) | B H (cm) | C C (cm) | A C (cm) | T C (cm) | L H C (cm) | L B T (mm) |
|---|---|---|---|---|---|---|---|---|---|
| Hap1/3 | 19.0 | 138.37 ± 3.88$^{ab}$ | 129.80 ± 1.21$^{b}$ | 71.60 ± 1.50$^{ab}$ | 119.77 ± 1.85 | 128.57 ± 3.08 | 19.71 ± 0.83$^{ab}$ | 88.29 ± 2.23$^{ab}$ | 14.38 ± 0.97 |
| Hap2/2 | 12.0 | 135.86 ± 3.91$^{b}$ | 130.63 ± 1.18$^{ab}$ | 71.09 ± 1.60$^{b}$ | 120.45 ± 2.06 | 126.68 ± 3.26 | 19.41 ± 0.67$^{b}$ | 86.41 ± 2.79$^{b}$ | 13.61 ± 1.08 |
| Hap2/5 | 7.6 | 138.00 ± 3.80$^{ab}$ | 130.43 ± 1.28$^{ab}$ | 72.50 ± 1.40$^{a}$ | 120.00 ± 2.04 | 126.07 ± 3.67 | 20.50 ± 2.85$^{a}$ | 87.71 ± 2.55$^{ab}$ | 14.23 ± 0.98 |
| Hap2/4 | 6.0 | 136.91 ± 3.91$^{ab}$ | 131.18 ± 0.87$^{a}$ | 72.18 ± 1.66$^{ab}$ | 119.45 ± 2.25 | 127.27 ± 3.32 | 19.73 ± 0.79$^{ab}$ | 86.73 ± 2.33$^{ab}$ | 13.90 ± 1.28 |
| Hap1/1 | 5.4 | 139.10 ± 3.78$^{a}$ | 130.60 ± 1.35$^{ab}$ | 72.30 ± 1.57$^{ab}$ | 120.40 ± 2.37 | 128.20 ± 3.58 | 19.70 ± 0.82$^{ab}$ | 88.40 ± 2.22$^{a}$ | 14.40 ± 0.94 |

Note: BW, Body weight/kg; BL, Body length/cm; BH, Body height/cm; CC, Chest circumference/cm; AC, Abdominal circumference/cm; TC, Tube circumference/cm; LHC, Leg and hip circumference/cm; LBT, Living backfat thickness/mm. The data are expressed as mean ± standard deviation.Different lowercase letters represent significant differences at 0.05 level ($p < 0.05$), different uppercase letters indicate significant differences at 0.01 level ($p < 0.01$), and the same letters (case-insensitive) show no significant difference ($p > 0.05$).

was expressed in all examined tissues, including the heart, liver, spleen, lung, kidney, and longissimus dorsi muscle of Xiangsu pigs. Among them, the expression of *ACAA1* gene in liver was significantly different from that in other tissues ($p < 0.01$). ACAA1 regulates fatty acid oxidation and lipid metabolism by catalyzing peroxisomal fatty acid β-oxidation (Luo et al., 2018). Lipid metabolism is tightly linked to fat deposition in muscle, which directly influences the meat taste of pork products. Liver and muscle are the main metabolic organs involved in the regulation of lipid metabolism, and investigation of the expression level and spatial and temporal changes of the *ACAA1* gene in critical visceral organs, and longissimus dorsi muscle of Xiangsu pigs is highly imperative.

We further examined the expression trend of ACAA1 mRNA in longissimus dorsi muscle of Xiangsu pigs at different ages (newborn, 6-month-old and 12-month-old). Interestingly, the expression level of the *ACAA1* gene in the longissimus dorsi muscle decreased with age. Previous studies have shown that gene expression in tissues varies during different growth stages. The intramuscular fat content in the longissimus thoracis muscle of Tibetan sheep shows an increasing trend from 4 months to 1.5 years old ($p < 0.05$), while the *MYH4* gene exhibits differential expression between the longissimus thoracis muscles at 4 months and 1.5 years old (Wen et al., 2022). Transcriptional analysis was conducted on the longissimus dorsi muscle of pigs at different growth stages, identifying many differentially expressed genes (DEGs) related to lipid metabolism and muscle development, the majority of which are involved in intramuscular fat (IMF) deposition (Li et al., 2023). Knockdown of the *ACAA1* gene promoted lipid droplet formation and lipid accumulation in sheep preadipocytes (Wang et al., 2021), inhibiting the *ACAA1* gene expression promoted intramuscular fat deposition in chicken (Li et al., 2019; Xie et al., 2014). In another investigation, upregulated expression of the *ACAA1* gene in mice was revealed to inhibit abdominal fat and liver lipid accumulation in high-fat diet mice (Xie et al., 2014). Therefore, the *ACAA1* gene could influence fat deposition in the longissimus dorsi muscle of Xiangsu pigs at different ages by regulating lipid metabolism.

In the present investigation, we analyzed blood DNA extracted from 184 Xiangsu pigs to determine the effect of *ACAA1* gene polymorphism on fat deposition in the longissimus dorsi muscle of

Xiangsu pigs. Amplification of the *ACAA1* gene sequence yielded four SNP loci: g.48810 A>G (rs343060194), g.51546 T>C (rs319197012), g.55035 T>C (rs333279910), and g.55088 C>T (rs322138947). g.51546 T>C (rs319197012) is an intron mutation, g.48810 A>G (rs343060194), g.55035 T>C (rs333279910), and g.55088 C>T (rs322138947) are exon synonymous mutations identified using gene polymorphism parameter evaluation. The four mutation sites were consistent with HWE ($p > 0.05$) and had moderate polymorphism ($0.25<PIC < 0.50$). At the same time, synonymous mutations have been shown to alter mRNA splicing and secondary structure, as well as amino acid co-translation and post-translational folding pathways (Sharma et al., 2019; Supek et al., 2014). Human cancer research has also demonstrated that synonymous mutations potentially change RNA binding proteins and miRNA binding sites (Teng et al., 2020). In this view, it is critical to investigate the association between SNPs in introns and exons of the *ACAA1* gene and backfat deposition in Xiangsu pigs.

LD analysis of the four SNPs revealed that the g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) had the strongest linkage and belonged to a strong LD ($r^2 = 1.000$) (Guryev et al., 2006; Slatkin, 2008). Furthermore, previous studies revealed a strong LD between gene exon mutations, which exert a potential synergistic effect on animal phenotypes (Zhao et al., 2021). Therefore, we hypothesize that the strong linkage mutation sites of g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) in the *ACAA1* gene may influence pig growth traits.

The relationship between four SNPs of the *ACAA1* gene and the growth traits of Xiangsu pigs revealed that the strong linkage imbalance sites g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) significantly differed from the body weight ($p < 0.05$) and the living backfat thickness of Xiangsu pigs ($p < 0.01$). Moreover, the heterozygous genotypes of the two loci revealed a dominant genotype in the body weight and living backfat thickness of 12-month-old Xiangsu pigs. These data provided more evidence that these two sites may have a synergistic effect on pig growth and backfat deposition. The g.48810 A>G (rs343060194) locus may primarily influence pig body weight, body length and abdominal circumference, while the g.51546 T>C (rs319197012) locus may influence the growth and backfat deposition of pigs in the later stages of fattening. Emerging evidence indicates that genes related to adipogenesis (Martinez-Montes et al., 2018; Shi et al., 2019) and fatty acid metabolism (Chen et al., 2019; Guo et al., 2017) signaling pathways play a role in pig backfat development (Gozalo-Marcilla et al., 2021) and that pig backfat thickness and body weight are moderately positively correlated (r = 0.632) (Hoa et al., 2021). Diplotypes analysis revealed that Hap1/1, -GCCT/GCCT- were beneficial to growth traits at 6 months of age. Hap1/1, -GCCT/GCCT-, Hap2/5, -GTTC/GCTC- were favorable for growth traits at 12 months of age and could be utilized as advantageous genotype combinations for breeding. The *ACAA1* gene mutations g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) may be linked to the body weight and living backfat thickness of Xiangsu pigs. Therefore, the *ACAA1* gene is a promising candidate gene for pig growth and development and backfat deposition.

## 5 Conclusion

This study investigated the tissue-specific expression of the *ACAA1* gene in Xiangsu pigs. The results showed that the expression level of *ACAA1* gene mRNA was highest in the liver of 6-month-old pigs. The expression level of *ACAA1* gene mRNA in the longissimus dorsi muscle of Xiangsu pigs decreased with age. In addition, this study conducted an association analysis of *ACAA1* gene SNPs with the growth traits of Xiangsu pigs. The results showed that there are four SNPs in the *ACAA1* gene of Xiangsu pigs; g.55035 T>C (rs333279910) and g.55088 C>T (rs322138947) were strongly linked ($r^2 = 1.000$). The strong linkage loci exhibited significant differences in body weight and body height and living backfat thickness. These SNPs potentially influence the growth traits of Xiangsu pigs and are valuable SNP markers for improving the growth performance of Xiangsu pigs.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author.

## Ethics statement

The animal studies were approved by the Animal Welfare Committee of Guizhou University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was obtained from the owners for the participation of their animals in this study.

## Author contributions

MX: Conceptualization, Writing–original draft. YR: Data curation, Methodology, Writing–review and editing. JH: Methodology, Resources, Writing–review and editing. LD: Investigation, Methodology, Writing–review and editing. JX: Data curation, Software, Writing–review and editing. HX: Funding acquisition, Supervision, Validation, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Guizhou Provincial Science and Technology Project (QKHFQ-2018,4007, (002)), and the Guizhou Provincial Agricultural Major Industrial Scientific Research Project (QKHKYZ-2019,011).

## Acknowledgments

Thanks to the Xiangsu Pig breeding Farm of Guizhou University, China for providing Xiangsu pigs.

## 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.2024.1346903/full#supplementary-material

## References


Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3 (4), 299–309. doi:10.1038/nrg777

Bao, W. B., Ye, L., Zhu, J., Pan, Z. Y., Zhu, G. Q., Huang, X. G., et al. (2012). Evaluation of m307 of fut1 gene as a genetic marker for disease resistance breeding of sutai pigs. *Mol. Biol. Rep.* 39 (4), 4223–4228. doi:10.1007/s11033-011-1208-1

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169 (7), 1177–1186. doi:10.1016/j.cell.2017.05.038

Bozorgmehr, A., Moayedi, R., Sadeghi, B., Ghadirivasfi, M., Joghataei, M. T., and Shahbazi, A. (2020). A novel link between the oxytocin receptor gene and impulsivity. *Neuroscience* 444, 196–208. doi:10.1016/j.neuroscience.2020.07.033

Chen, G., Cheng, X., Shi, G., Zou, C., Chen, L., Li, J., et al. (2019). Transcriptome analysis reveals the effect of long intergenic noncoding rnas on pig muscle growth and fat deposition. *Biomed. Res. Int.* 2019, 2951427–2951515. doi:10.1155/2019/2951427

Deng, T., Wu, J., Abdel-Shafy, H., Wang, X., Lv, H., Shaukat, A., et al. (2023). Comparative genomic analysis of the thiolase family and functional characterization of the acetyl-coenzyme a acyltransferase-1 gene for milk biosynthesis and production of buffalo and cattle. *J. Agric. Food. Chem.* 71 (7), 3325–3337. doi:10.1021/acs.jafc.2c07763

Du, A. C., Clutter, A. C., and Lohuis, M. M. (2007). Characterizing linkage disequilibrium in pig populations. *Int. J. Biol. Sci.* 3 (3), 166–178. doi:10.7150/ijbs.3.166

Gozalo-Marcilla, M., Buntjer, J., Johnsson, M., Batista, L., Diez, F., Werner, C. R., et al. (2021). Genetic architecture and major genes for backfat thickness in pig lines of diverse genetic backgrounds. *Genet. Sel. Evol.* 53 (1), 76. doi:10.1186/s12711-021-00671-w

Guo, Y., Qiu, H., Xiao, S., Wu, Z., Yang, M., Yang, J., et al. (2017). A genome-wide association study identifies genomic loci associated with backfat thickness, carcass weight, and body weight in two commercial pig populations. *J. Appl. Genet.* 58 (4), 499–508. doi:10.1007/s13353-017-0405-6

Guryev, V., Smits, B. M. G., de Belt, J. V., Verheul, M., Hubner, N., and Cuppen, E. (2006). Haplotype block structure is conserved across mammals. *PLoS Genet.* 2 (7), e121. doi:10.1371/journal.pgen.0020121

Hoa, V. B., Seo, H. W., Seong, P. N., Cho, S. H., Kang, S. M., Kim, Y. S., et al. (2021). Back-fat thickness as a primary index reflecting the yield and overall acceptance of pork meat. *Anim. Sci. J.* 92 (1), e13515. doi:10.1111/asj.13515

Kumar, A., Shiloach, J., Betenbaugh, M. J., and Gallagher, E. J. (2015). The beta-3 adrenergic agonist (cl-316,243) restores the expression of down-regulated fatty acid oxidation genes in type 2 diabetic mice. *Nutr. Metab.* 12, 8. doi:10.1186/s12986-015-0003-8

Li, G., Fu, S., Chen, Y., Jin, W., Zhai, B., Li, Y., et al. (2019). Microrna-15a regulates the differentiation of intramuscular preadipocytes by targeting acaa1, acox1 and scp2 in chickens. *Int. J. Mol. Sci.* 20 (16), 4063. doi:10.3390/ijms20164063

Li, X., Yang, Y., Li, L., Ren, M., Zhou, M., and Li, S. (2023). Transcriptome profiling of different developmental stages on longissimus dorsi to identify genes underlying intramuscular fat content in wannanhua pigs. *Genes.* 14 (4), 903. doi:10.3390/genes14040903

Li, Y., Liu, X., Niu, L., and Li, Q. (2017). Proteomics analysis reveals an important role for the ppar signaling pathway in dbdct-induced hepatotoxicity mechanisms. *Molecules* 22 (7), 1113. doi:10.3390/molecules22071113

Li, Z., Zhang, Z., He, Z., Tang, W., Li, T., Zeng, Z., et al. (2009). A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn). *Cell Res.* 19 (4), 519–523. doi:10.1038/cr.2009.33

Lisowski, P., Kosciuczuk, E. M., Goscik, J., Pierzchala, M., Rowinska, B., and Zwierzchowski, L. (2014). Hepatic transcriptome profiling identifies differences in expression of genes associated with changes in metabolism and postnatal growth between hereford and holstein-friesian bulls. *Anim. Genet.* 45 (2), 288–292. doi:10.1111/age.12116

Liu, C., Ran, X., Wang, J., Li, S., and Liu, J. (2018). Detection of genomic structural variations in guizhou indigenous pigs and the comparison with other breeds. *PLoS One* 13 (3), e0194282. doi:10.1371/journal.pone.0194282

Liu, F., Li, H., Chang, H., Wang, J., and Lu, J. (2015). Identification of hepatocellular carcinoma-associated hub genes and pathways by integrated microarray analysis. *Tumori* 101 (2), 206–214. doi:10.5301/tj.5000241

Liu, H., Song, H., Jiang, Y., Jiang, Y., Zhang, F., Liu, Y., et al. (2021). A single-step genome wide association study on body size traits using imputation-based whole-genome sequence data in yorkshire pigs. *Front. Genet.* 12, 629049. doi:10.3389/fgene.2021.629049

Luo, C., Zhao, S., Dai, W., Zheng, N., and Wang, J. (2018). Proteomic analysis of lysosomal membrane proteins in bovine mammary epithelial cells illuminates potential novel lysosome functions in lactation. *J. Agric. Food. Chem.* 66 (49), 13041–13049. doi:10.1021/acs.jafc.8b04508

Luo, R., Fan, Y., Yang, J., Ye, M., Zhang, D. F., Guo, K., et al. (2021). A novel missense variant in ACAA1 contributes to early-onset Alzheimer's disease, impairs lysosomal function, and facilitates amyloid-β pathology and cognitive decline. *Signal Transduct. Target. Ther.* 6 (1), 325. doi:10.1038/s41392-021-00748-4

Martinez-Montes, A. M., Fernandez, A., Munoz, M., Noguera, J. L., Folch, J. M., and Fernandez, A. I. (2018). Using genome wide association studies to identify common qtl regions in three different genetic backgrounds based on iberian pig breed. *PLoS One* 13 (3), e0190184. doi:10.1371/journal.pone.0190184

Naicy, T., Venkatachalapathy, R. T., Aravindakshan, T. V., and Kurian, E. (2017). Association of a cac8i polymorphism in the igf1 gene with growth traits in indian goats. *J. Genet. Eng. Biotechnol.* 15 (1), 7–11. doi:10.1016/j.jgeb.2017.04.002

Nei, M., and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* 76 (2), 379–390. doi:10.1093/genetics/76.2.379

Nwosu, Z. C., Battello, N., Rothley, M., Pioronska, W., Sitek, B., Ebert, M. P., et al. (2018). Liver cancer cell lines distinctly mimic the metabolic gene expression pattern of the corresponding human tumours. *J. Exp. Clin. Cancer Res.* 37 (1), 211. doi:10.1186/s13046-018-0872-6

Peakall, R., and Smouse, P. E. (2012). Genalex 6.5: genetic analysis in excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28 (19), 2537–2539. doi:10.1093/bioinformatics/bts460

Peng, W. T., Jin, X., Xu, X. E., Yang, Y. S., Ma, D., Shao, Z. M., et al. (2023). Inhibition of acaa1 restrains proliferation and potentiates the response to cdk4/6 inhibitors in triple-negative breast cancer. *Cancer Res.* 83 (10), 1711–1724. doi:10.1158/0008-5472.CAN-22-2143

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Serrote, C., Reiniger, L., Silva, K. B., Rabaiolli, S., and Stefanel, C. M. (2020). Determining the polymorphism information content of a molecular marker. *Gene* 726, 144175. doi:10.1016/j.gene.2019.144175

Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Gross, M., et al. (2019). A pan-cancer analysis of synonymous mutations. *Nat. Commun.* 10 (1), 2569. doi:10.1038/s41467-019-10489-2

Shi, G., Chen, L., Chen, G., Zou, C., Li, J., Li, M., et al. (2019). Identification and functional prediction of long intergenic non-coding rnas related to subcutaneous adipose development in pigs. *Front. Genet.* 10, 160. doi:10.3389/fgene.2019.00160

Shi, L., Wang, L., Fang, L., Li, M., Tian, J., Wang, L., et al. (2022). Integrating genome-wide association studies and population genomics analysis reveals the genetic architecture of growth and backfat traits in pigs. *Front. Genet.* 13, 1078696. doi:10.3389/fgene.2022.1078696

Shi, Y. Y., and He, L. (2005). Shesis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res.* 15 (2), 97–98. doi:10.1038/sj.cr.7290272

Slatkin, M. (2008). Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9 (6), 477–485. doi:10.1038/nrg2361

Sordillo, J. E., Sharma, S., Poon, A., Lasky-Su, J., Belanger, K., Milton, D. K., et al. (2011). Effects of endotoxin exposure on childhood asthma risk are modified by a genetic polymorphism in acaa1. *BMC Med. Genet.* 12, 158. doi:10.1186/1471-2350-12-158

Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156 (6), 1324–1335. doi:10.1016/j.cell.2014.01.051

Tang, L. T., Ran, X. Q., Mao, N., Zhang, F. P., Niu, X., Ruan, Y. Q., et al. (2018). Analysis of alternative splicing events by rna sequencing in the ovaries of Xiang pig at estrous and diestrous. *Theriogenology* 119, 60–68. doi:10.1016/j.theriogenology.2018.06.022

Taylor, J. G., Choi, E. H., Foster, C. B., and Chanock, S. J. (2001). Using genetic variation to study human disease. *Trends Mol. Med.* 7 (11), 507–512. doi:10.1016/s1471-4914(01)02183-9

Teng, H., Wei, W., Li, Q., Xue, M., Shi, X., Li, X., et al. (2020). Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. *Nucleic. acids. Res.* 48 (3), 1192–1205. doi:10.1093/nar/gkaa019

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on snp and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34 (3), 275–305. doi:10.1186/1297-9686-34-3-275

Wanders, R. J., Vreken, P., Ferdinandusse, S., Jansen, G. A., Waterham, H. R., van Roermund, C. W., et al. (2001). Peroxisomal fatty acid alpha- and beta-oxidation in humans: enzymology, peroxisomal metabolite transporters and peroxisomal diseases. *Biochem. Soc. Trans.* 29 (Pt 2), 250–267. doi:10.1042/0300-5127:0290250

Wang, Y., Li, X., Cao, Y., Xiao, C., Liu, Y., Jin, H., et al. (2021). Effect of the acaa1 gene on preadipocyte differentiation in sheep. *Front. Genet.* 12, 649140. doi:10.3389/fgene.2021.649140

Wen, Y., Li, S., Bao, G., Wang, J., Liu, X., Hu, J., et al. (2022). Comparative transcriptome analysis reveals the mechanism associated with dynamic changes in meat quality of the longissimus thoracis muscle in Tibetan sheep at different growth stages. *Front. Vet. Sci.* 9, 926725. doi:10.3389/fvets.2022.926725

Xie, W., Zhang, S., Lei, F., Ouyang, X., and Du, L. (2014). Ananas comosus l. Leaf phenols and p-coumaric acid regulate liver fat metabolism by upregulating cpt-1 expression. *Evid.-based Complement. Altern. Med.* 2014, 903258. doi:10.1155/2014/903258

Xu, J., Ruan, Y., Sun, J., Shi, P., Huang, J., Dai, L., et al. (2022). Association analysis of prkaa2 and msmb polymorphisms and growth traits of xiangsu hybrid pigs. *Genes.* 14 (1), 113. doi:10.3390/genes14010113

Yan, H., Li, Z., Shen, Q., Wang, Q., Tian, J., Jiang, Q., et al. (2017). Aberrant expression of cell cycle and material metabolism related genes contributes to hepatocellular carcinoma occurrence. *Pathol. Res. Pract.* 213 (4), 316–321. doi:10.1016/j.prp.2017.01.019

Zhang, B., Wu, Q., Wang, Z., Xu, R., Hu, X., Sun, Y., et al. (2019). The promising novel biomarkers and candidate small molecule drugs in kidney renal clear cell carcinoma: evidence from bioinformatics analysis of high-throughput data. *Mol. Genet. Genom. Med.* 7 (5), e607. doi:10.1002/mgg3.607

Zhang, H., Zhuang, Z., Yang, M., Ding, R., Quan, J., Zhou, S., et al. (2021). Genome-wide detection of genetic loci and candidate genes for body conformation traits in duroc × landrace × yorkshire crossbred pigs. *Front. Genet.* 12, 664343. doi:10.3389/fgene.2021.664343

Zhao, H., Hu, R., Li, F., and Yue, X. (2021). Five snps within the fgf5 gene significantly affect both wool traits and growth performance in fine-wool sheep (ovis aries). *Front. Genet.* 12, 732097. doi:10.3389/fgene.2021.732097

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 10 May 2024
DOI 10.3389/fvets.2024.1337939



OPEN ACCESS

EDITED BY
Ana Fabrícia Braga Magalhães,
Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Brazil

REVIEWED BY
Ivan Cunha Bustamante-Filho,
University of Vale do Taquari, Brazil
Fabieli Loise Braga Feitosa,
Federal University of Bahia (UFBA), Brazil

*CORRESPONDENCE
Mustafa Hitit
muhitit@pvamu.edu;
vetdrmustafahitit@gmail.com

RECEIVED 13 November 2023
ACCEPTED 08 April 2024
PUBLISHED 10 May 2024

CITATION
Hitit M, Kaya A and Memili E (2024) Sperm long non-coding RNAs as markers for ram fertility. *Front. Vet. Sci.* 11:1337939. doi: 10.3389/fvets.2024.1337939



# Sperm long non-coding RNAs as markers for ram fertility

Mustafa Hitit[1,2]*, Abdullah Kaya[3] and Erdogan Memili[2]

[1]Department of Genetics, Faculty of Veterinary Medicine, Kastamonu University, Kastamonu, Türkiye, [2]College of Agriculture, Food and Natural Resources, Cooperative Agricultural Research Center, Prairie View A&M University, Prairie View, TX, United States, [3]Department of Animal and Dairy Sciences, College of Agricultural and Life Sciences, University of Wisconsin–Madison, Madison, WI, United States

It is critical in sheep farming to accurately estimate ram fertility for maintaining reproductive effectiveness and for production profitability. However, there is currently a lack of reliable biomarkers to estimate semen quality and ram fertility, which is hindering advances in animal science and technology. The objective of this study was to uncover long non-coding RNAs (lncRNAs) in sperm from rams with distinct fertility phenotypes. Mature rams were allocated into two groups: high and low fertility (HF; $n = 31$; 94.5 ± 2.8%, LF; $n = 25$; 83.1 ± 5.73%; $P = 0.028$) according to the pregnancy rates sired by the rams (average pregnancy rate; 89.4 ± 7.2%). Total RNAs were isolated from sperm of the highest- and lowest-fertility rams ($n = 4$, pregnancy rate; 99.2 ± 1.6%, and 73.6 ± 4.4%, respectively) followed by next-generation sequencing of the transcripts. We uncovered 11,209 lncRNAs from the sperm of rams with HF and LF. In comparison to each other, there were 93 differentially expressed (DE) lncRNAs in sperm from the two distinct fertility phenotypes. Of these, 141 mRNAs were upregulated and 134 were downregulated between HF and LF, respectively. Genes commonly enriched for 9 + 2 motile cilium and sperm flagellum were *ABHD2*, *AK1*, *CABS1*, *ROPN1*, *SEPTIN2*, *SLIRP*, and *TEKT3*. Moreover, *CABS1*, *CCDC39*, *CFAP97D1*, *ROPN1*, *SLIRP*, *TEKT3*, and *TTC12* were commonly enriched in flagellated sperm motility and sperm motility. Differentially expressed mRNAs were enriched in the top 16 KEGG pathways. Targets of the differentially expressed lncRNAs elucidate functions in *cis* and *trans* manner using the genetic context of the lncRNA locus, and lncRNA sequences revealed 471 mRNAs targets of 10 lncRNAs. This study illustrates the existence of potential lncRNA biomarkers that can be implemented in analyzing the quality of ram sperm and determining the sperm fertility and is used in breeding soundness exams for precision livestock farming to ensure food security on a global scale.

KEYWORDS

**ram, sperm, fertility, long non-coding RNAs, gene expression**

## 1 Introduction

Male fertility is crucial for animal reproduction as it encompasses the successful fertilization of an egg with a viable spermatozoon, subsequently leading to embryonic and fetal development, ensuring the species' continued existence. Fertility is an economically important trait, and ram flock represents more than 50% of the genetics of every sheep farmer's flock (1). Thus, proper management of rams for maximum performance and longevity is vital to the success and reproductive efficiency of sheep farming (2). Accordingly, advancing ram fertility in the livestock system is imperative to provide animal-based food demands to feed the ever-increasing human population on a global scale.

In addition to genetic evaluation and testing, rams have been selected according to breeding soundness exam (BSE), which requires a series of examinations, including physical exam, scrotal measurement, sperm morphology, and motility. Despite massive attempts to evaluate ram fertility using the BSE, the predictability of ram fertility still awaits improvement. Non-compensable factors, such as DNA damage, RNA molecules (3), and protein markers (4, 5) in sperm fertility, are attributed to minute sperm abnormalities that cannot be determined using conventional procedures (6). Advanced omics approaches have paved the way for underlying molecular mechanisms related to spermatogenesis, fertilization, and embryogenesis (7, 8). Such methods may be used more widely in the future for producing farm animals in combination with the evaluation of semen parameters (9, 10), while also estimating accurate sperm fertility markers in farm animals (11, 12).

Spermatozoa can transmit not just the paternal DNA but also certain RNA molecules and transcription factors, presumably inherited into the oocytes during fertilization (13). Researchers have recently used cutting-edge approaches for discovering small noncoding RNAs with various nucleotide lengths and biogenesis processes, such as PIWI-interacting RNAs (piRNA), microRNAs (miRNA), tRNAs, and long noncoding RNA (lncRNA)-derived short RNAs (14–17). Sperm bearing RNA molecules are implicated in spermatogenesis and embryo development at transcriptional and posttranscriptional levels (18–20). As such, they play roles in regulating spermatogenesis and fertilization by transferring small noncoding RNA (sncRNA) and lncRNA into the oocyte (3, 19, 21). In addition, paternally derived noncoding RNAs are key regulators of preimplantation embryos (22) since some of them are involved in the control of gene expression in zygotic and early embryonic development (23). Accordingly, sperm noncoding RNAs can enhance the transmission of epigenetic information to the offspring (13) because environmental and metabolic-induced modifications of sperm cells may influence the epigenetic modulation of embryo development by changing the gene expression through noncoding RNAs (24–27).

LncRNAs comprise nucleotides located in the cytoplasm and nucleus, which are transcribed by RNA polymerase II and are longer than 200 nucleotides and lack protein-coding capacity. LncRNAs, with their higher and more stable structures, modulate gene expression (*cis* and *trans* manner) at several levels, including epigenetic, transcriptional, posttranscriptional, and posttranslational, through their interactions with mRNA, proteins, and other sncRNAs. They have significant functions in biological processes such as modifying the chromatin structure, activating gene expression, inhibiting gene expression, and translating mRNA molecules. LncRNAs have crucial roles in regulating the many biological processes that are highly expressed in bovine and mouse testis and mature sperm (28, 29). There is a linkage between noncoding RNA markers and male fertility, as the transcripts can be involved in the prediction of fertility (30, 31). The lncRNAs differentially expressed in distinct motility phenotypes of bovines and humans (32) imply that sperm lncRNA may possess functional roles in fertility (33). Using RNA sequencing technologies, sperm bearing RNA was found to be conversed among many species, such as stallions, goats, and boar (34–36). The unique male sperm ncRNAs with consistent fertility phenotypes can be valuable as potent fertility biomarkers. Accordingly, this study aimed to uncover long noncoding RNA profiles from ram sperm possessing distinct fertility phenotypes.

# 2 Material and methods

## 2.1 Ram fertility assessment and experimental design

The Republic of Turkey's Ministry of Agriculture and Forestry's Institute of Bahri-Dagdaş International Agricultural Research supplied information on the fertility phenotypes of adult rams. Pregnancy rates from natural mating were used to determine the fertility of mature rams ($n = 66$) at least 3–4 years old during the breeding seasons of 2017–2018–2019. The ewes' estrus was detected using teaser rams, who were not permitted to mate, by covering the prepuce area. The teaser rams were presented to the ewes early in the morning for about 30 min. The estrus was considered to be ewes seeking, standing for teasing, and allowing mount attempts by teaser rams. A handler selected estrous ewes and brought them into an enclosure along with a randomly selected single ram for natural mating. Throughout the breeding season, estrus detection was maintained, and ewes were accepted to mate with the chosen ram at random. Ewes were regarded as pregnant if they did not return to estrus within 35 days of mating. In addition, the number of pregnant and non-pregnant ewes for each ram was confirmed by matching the mating and lambing dates according to the duration of pregnancies. The rams' fertility scores were rated based on their conception rates. The rams were split into two groups based on their fertility levels: the high-fertility (HF) group ($n = 31$; 94.5 $\pm$ 2.8%) and the low-fertility (LF) group ($n = 25$; 83.1 $\pm$ 5.73%). The average pregnancy rate was 89.4 $\pm$ 7.2% ($n = 66$), and the rams were divided into these groups based on their fertility levels. We conducted an *a priori* power analysis using G*Power3 (V3.1.9.4) to test the differences between two independent group means using a two-tailed test, an effect size ($d = 2.52$), and an alpha level of .05. The result showed that a total sample of eight animals with two equal-sized groups of $n = 4$ was required to achieve a power of .80 for RNA profiling. However, we excluded the rams that did not have 50 mating and were not used, so we profiled four animals for each group out of a total of 56. During the breeding season, each ram served at least 50 ewes in both groups.

## 2.2 Semen collection

The Bahri-Dagdaş Research Center Ethical Committee, Turkey (Number: 22.12.2016/58), approved the animal procedures. We trained the rams to obtain sperm using an artificial vagina (AV) that enabled them to mount on teaser ewes during estrus. Rams were permitted to ejaculate into the AV upon mounting. The first three collections were discarded prior to the collection of research samples, which was followed by semen collection and processing for use in research. Fertility rates that were 1 standard deviation above or below the mean were termed outliers. Four rams with the highest fertility (pregnancy rate; % 99.2 $\pm$ 1.6) and four rams with

the lowest fertility (pregnancy rate; % 73.6 ± 4.4) were selected for lncRNA sequencing with high confidence. Subsequently, we collected about $2 \times 10^9$/ml spermatozoa per ejaculate, and then the aliquots from each sample were adjusted to a final concentration of $10^7$/ml in straws and frozen at −80°C until lncRNA analysis.

## 2.3 RNA isolation

Prior to RNA isolation, we purified sperm by filtering the semen samples with a 500-mesh sieve to eliminate cell debris. Then, samples were treated with a somatic cell lysis solution (0.3% Triton X-100 and 0.1% SDS in DEPC-treated $H_2O$) for 30 min on ice to eradicate somatic cells, followed by microscopic analysis of non-sperm cell contamination. We isolated total RNA from the purified ram sperm ($n = 4$, for each group) samples using the SanPrep column microRNA miniprep kit (Bio Basic Inc, Canada) with slight modification using the manufacturer's protocols. We added 800 μl of a guanidine–thiocyanate lysis buffer enriched in 20 mM DL-dithiothreitol onto the pellet, and then sperm cells were homogenized by passing the samples through a 26-G needle syringe 20–25 times. After other contaminants were thoroughly removed and total RNA was attached to the membrane, an on-column DNase digestion was carried out to remove any traces of DNA contamination. We evaluated the concentration and integrity of the total RNA samples using a NanoDrop (Colibri Microvolume Spectrometer, Titertek-Berthold, Germany) and a 2100-Bioanalyzer with the RNA 2100 Nano Chip (Applied Biosystems, Carlsbad, CA, USA), respectively.

## 2.4 Library preparation for lncRNA sequencing

Each RNA sample was utilized to prepare 2 g of RNA for the RNA library, and ribosomal RNA was first eliminated using the Ribo-ZeroTM rRNA Removal Kit from Epicenter. Then, using the NEBNext® UltraTM-Directional RNA Library Prep Kit from Illumina® (NEB, USA) in compliance with the commercial kit protocol, sequencing libraries were developed utilizing the rRNA-depleted RNA. Reverse transcription was used to create the first strand of the cDNA following fragmentation with an average length of 200 bp. The Agilent Bioanalyzer 2100 equipment was used to evaluate library quality following product purification using the AMPure XP system. As a result, libraries underwent sequencing using the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, United States), and Novogene Corporation (Beijing, China) generated 150-bp paired-end reads.

## 2.5 RNA-Seq read alignment and transcript assembly

Initially, rRNA, adapter sequences, empty reads, and low-quality reads were eliminated from the raw data. All trimmed reads were confirmed to satisfy the quality threshold (*Q*-score; Q20 and Q30) to ensure that there was no bias in the evaluation step toward approaches that favor maximum read. The Phred scale indicating the reliability of base-calling, with Q20 representing a base call accuracy of 99% (or a 1% chance of error) and Q30 representing a base call accuracy of 99.9% (or a 0.1% chance of error) was used as the quality score. The Ovis aries (v4.0) reference genome was indexed with Bowtie v2.0.6, and the processed paired-end reads were mapped to that genome with HISAT2 2.1.0 (37). Each sample's mapped reads were constructed using StringTie (v1.3.1) (38). Finally, Cuffcompare, a program included in Cufflinks (v2.1.1), was used to annotate the assembled transcripts.

## 2.6 Putative lncRNA identification

To classify newly screened lncRNAs with respect to their positional relationship with known mRNAs, putative lncRNAs were identified. To minimize the false-positive rate (FDR), assembled transcripts were classified to retrieve putative lncRNAs, such as lincRNA, antisense lncRNA, intronic lncRNA, and sense-overlapping. (A) Transcripts with a single exon were eliminated. (B) Transcripts with fewer than 200 nucleotides were eliminated. (C) Using Cufflinks v2.1.1, the annotated lncRNAs in the database were used to exclude the transcripts that overlapped with the exon region of the database annotation. (D) All transcripts with modest levels of expression [FPKM thresholds were set for the categorization of transcript expression levels; genes with very low or no expression (FPKM $< 0.5$), and FPKM $= 0.5$ was chosen as the cutoff to filter out the average read coverage per transcript which was much higher than the other transcript] were omitted (FPKM of a single exon transcript). Using three methods, namely, Coding-Non-Coding-Index (CNCI) (39), Pfam-scan (40), and coded potential calculator (CPC) (41), all estimated transcripts with coding potential were filtered out, and a set of putative lncRNAs was compiled from those with noncoding potential. Using Cuffcompare, the distinct categories of lncRNAs were obtained.

## 2.7 Analysis of mRNA and lncRNA expression levels

The FPKM value was used to evaluate levels of mRNA and lncRNA expression. Cuffdiff (v2.1.1) was used to determine lncRNA FPKM values. Later, the statistically significant DE genes were quantified by a $\log_2$-fold change higher or equal to 2 ($P$-value $< 0.05$) or $P$-adjust $< 0.05$ (applied correction for multiple testing to the $P$-values to FDR), using Ballgown (42).

## 2.8 Target gene prediction

DE lncRNAs were selected to determine target genes. Pearson's correlation was used to assess potential coexpression between lncRNAs and mRNAs. A Pearson correlation >0.7 and a $P$-value of 0.05 were used to determine a positive association between an lncRNA and an mRNA. LncRNAs can act as *cis* regulators by remodeling factors onto local chromatin. We described *cis*-modulated genes as protein-coding genes that were coexpressed

TABLE 1 Summary of data production.

| Sample | Raw_reads | Clean_reads | Raw_data (G) | Clean_data (G) | Error_rate (%) | Q20 (%) | Q30 (%) | GC_ content (%) |
|---|---|---|---|---|---|---|---|---|
| LF | 41,020,716 | 40,183,273 | 12.3 | 12.1 | 0.03 | 96.06 | 90.96 | 60.89 |
| HF | 50,087,458 | 49,069,550 | 15 | 14.7 | 0.03 | 95.39 | 89.5 | 62.37 |

Raw read statistics: the total number of reads for each file is determined, and each set of four consecutive lines represents the information for one read. Clean reads are the same as raw reads, but only the filtered reads are used to calculate the results of any further study; Clean reads (G) the sum of a sequence's length and number, expressed in giga bases; Error rate: the sequencing error rate; Q20, Q30: the proportion of all bases for which the Phred score is 20 or 30 or above, indicating base call accuracy; the proportion of guanine (G) and cytosine (C) in all bases is known as GC content.

with one dysregulated lncRNA and were within 30 kb upstream or downstream in genomic distance in the same allele. To participate in certain biological processes, key transcription factors (TFs) are regulated in a *trans* manner by unique lncRNAs. As a result, we matched these lncRNAs' coexpressed mRNAs to mRNAs that were regulatory targets of specific TFs to anticipate that these lncRNAs might be involved in pathways controlled by these TFs.

## 2.9 Gene ontology terms and KEGG pathway enrichment

We assessed DE mRNAs for gene ontology (GO) enrichment analysis using g:Profiler (43). GO terms possessing a corrected *P*-value $< 0.05$ were accepted as significantly enriched by DE genes. The statistical enrichment of lncRNA target genes was examined for KEGG pathway functional analysis based on the reactome pathway database by WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) with $P < 0.05$ and FDR <5.0% (44). The following parameters were specified for the enrichment analysis. A particular organism was labeled as *Ovis aries* (sheep). Sequential GO analyses (biological process; BP, cellular component; CC, and molecular function; MF) were conducted. The g:SCS method is used to compute multiple testing corrections for *P*-values based on GO and pathway enrichment analysis, with the user threshold set at 0.05.

## 2.10 Statistical analysis

Data were analyzed through SPSS software (version 22.0). Statistical plots were generated using GraphPad Prism 9 (GraphPad Software, USA). During the experimental process, four biological replicates were included with the measurement repeated twice. LF vs. HF groups were analyzed using an independent *t*-test. Significance was accepted at *P*-value = 0.05.

# 3 Results

## 3.1 Overview of sequencing data in ram sperm between LF and HF

In the current study, we pooled two cDNA libraries out of the eight total RNAs (LF and HF, each = 4) obtained from low- and high-fertility ram sperm. The total number of raw readings obtained from all cDNA libraries was 91,108,174. Upon filtering the reads, a total of 89,252,823 clean reads were obtained. The Q20 (%) percentages were 96.06 and 95.39 for LF and HF, respectively. The Q30 (%) percentages were shown to be 90.96 and 89.50 for LF and HF, respectively. The percentages of GC content (%) were shown to be LF, 60.89, and HF, 62.37 (Table 1). Furthermore, after aligning the clean reads with the ovine reference genome through the TopHat2 algorithm, we discovered that the total mapped reads or fragments referring to all samples exceeded 75% and were mapped in the reference genome (Table 2).

TABLE 2 A list of reads that were mapped to the reference genome.

| Sample name | HF | LF |
|---|---|---|
| Total reads | 98,139,100 | 80,366,546 |
| Total mapped | 82,508,774 (84.07%) | 69,017,041 (85.88%) |
| Multiple mapped | 3,526,455 (3.59%) | 2,394,038 (2.98%) |
| Uniquely mapped | 78,982,319 (80.48%) | 66,623,003 (82.90%) |
| Read-1 | 40,351,963 (41.12%) | 33,802,488 (42.06%) |
| Read-2 | 38,630,356 (39.36%) | 32,820,515 (40.84%) |
| Reads map to "+" | 39,407,126 (40.15%) | 33,255,875 (41.38%) |
| Reads map to "−" | 39,575,193 (40.33%) | 33,367,128 (41.52%) |
| Non-splice reads | 77,745,931 (79.22%) | 65,903,933 (82.00%) |
| Splice reads | 1,236,388 (1.26%) | 719,070 (0.89%) |

Total reads: the number of reads after data filtration (clean data); Total mapped represents the total quantity of mappable reads. If a suitable reference genome is available and no contamination occurs during the experimental procedure, the percentage will typically exceed 70%. Numerous mappings: the number of sequences mapped to multiple reference sequence positions; specifically mapped: the number of reads that map to a particular position in the reference sequences. Map to "+"; Reads that map to "−": the number of reads that are mapped to the minus strand. Quantity of reads mapped to two exons; also known as junction reads. Similarly, non-splice reads are those that are completely mapped to a single exon. The ratio of splice readings to total read length is proportional.

TABLE 3 A list of reads that were mapped to the reference genome.

| Classification of mapped reads (LF) | Classification of mapped reads (HF) |
|---|---|
| Others (49,390,745 [78.4%]) | Others (57,960,150 [77.95%]) |
| Protein_coding (12,785,367 [20.3%]) | Protein_coding (15,352,804 [20.65%]) |
| lncRNA (697,609 [1.1%]) | lncRNA (855,897 [1.15%] |
| pseudogene (125,827 [0.2%]) | pseudogene (137,580 [0.2%]) |
| rRNA (10,395 [0.02%]) | rRNA (11,500 [0.02%]) |
| miRNA (5,685 [0.01%]) | miRNA (7,407 [0.01%]) |
| misc_RNA (3,588 [0.0%]) | misc_RNA (13,778 [0.0%]) |
| ribozyme (2,495 [0.0%]) | ribozyme (7,883 [0.0%]) |

Distribution of mapped reads in known types of RNAs for low fertility (LF) and high fertility (HF).

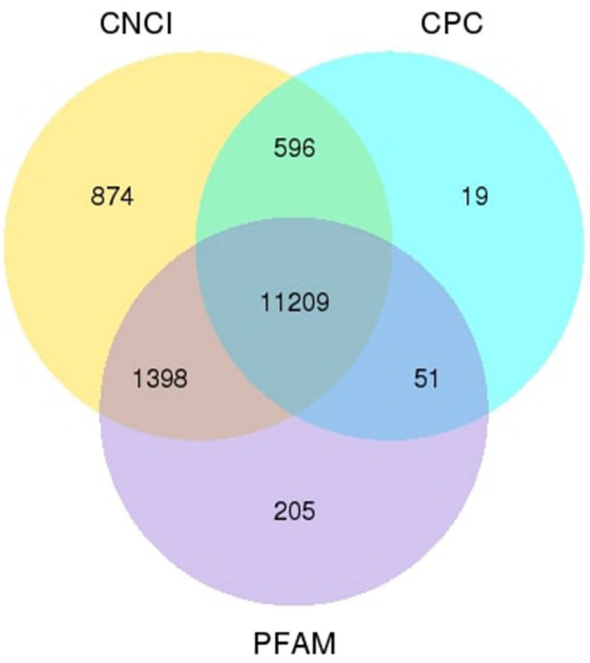



FIGURE 1
Venn diagram. The numerical values within each circle and the areas of overlap symbolize the total count and the number of noncoding transcripts that are found out by the software.

## 3.2 Chromosome read distribution and known RNA types

The distribution intensity of the total mapped reads was split up and computed for both the plus and minus strands within each chromosome (Supplementary Table 1). The distributions of reads in the known RNA types for LF and HF are depicted in Table 3. The corresponding data related to these distributions can be found in Supplementary Table 2.

## 3.3 Novel lncRNA identification in ram sperm between LF and HF

We applied the lncRNA filtering method to determine the novel lncRNA candidates in ram sperm. The specific filtering technique is outlined in Supplementary Table 3. Using this lncRNA filtering process, we identified a total of 14,352 lncRNAs, of which 11,209 were novel candidates, the data for which are supplied in Supplementary Table 4. We utilized three distinct software packages (CNCI, CPC, and Pfam) to estimate the potential protein-coding ability of the transcripts (Figure 1).

## 3.4 Identification and characteristics of lncRNAs between LF and HF in ram sperm

The primary characteristics of the lncRNAs' biotype distribution and length, exon intensity, and ORF length were examined as a consequence of the sequence analysis. Using Cufflinks, we divided lncRNAs into three groups: long intervening noncoding RNA (lincRNA), antisense lncRNA, and intronic lncRNA. Of these, 8.3%, 8.5%, 9.8%, 72.6%, and 0.8% of the attained lncRNAs were sense intronic, antisense, sense overlapping, lincRNA, and others, respectively (Figure 2A). Using this classification scheme, we revealed that the majority of the sperm-specific lncRNAs (72.6%) were lincRNAs. The exon lengths of the obtained lncRNAs were between 126 and 20,040 bp, and approximately 37% of the lncRNAs were demonstrated to be intense in 946–1,356 bp length (Figure 2B), with 1,542 bp being the median value. The exon length of the obtained mRNAs ranged between 75 and 26,726 bp, with around 10% of the mRNAs being intense at 835–1,025 bp (Figure 2B). Furthermore, lncRNAs have been shown to be more abundant in the first and fourth exons, whereas mRNAs were found to be more abundant in the first and seventh exons (Figure 2C). The length of the ORF of the lncRNAs was between 54 and 20,037 bp, with a median value of 1,461 bp (Figure 2D). For mRNAs, the ORF length was reported as 3–26,460, and the median value was 1,332.

The expressions associated with transcripts of lncRNA, mRNA, novel lncRNAs, and TUCP were evaluated using Cuffdiff. The most common approach for predicting gene expression levels, FPKM, is based on the effects of gene length on sequence depth and fragments described in the RNA sequence. The FPKM values associated with samples were calculated (Supplementary Table 5). Upon comparing FPKM values between LF and HF, no significant change was observed (Figures 3A, B). Moreover, the expression levels related to novel lncRNAs were the highest among the transcripts. TUCP and novel mRNA transcripts were greater than those of lncRNA and mRNAs, while lncRNA transcripts were similar to those of mRNA (Figures 3C, D).

## 3.5 LncRNAs and mRNAs that are differentially expressed in LF and HF

Cuffdiff was employed to identify differentially expressed lncRNAs, mRNAs, and TUCPs. As a result, in HF and LF ram sperm, 93 lncRNAs and 275 mRNAs were reported to be differentially expressed (DE). We discovered 49 lncRNAs that are significantly upregulated and 44 that are downregulated between LF and HF groups. In addition, we discovered that 141 mRNAs were upregulated, whereas 134 were downregulated. Volcano plots show the up and downregulation highlights (Figures 4A, B; Supplementary Table 6). We also examined the expression patterns of DE lncRNAs and mRNAs using hierarchical clustering analysis, which serves as an additional approach for shedding light on differentially expressed genes by grouping genes with comparable expression patterns. The transcript FPKMs were utilized for hierarchical clustering, with discrete colors indicating the direction of the expression level. The clustering of genes on the left was caused by similar expressions (fold change >2, $P < 0.05$) and LF and HF ram sperm, whereas the expression change from blue to red indicated them as increasingly increased (Figures 4C, D).

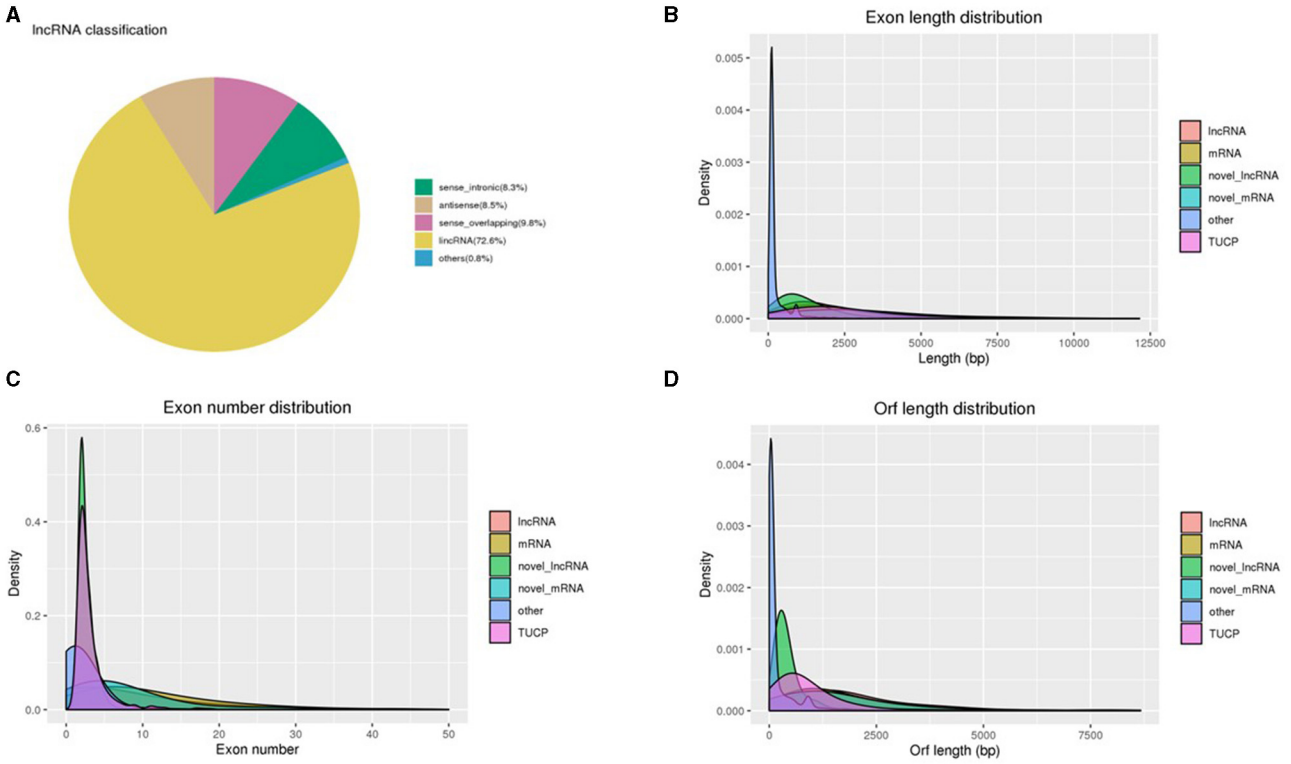



FIGURE 2
**(A)** Distribution of various lncRNA types. **(B)** Comparison of the lengths of lncRNAs, mRNAs, novel lncRNA–mRNA, and TUCP. **(C)** A comparison of the exon counts of lncRNAs and mRNAs. **(D)** A comparison of the ORF lengths of lncRNAs and mRNAs.

## 3.6 DE mRNA functional annotation and KEGG pathways analysis

Following a study at the transcriptome level, 275 mRNAs were selected for enrichment analysis due to their differential expression levels (141 upregulated and 134 downregulated). The top eight GO annotations terms in biological process (BP) were significantly enriched in the DE mRNAs, namely, the biological process (GO:0008150), cellular process (GO:0009987), regulation of cellular process (GO:0050794), organelle organization (GO:0006996), flagellated sperm motility (GO:0030317), sperm motility (GO:0097722), and cilium movement involved in cell motility (GO:0060294). In addition, in cellular component, intracellular anatomical structure (GO:0005622), sperm flagellum (GO:0036126), and motile cilium (GO:0031514) were among the top GO terms and binding (GO:0005488), protein binding (GO:0005515), enzyme binding (GO:0019899), and protein C-terminus binding (GO:0008022) in molecular function (MF; Figure 5A). The detailed definition of the GO terms is presented in Supplementary Figure 1. According to the statistics of the pathway enrichment, the top 16 KEGG pathways, including complex I biogenesis, the citric acid (TCA) cycle, respiratory electron transport, respiratory electron transport, mitochondrial translation, and gene silencing by RNA, had higher concentrations of DE mRNAs. *ABHD2, AK1, CABS1, ROPN1, SEPTIN2, SLIRP,* and *TEKT3* genes were commonly enriched for 9 + 2 motile cilium, sperm flagellum, and motile cilium. Moreover, *CABS1, CCDC39, CFAP97D1, ROPN1, SLIRP, TEKT3,* and *TTC12* were commonly enriched in flagellated sperm motility, sperm motility, and cilium movement involved in cell motility (Supplementary Figure 1).

## 3.7 Target genes of cis- or trans-regulated by lncRNAs

To better understand how lncRNAs act in both *cis* and *trans* manner, we predicted their targets. There were 471 mRNAs that were discovered as targets of 10 lncRNAs (five upregulated and five downregulated), using 30 kb as the cutoff (Supplementary Figure 2). According to the GO enrichment study results, 55 significant GO terms were found (corrected *P*-value 0.05). In MF, the top five GO keywords were nucleoside phosphate binding, protein binding, small-molecule binding, binding, and nucleotide binding (Supplementary Figure 3). In BP, negative regulation of the cellular metabolic process, system development, response to chemical stimuli, organonitrogen compound metabolic process, and multicellular organism development were among the top five GO terms and cytoplasm, cell junction, cytosol, nucleoplasm, and synapse in CC (Figure 5B).

# 4 Discussion

The quality of sperm on a cellular level alone is no longer considered to be a reliable predictor of male fertility in selecting

10.3389/fvets.2024.1337939

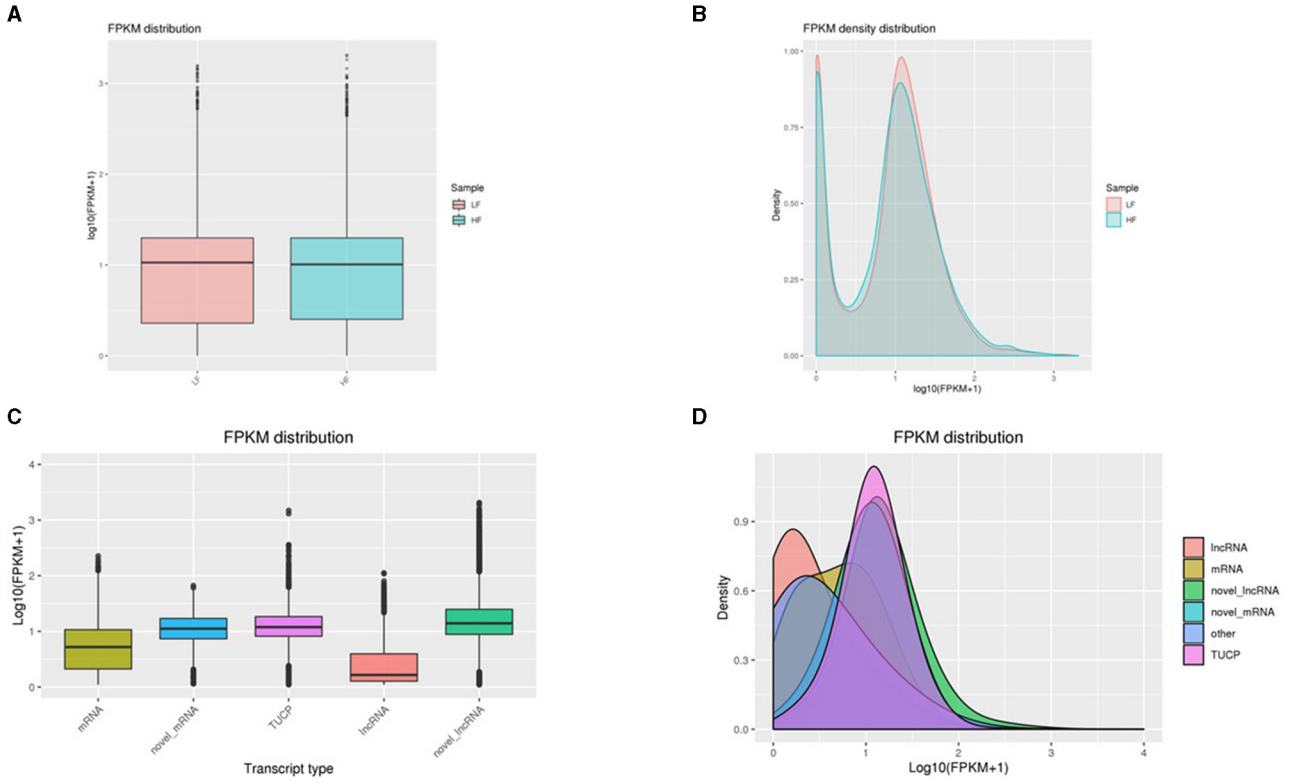



FIGURE 3
**(A, B)** A comparison of gene expressions. Box plot depicting FPKM values. The *X* and *Y* axes reflect the corresponding sample name and $\log_{10}$(FPKM + 1) value, respectively. For each sample, the plot region reflects, from top to bottom, the maximum, upper quartile, median, lower quartile, and minimum statistics. (2) Density distribution of Fragments Per Kilobase of transcript per Million mapped reads. The *X* and *Y* axes show, respectively, the value of $\log_{10}$(FPKM + 1) and the density of genes. **(C, D)** Violin plot for distinct forms of lncRNA, mRNA, and TUCP transcription. The *X*-axis of the FPKM violin plot displays the sample names, while the *Y*-axis displays the $\log_{10}$(FPKM + 1). Each violin plot possesses five statistical parameters (max value, upper quartile, median, lower quartile, and min value). The breadth of the violin plot indicates gene density.

superior male prospects in livestock as a source of frozen semen (12, 31). Methods for analyzing the sperm transcriptome, such as measuring messenger RNA levels, have been linked to increased fertility (45). Analysis of the sperm transcriptome has become the primary tool for predicting male fertility potential in the livestock business, due in part to the widespread application of this technology in livestock growth (45, 46). Accordingly, this study demonstrated long noncoding RNA profiles in sperm from rams with HF and LF phenotypes to uncover potential lncRNAs associated with fertility.

The comparative analysis across multiple species indicates a considerable gap in our understanding of the functional roles of lncRNAs within sperm cells despite their prevalence in male germ cell development. Previous studies on cattle, boar, sheep, mice, and humans have collectively identified substantial numbers of potential lncRNAs in sperm cells (46–48). However, the functional annotation and investigation of these lncRNAs have remained limited. We obtained a total of 91,108,174 cDNA libraries, of which 89,252,823 were clean reads from ram spermatozoa. This provides a robust dataset for comparative analysis, especially when considering the consistency with similar studies conducted, such as on human, turkey, and boar (79.8, 84, and about 65.5 million, respectively) spermatozoa (49, 50). Furthermore, the range of unique mapped reads, ranging from 66,623,003 to 78,982,319, indicated a strong alignment of the sequenced data to the reference genome of sheep. The high percentage of over 80% uniquely mapped reads demonstrated a significant mapping consistency comparable to that observed in bull and stallion spermatozoa (32, 36). This consistency across species further confirms the reliability and quality of the obtained sequence data for ram spermatozoa, supporting the validity of results.

In this study on ram fertility, we identified 11,209 sperm lncRNAs, a subset comprising 93 differentially expressed lncRNAs associated with LF and HF phenotypes in ram sperm. Comparatively, in mouse mature spermatozoa, 4,088 novel lncRNA transcripts were identified out of 20,907 known lncRNA transcripts (29), demonstrating the complexity and diversity of these transcripts in different species' sperm cells. Similarly, in human sperm, 27,472 novel lncRNAs were discovered (29), with 19 differential expressions of lncRNA out of 11,561 lncRNA transcripts in mature bull spermatozoa (32). Analysis of goat spermatozoa revealed 655 lncRNA transcripts relevant to spermatogenesis from sequencing of the cDNA library, of which 1% annotated to lncRNAs was similar to ram spermatozoa with a classification of 1.1% (51). We also showed that the type of lincRNA seemed to be closer between ram and goat spermatozoa

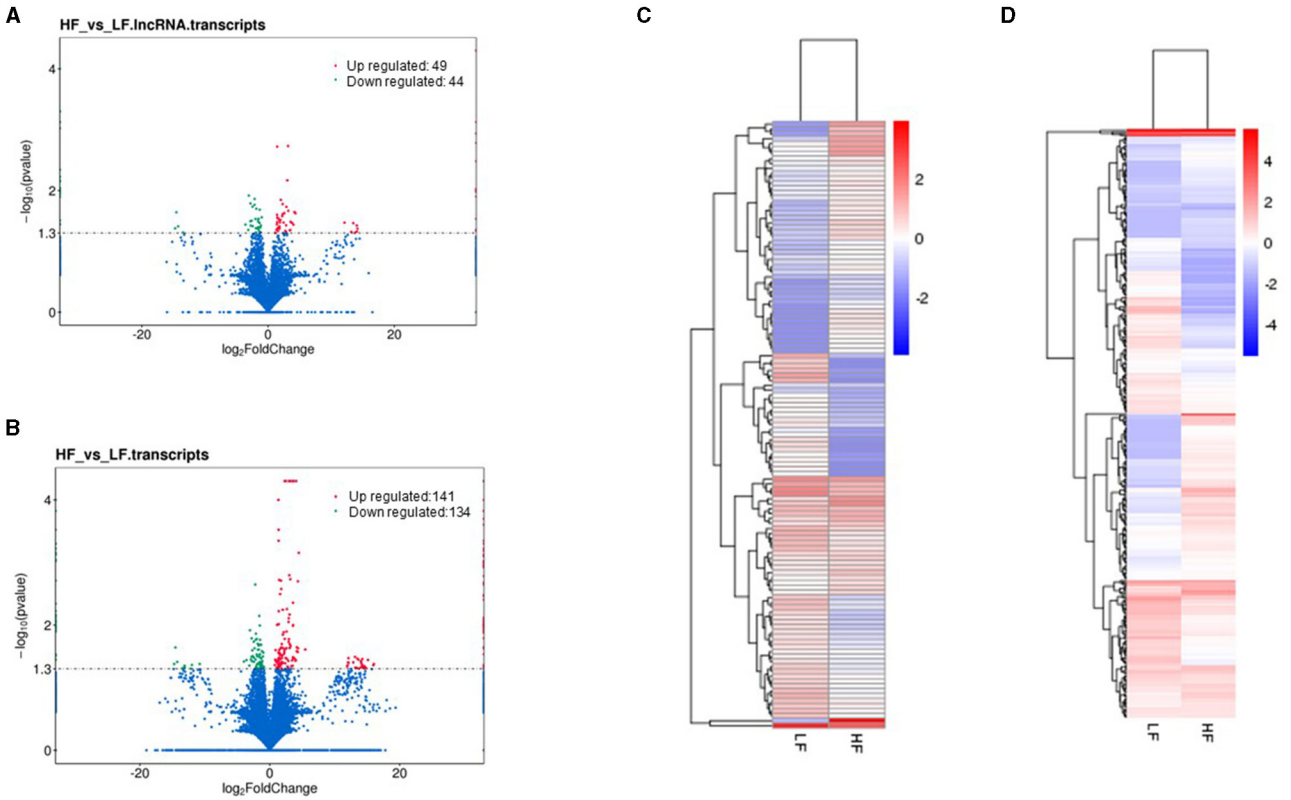


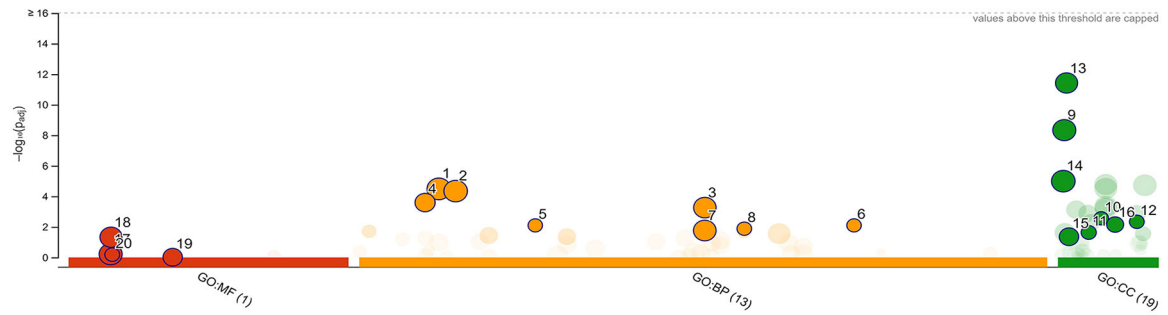

FIGURE 4
Volcano plot: **(A, B)** show differentially expressed lncRNAs and mRNAs in sperm between LF and GF rams, respectively. Heat map of clustering of genes; **(C, D)** show differentially expressed mRNAs and lncRNAs between LF and HF ram sperm (fold change >2, $p < 0.05$), respectively. Colors indicate the following red; up and blue; down.

(34), both revealing a 72.6 and 76.64% annotation to lncRNAs, respectively, which highlights the relevance of these findings across species and their potential implications for understanding male fertility mechanisms.

The GO and KEGG investigations were carried out for associated genes of DE mRNAs and lncRNAs to completely examine the functional roles of mRNAs and lncRNAs in ram sperm fertility. Candidate genes have been identified using this bioinformatics approach, and they were related to male reproductive biology. Our findings demonstrated that 275 mRNA transcripts were enriched for the biological process, cellular component, and molecular function GO characterizations. It has become apparent that several metabolic pathways and regulatory mechanisms have crucial roles in fertility, as related to sperm motility. We demonstrated that *ABHD2*, *AK1*, *CABS1*, *ROPN1*, *SEPTIN2*, *SLIRP*, and *TEKT3* genes were commonly enriched for the 9 + 2 motile cilium, sperm flagellum, and motile cilium. Of these, *ABHD2*, an isolated molecule from sperm tails, is needed to activate sperm (52). *AK1* is an enzyme that is often responsible for cellular energy balance. It is detected in the flagella of murine and bovine sperm, which suggests that it is involved in sperm motility and is also demonstrated to be directly associated with bull fertility (53–55). Moreover, *CABS1*, *CCDC39*, *CFAP97D1*, *ROPN1*, *SLIRP*, *TEKT3*, and *TTC12* were commonly enriched in flagellated sperm motility, sperm motility, and cilium movement involved in cell motility. As a $Ca^{2+}$ storage protein in mature sperm, *CABS1* is a crucial factor in the regulation of calcium signaling and has been shown to preserve sperm flagella structure (56, 57). *ROPN1*, implicated in fibrous sheath integrity and sperm motility, is engaged in PKA-dependent signaling for spermatozoa capacitation; therefore, mutation and lack of its expression in murine sperm cells cause impaired fertility (58, 59). We showed that differentially abundant protein profiles of sperm from rams with contrasting fertility phenotypes were associated with metabolic energy generation by sperm cells along with the motility signaling pathway (4). Our results are consistent with results from the previous study that mRNA–lncRNA interaction seemed to regulate signaling pathways for functional motility.

The primary potential roles of lncRNAs are to control the expression of neighboring protein-coding genes through both *cis* and *trans* manners, integrating transcriptional coactivation or repression. Consequently, we conducted an in-depth analysis and determined the mRNAs situated within the 30 kilobase (kb) threshold upstream and downstream regions of the differentially expressed lncRNAs. We then employed GO and KEGG analyses on the target genes to determine the undertaken lncRNAs. We found that lncRNAs with different levels of expression are involved in a number of important biological processes. We demonstrated that the identified lncRNA TCONS_00136350 may regulate the differentially expressed coding gene ADAM metallopeptidase

A




| ID | Source | Term ID | Term Name | $p_{adj}$ (query... |
|---|---|---|---|---|
| 1 | GO:BP | GO:0008150 | biological_process | $3.254\times10^{-5}$ |
| 2 | GO:BP | GO:0009987 | cellular process | $4.535\times10^{-5}$ |
| 3 | GO:BP | GO:0050794 | regulation of cellular process | $5.474\times10^{-4}$ |
| 4 | GO:BP | GO:0006996 | organelle organization | $2.551\times10^{-4}$ |
| 5 | GO:BP | GO:0030317 | flagellated sperm motility | $7.971\times10^{-3}$ |
| 6 | GO:BP | GO:0097722 | sperm motility | $7.971\times10^{-3}$ |
| 7 | GO:BP | GO:0050789 | regulation of biological process | $1.730\times10^{-2}$ |
| 8 | GO:BP | GO:0060294 | cilium movement involved in cell motility | $1.311\times10^{-2}$ |
| 9 | GO:CC | GO:0005622 | intracellular anatomical structure | $4.664\times10^{-9}$ |
| 10 | GO:CC | GO:0036126 | sperm flagellum | $2.743\times10^{-3}$ |
| 11 | GO:CC | GO:0031514 | motile cilium | $2.320\times10^{-2}$ |
| 12 | GO:CC | GO:0097729 | 9+2 motile cilium | $4.618\times10^{-3}$ |
| 13 | GO:CC | GO:0005737 | cytoplasm | $3.801\times10^{-12}$ |
| 14 | GO:CC | GO:0005575 | cellular_component | $1.016\times10^{-5}$ |
| 15 | GO:CC | GO:0005856 | cytoskeleton | $4.426\times10^{-2}$ |
| 16 | GO:CC | GO:0048471 | perinuclear region of cytoplasm | $7.027\times10^{-3}$ |
| 17 | GO:MF | GO:0005488 | binding | $6.164\times10^{-1}$ |
| 18 | GO:MF | GO:0005515 | protein binding | $4.730\times10^{-2}$ |
| 19 | GO:MF | GO:0019899 | enzyme binding | $9.666\times10^{-1}$ |
| 20 | GO:MF | GO:0008022 | protein C-terminus binding | $6.585\times10^{-1}$ |

B

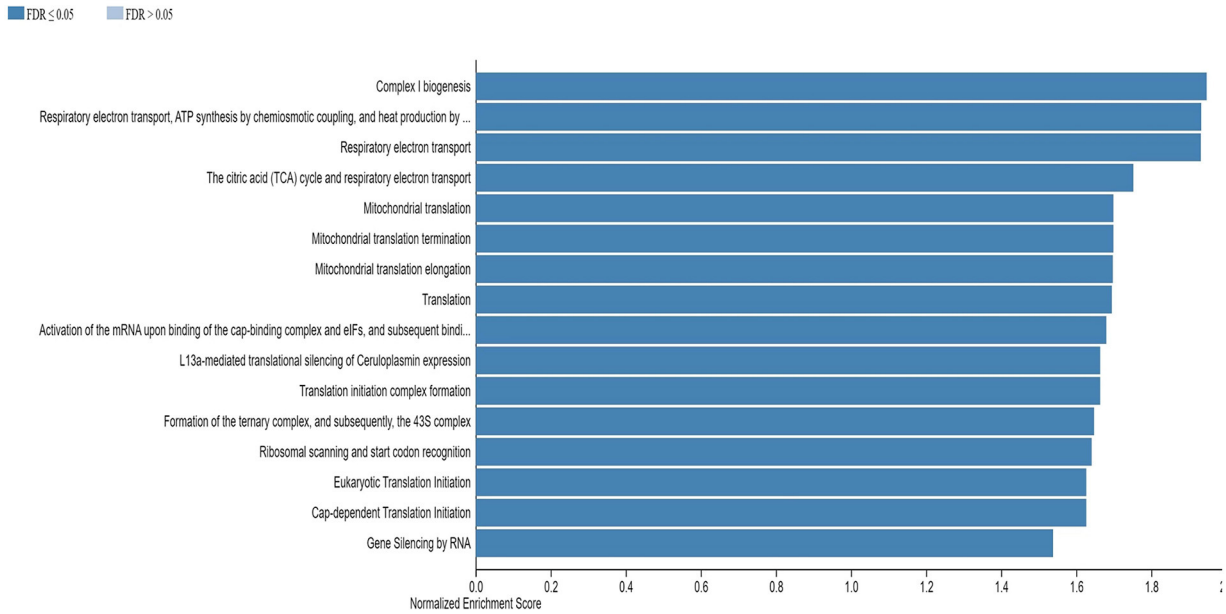



FIGURE 5
**(A)** Functional enrichment analysis of DE mRNAs in LF and HF sperm of ram. Significantly enriched target gene terms are shown. GO keywords are represented by red color codes, molecular functions by orange ones, biological processes by green ones, and cellular components by green ones (CC). **(B)** Functional enrichment analysis of target genes of *cis- or trans*-regulated by lncRNAs. Illustrated are normalized enrichment scores for specific sets of target genes.

domain 32 (*ADAM32*). *ADAM32* is a member of the disintegrin family of membrane-anchored proteins and is detected on the surface of mature sperm (60). *ADAM32* has been demonstrated to be implicated in sperm–egg plasma membrane interaction, and (61) its expression levels are correlated with the quality of human sperm (62). Although *ADAM32* is not required for normal male fertility (63), we showed that it was identified in the top 10 differentially abundant proteins in ram sperm (4). We also discovered that differentially expressed lncRNA TCONS_00035618 modulated enolase 1 (*ENO1*), the protein expression of which was higher in increased fertility of ram sperm (4). *ENO1* plays an important role in the process of metabolism of monocarboxylic acids, and it is a component in the pathway that leads to glycolysis and gluconeogenesis (KEGG:00010) associated with bull fertility (64), and its lower expression levels lead to low motility (65). In line with our study and of others, lncRNAs are potentially important for the coactivation/repression of target genes that regulate sperm fertility and motility.

## 5 Conclusions

Even though several potential lncRNAs have already been identified through cutting-edge technology and accessible lncRNA annotation tools, functional annotation of lncRNAs in sperm biology is still in its infancy and holds great promise. In this study, differentially expressed lncRNAs in ram sperm were ascertained along with their associations with low *vs.* high fertility phenotypes. These findings are important because they help advance both the fundamental science of mammalian male gamete biology and applied science that may provide practical value for potential male fertility markers.

## Data availability statement

The datasets generated and/or analyzed during the current study are available in the EMBL-EBI HYPERLINK http://www.ebi.ac.uk/biostudies/studies/S-BSST1231 accessible at this accession number "S-BSST1231". The data encompass all raw relevant data.

## Ethics statement

The animal study was approved by Bahri-Dagdas, Research Center Ethical Committee, Turkey (Number: 22.12.2016/58). The study was conducted in accordance with the local legislation and institutional requirements.

## Author contributions

MH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AK: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. EM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the Kastamonu University, Scientific Research Coordination Unit with a project number of KÜ-BAP01/2020-63. Animal materials provided were by the Scientific and Technological Research Council of Turkey (TUBITAK) with project no: (117O992) grant to AK.

## 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



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2024.1337939/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Distribution of lncRNA on chromosomes. **(A)** Distribution of lncRNAs in low fertility (LF) ram group. **(B)** Distribution of lncRNAs in high fertility (HF) ram group.

SUPPLEMENTARY FIGURE 2
Identification pipeline for lncRNAs Identification pipeline for lncRNAs. Each step is documented in detail in the Methods section.

SUPPLEMENTARY FIGURE 3
Functional enrichment analysis of target genes of Cis- or Trans-regulated by lncRNAs. Illustrated are normalized enrichment scores for specific sets of target genes. Significantly enriched target gene terms were shown. GO keywords are represented by red color codes, molecular functions by orange ones, biological processes by green ones, and cellular components by green ones (CC).

# References


1. MacLaren APC. Ram fertility in South-West Scotland. *Br Vet J.* (1988) 144:45–54. doi: 10.1016/0007-1935(88)90151-0

2. Pardos L, Rubio MTM, Fantova E. The diversity of sheep production systems in Aragón (Spain): characterisation and typification of meat sheep farms. *Spanish J Agric Res.* (2008) 6:497–507. doi: 10.5424/sjar/2008064-344

3. Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, et al. Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res.* (2013) 41:4104–17. doi: 10.1093/nar/gkt132

4. Hitit M, Özbek M, Ayaz-Guner S, Guner H, Oztug M, Bodu M, et al. Proteomic fertility markers in ram sperm. *Anim Reprod Sci.* (2021) 235:106882. doi: 10.1016/j.anireprosci.2021.106882

5. Kaya A, Dogan S, Peter V, Kutchy NA, Ross P, Topper E, et al. Sperm proteins ODF2 and PAWP as markers of fertility in breeding bulls. *Cell Tissue Res.* (2021) 387:159. doi: 10.1007/s00441-021-03529-1

6. Ugur MR, Saber Abdelrahman A, Evans HC, Gilmore AA, Hitit M, Arifiantini RI, et al. Advances in Cryopreservation of Bull Sperm. *Front Vet Sci.* (2019) 6:268. doi: 10.3389/fvets.2019.00268

7. Long JA. The 'omics' revolutionse of genomic, transcriptomic, proteomic and metabolomic tools to predict male reproductive traits that impact fertility in livestock and poultry. *Anim Reprod Sci.* (2020) 220:106354. doi: 10.1016/j.anireprosci.2020.106354

8. Carrell DT, Aston KI, Oliva R, Emery BR, De Jonge CJ. The "omics" of human male infertility: integrating big data in a systems biology approach. *Cell Tissue Res.* (2016) 363:295–312. doi: 10.1007/s00441-015-2320-7

9. de Azevedo Viana AG, Ribeiro IM, Rodrigues Carvalho RP, Memili E, Moura AA, Machado-Neves M. Functional attributes of seminal proteins in bull fertility: a systematic review. *Reproduction.* (2021) 161:459–75. doi: 10.1530/REP-20-0392

10. Ugur MR, Dinh T, Hitit M, Kaya A, Topper E, Didion B, et al. Amino acids of seminal plasma associated with freezability of bull sperm. *Front Cell Dev Biol.* (2020) 7:347. doi: 10.3389/fcell.2019.00347

11. Parisi AM, Thompson SK, Kaya A, Memili E. Molecular, cellular, and physiological determinants of bull fertility. *Turkish J Vet Anim Sci.* (2014) 38:8. doi: 10.3906/vet-1404-76

12. Özbek M, Hitit M, Kaya A, Jousan FD, Memili E. Sperm functional genome associated with bull fertility. *Front Vet Sci.* (2021) 8:610888. doi: 10.3389/fvets.2021.610888

13. Chen Q, Yan W, Duan E. Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. *Nat Rev Genet.* (2016) 17:733–43. doi: 10.1038/nrg.2016.106

14. Feugang JM, Rodriguez-Osorio N, Kaya A, Wang H, Page G, Ostermeier GC, et al. Transcriptome analysis of bull spermatozoamplications for male fertility. *Reprod Biomed Online.* (2010) 21:312–24. doi: 10.1016/j.rbmo.2010.06.022

15. Krawetz SA, Kruger A, Lalancette C, Tagett R, Anton E, Draghici S, et al. survey of small RNAs in human sperm. *Hum Reprod.* (2011) 26:3401–12. doi: 10.1093/humrep/der329

16. Jodar M, Anton E. Small RNAs present in semen and their role in reproduction. Inorcajadas JA, Gosálvez J, editors. *Reproductomics. The Omics Revolution and its Impact on Human Reproductive Medicine.* Cambridge, MA: Academic Press (2018). p. 109–23. doi: 10.1016/B978-0-12-812571-7.00008-3

17. Peng H, Shi J, Zhang Y, Zhang H, Liao S, Li W, et al. A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. *Cell Res.* (2012) 22:1609–12. doi: 10.1038/cr.2012.141

18. Hayashi K, Lopes SMCS, Tang F, Surani MA. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell.* (2008) 3:391–401. doi: 10.1016/j.stem.2008.07.027

19. Luo LF, Hou CC, Yang WX. Small non-coding RNAs and their associated proteins in spermatogenesis. *Gene.* (2016) 578:141–57. doi: 10.1016/j.gene.2015.12.020

20. Watanabe T, Cheng EC, Zhong M, Lin H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* (2015) 25:368–80. doi: 10.1101/gr.180802.114

21. García-López J, Alonso L, Cárdenas DB, Artaza-Alvarez H, De Dios Hourcade J, Martínez S, et al. Diversity and functional convergence of small noncoding RNAs in male germ cell differentiation and fertilization. *RNA.* (2015) 21:946–62. doi: 10.1261/rna.048215.114

22. Liu WM, Pang RTK, Chiu PCN, Wong BPC, Lao K, Lee KF, et al. Sperm-borne microRNA-34c is required for the first cleavage division in mouse. *Proc Natl Acad Sci U S A.* (2012) 109:490–4. doi: 10.1073/pnas.1110368109

23. Yuan S, Schuster A, Tang C, Yu T, Ortogero N, Bao J, et al. Sperm-borne miRNAs and endo-siRNAs are important for fertilization and preimplantation embryonic development. *Dev.* (2016) 143:635–47. doi: 10.1242/dev.131755

24. Sharma U. Paternal contributions to offspring health: role of sperm small rnas in intergenerational transmission of epigenetic information. *Front Cell Dev Biol.* (2019) 7:215. doi: 10.3389/fcell.2019.00215

25. Gapp K, Jawaid A, Sarkies P, Bohacek J, Pelczar P, Prados J, et al. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat Neurosci.* (2014) 17:667–9. doi: 10.1038/nn.3695

26. Rodgers AB, Morgan CP, Leu NA, Bale TL. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. *Proc Natl Acad Sci U S A.* (2015) 112:13699–704. doi: 10.1073/pnas.1508347112

27. Skinner MK, Ben Maamar M, Sadler-Riggleman I, Beck D, Nilsson E, McBirney M, et al. Alterations in sperm DNA methylation, non-coding RNA and histone retention associate with DDT-induced epigenetic transgenerational inheritance of disease. *Epigenet Chromatin.* (2018) 11:8. doi: 10.1186/s13072-018-0178-0

28. Gao Y, Li S, Lai Z, Zhou Z, Wu F, Huang Y, et al. Analysis of long non-coding RNA and mRNA expression profiling in immature and mature bovine (*Bos taurus*) testes. *Front Genet.* (2019) 10:646. doi: 10.3389/fgene.2019.00646

29. Zhang X, Gao F, Fu J, Zhang P, Wang Y, Zeng X. Systematic identification and characterization of long non-coding RNAs in mouse mature sperm. *PLoS ONE.* (2017) 12:e0173402. doi: 10.1371/journal.pone.0173402

30. Menezes ESB, Badial PR, El Debaky H, Husna AU, Ugur MR, Kaya A, et al. Sperm miR-15a and miR-29b are associated with bull fertility. *Andrologia.* (2020) 52:e13412. doi: 10.1111/and.13412

31. Kaya A, Memili E. Sperm macromolecules associated with bull fertility. *Anim Reprod Sci.* (2016) 169:88–94. doi: 10.1016/j.anireprosci.2016.02.015

32. Wang X, Yang C, Guo F, Zhang Y, Ju Z, Jiang Q, et al. Integrated analysis of mRNAs and long noncoding RNAs in the semen from Holstein bulls with high and low sperm motility. *Sci Rep.* (2019) 9:2092. doi: 10.1038/s41598-018-38462-x

33. Dai YB, Lin Y, Song N, Sun F. LncRNA4667 is dispensable for spermatogenesis and fertility in mice. *Reprod Dev Med.* (2019) 3:18. doi: 10.4103/2096-2924.255985

34. Chen H, Miao X, Xu J, Pu L, Li L, Han Y, et al. Alterations of mRNA and lncRNA profiles associated with the extracellular matrix and spermatogenesis in goats. *Anim Biosci.* (2022) 35:544–55. doi: 10.5713/ab.21.0259

35. Fraser L, Paukszto Ł, Mańkowska A, Brym P, Gilun P, Jastrzebski JP, et al. Regulatory potential of long non-coding RNAs (lncRNAs) in boar spermatozoa with good and poor freezability. *Life.* (2020) 10:1–26. doi: 10.3390/life10110300

36. Das PJ, McCarthy F, Vishnoi M, Paria N, Gresham C, Li G, et al. Stallion sperm transcriptome comprises functionally coherent coding and regulatory RNAs as revealed by microarray analysis and RNA-seq. *PLoS ONE.* (2013) 8:e56535. doi: 10.1371/journal.pone.0056535

37. 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

38. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* (2016) 11:1650–67. doi: 10.1038/nprot.2016.095

39. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* (2013) 41:e166. doi: 10.1093/nar/gkt646

40. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res.* (2014) 42:D222–30. doi: 10.1093/nar/gkt1223

41. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. 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

42. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol.* (2015) 33:243–246. doi: 10.1038/nbt.3172

43. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* (2010) 11:R14. doi: 10.1186/gb-2010-11-2-r14

44. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* (2019) 47:W199–205. doi: 10.1093/nar/gkz401

45. Indriastuti R, Pardede BP, Gunawan A, Ulum MF, Arifiantini RI, Purwantara B. Sperm transcriptome analysis accurately reveals male fertility potential in livestock. *Animals.* (2022) 12:2955. doi: 10.3390/ani12212955

46. Selvaraju S, Parthipan S, Somashekar L, Kolte AP, Krishnan Binsila B, et al. Occurrence and functional significance of the transcriptome in bovine (*Bos taurus*.) spermatozoa. *Sci Rep*. (2017) 7:2392. doi: 10.1038/srep42392

47. Yang H, Wang F, Li F, Ren C, Pang J, Wan Y, et al. Comprehensive analysis of long noncoding RNA and mRNA expression patterns in sheep testicular maturation. *Biol Reprod*. (2018) 99:650–61. doi: 10.1093/biolre/ioy088

48. Gòdia M, Estill M, Castelló A, Balasch S, Rodríguez-Gil JE, Krawetz SA, et al. RNA-seq analysis to describe the boar sperm transcriptome and its seasonal changes. *Front Genet*. (2019) 10:299. doi: 10.3389/fgene.2019.00299

49. Fraser L, Brym P, Pareek CS, Mogielnicka-Brzozowska M, Paukszto Ł, Jastrzebski JP, et al. Transcriptome analysis of boar spermatozoa with different freezability using RNA-Seq. *Theriogenology*. (2020) 142:400–13. doi: 10.1016/j.theriogenology.2019.11.001

50. Jastrzebski JP, Lipka A, Majewska M, Makowczenko KG, Paukszto L, Bukowska J, et al. *In silico* Identification of lncRNAs regulating sperm motility in the Turkey (*Meleagris gallopavo* L.). *Int J Mol Sci*. (2022) 23:7642. doi: 10.3390/ijms23147642

51. Sahoo B, Gupta MK. Transcriptome analysis reveals spermatogenesis-related CircRNAs and LncRNAs in goat spermatozoa. *Biochem Genet*. (2023) 1–23. doi: 10.1007/s10528-023-10520-8

52. Miller MR, Mannowetz N, Iavarone AT, Safavi R, Gracheva EO, Smith JF, et al. Unconventional endocannabinoid signaling governs sperm activation via the sex hormone progesterone. *Science*. (2016) 352:555–9. doi: 10.1126/science.aad6887

53. Kasimanickam V, Kasimanickam R, Arangasamy A, Saberivand A, Stevenson JS, Kastelic JP. Association between mRNA abundance of functional sperm function proteins and fertility of Holstein bulls. *Theriogenology*. (2012) 78:2007–19.e2. doi: 10.1016/j.theriogenology.2012.07.016

54. Schoff PK, Cheetham J, Lardy HA. Adenylate kinase activity in ejaculated bovine sperm flagella. *J Biol Chem*. (1989) 264:6086–91. doi: 10.1016/S0021-9258(18)83316-6

55. Cao W, Haig-Ladewig L, Gerton GL, Moss SB. Adenylate kinases 1 and 2 are part of the accessory structures in the mouse sperm flagellum. *Biol Reprod*. (2006) 75:492–500. doi: 10.1095/biolreprod.106.053512

56. Kawashima A, Osman BAH, Takashima M, Kikuchi A, Kohchi S, Satoh E, et al. CABS1 is a novel calcium-binding protein specifically expressed in elongate spermatids of mice. *Biol Reprod*. (2009) 80:1293–304. doi: 10.1095/biolreprod.108.073866

57. Zhang X, Zhou W, Zhang P, Gao F, Zhao X, Shum WW, et al. Cabs1 maintains structural integrity of mouse sperm flagella during epididymal transit of sperm. *Int J Mol Sci*. (2021) 22:652. doi: 10.3390/ijms22020652

58. Fiedler SE, Carr DW. ROPN1, a Protein kinase A-like (R2D2) protein, regulates sperm motility. *Biol Reprod*. (2011) 85:804. doi: 10.1093/biolreprod/85.s1.804

59. Fiedler SE, Dudiki T, Vijayaraghavan S, Carr DW. Loss of R2D2 proteins ROPN1 and ROPN1L causes defects in murine sperm motility, phosphorylation, and fibrous sheath integrity. *Biol Reprod*. (2013) 41:88. doi: 10.1095/biolreprod.112.105262

60. Choi I, Woo JM, Hong S, Jung YK, Kim DH, Cho C. Identification and characterization of ADAM32 with testis-predominant gene expression. *Gene*. (2003) 304:151–62. doi: 10.1016/S0378-1119(02)01202-7

61. Civetta A. Positive selection within sperm-egg adhesion domains of fertilinn ADAM gene with a potential role in fertilization. *Mol Biol Evol*. (2003) 20:21–9. doi: 10.1093/molbev/msg002

62. Netherton JK, Hetherington L, Ogle RA, Velkov T, Baker MA. Proteomic analysis of good-and poor-quality human sperm demonstrates that several proteins are routinely aberrantly regulated. *Biol Reprod*. (2018) 99:395–408. doi: 10.1093/biolre/iox166

63. Lee S, Hong SH, Cho C. Normal fertility in male mice lacking ADAM32 with testis-specific expression. *Reprod Biol*. (2020) 20:589–94. doi: 10.1016/j.repbio.2020.09.001

64. Park YJ, Pang WK Ryu DY, Song WH, Rahman MS, Pang MG. Optimized combination of multiple biomarkers to improve diagnostic accuracy in male fertility. *Theriogenology*. (2019) 139:106–12. doi: 10.1016/j.theriogenology.2019.07.029

65. Li H, He Y, Yan J, Zhao Q, Di C, Zhang H. Comparative proteomics reveals the underlying toxicological mechanism of low sperm motility induced by iron ion radiation in mice. *Reprod Toxicol*. (2016) 65:148–58. doi: 10.1016/j.reprotox.2016.07.014

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 August 2024
DOI 10.3389/fgene.2024.1368710



OPEN ACCESS

EDITED BY
Nuno Carolino,
National Institute for Agricultural and Veterinary Research (INIAV), Portugal

REVIEWED BY
Mohammad Hossein Banabazi,
Swedish University of Agricultural Sciences, Sweden
Diercles Francisco Cardoso,
Consultant, Brazil

*CORRESPONDENCE
Cedric Gondro,
gondroce@msu.edu

RECEIVED 11 January 2024
ACCEPTED 09 July 2024
PUBLISHED 02 August 2024

CITATION
Nawaz MY, Savegnago RP, Lim D, Lee SH and Gondro C (2024), Signatures of selection in Angus and Hanwoo beef cattle using imputed whole genome sequence data.
*Front. Genet.* 15:1368710.
doi: 10.3389/fgene.2024.1368710



# Signatures of selection in Angus and Hanwoo beef cattle using imputed whole genome sequence data

Muhammad Yasir Nawaz[1,2], Rodrigo Pelicioni Savegnago[1], Dajeong Lim[3], Seung Hwan Lee[3] and Cedric Gondro[1,2]*

[1]Department of Animal Science, Michigan State University, East Lansing, MI, United States, [2]Genetics and Genome Sciences Graduate Program, Michigan State University, East Lansing, MI, United States, [3]Division of Animal and Dairy Science, Chungnam National University, Daejeon, Republic of Korea

In this study, we detected signatures of selection in Hanwoo and Angus beef cattle using allele frequency and haplotype-based methods based on imputed whole genome sequence variants. Our dataset included 13,202 Angus animals with 10,057,633 imputed SNPs and 10,437 Hanwoo animals with 13,241,550 imputed SNPs. The dataset was subset down to 6,873,624 SNPs in common between the two populations to identify within population (runs of homozygosity, extended haplotype homozygosity) and between population signals of selection (allele fixation index, extended haplotype homozygosity). Assuming these selection signals were complementary to each other, they were combined into a decorrelated composite of multiple signals to identify regions under selection for each of the breeds. 27 genomic regions spanning 25.15 Mb and harboring 360 genes were identified in Angus on chromosomes 1,3, 4, 5, 6, 7, 8, 12, 13, 14, 16, 20, 21 and 28. Similarly, in Hanwoo, 59 genes and 17 genomic regions spanning 5.21 Mb on chromosomes 2, 4, 5, 6, 7, 8, 9, 10, 13, 17, 20 and 24 were identified. Apart from a small region on chromosome 13, there was no major overlap of selection signals between the two breeds reflecting their largely different selection histories, environmental challenges, breeding objectives and breed characteristics. Positional candidate genes identified in selected genomic regions in Angus have been previously associated with growth, immunity, reproductive development, feed efficiency and adaptation to environment while the candidate genes identified in Hanwoo included important genes regulating meat quality, fat deposition, cholesterol metabolism, lipid synthesis, neuronal development, and olfactory reception.

KEYWORDS

**signatures of selection, Hanwoo, Angus, WGS, beef cattle**

## 1 Introduction

Natural selection is an adaptive response to the environment a population inhabits, which drives its evolutionary changes by favoring traits that are advantageous and increases their prevalence in the population. Very recently, at least on an evolutionary scale, human driven artificial selection has also become a primary driver of changes in populations by exerting selective pressure on traits of human interest. A prime example of artificial selection is dog breeding: dogs have been bred for various desireable characteristics

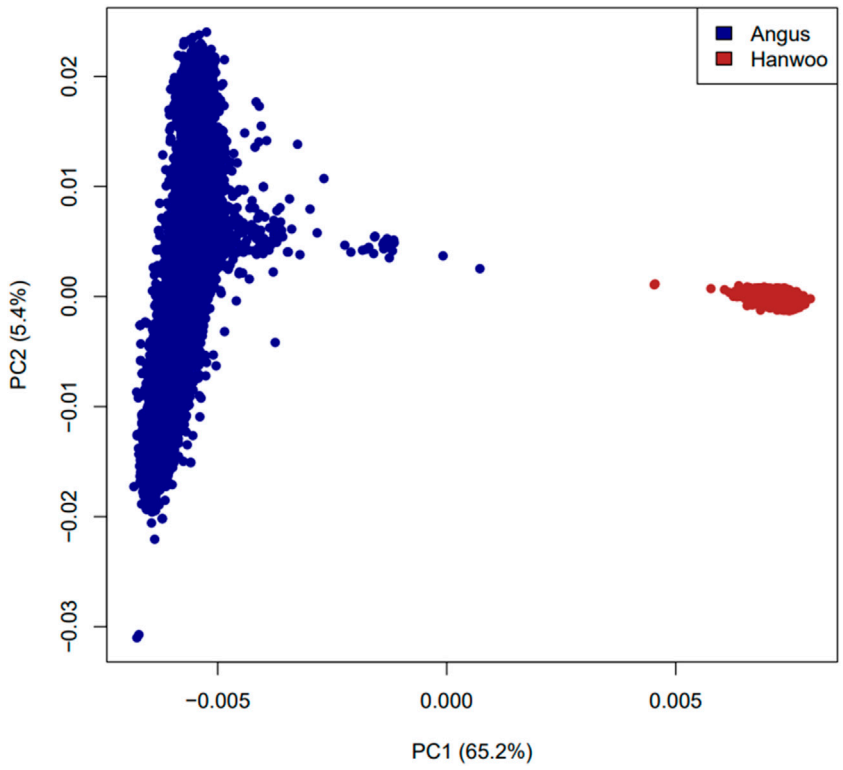



FIGURE 1
Plot of first two principal components based on a relationship matrix constructed from 6,873,624 SNPs common between Angus and Hanwoo.

which led to a wide variety of breeds from the tiny Chihuahua to the massive Great Dane. Such selection processes change allele frequencies in populations and leave traceable marks across the genome. Genomic regions under selective pressure can be identified by their allele frequency distributions, measures of linkage disequilibrium between loci and the structure of their haplotypes. Identification of these genetic patterns or *signatures of selection* (SOS) help us understand the underlying biological processes of adaptation in different environments and provide insights into the domestication history of agricultural species. They can also help us identify genes or genomic regions that regulate the phenotypic expression of traits of economic importance. For example, studies of signatures of selection have been used to identify genes that regulate coat color and body size in dogs (Pollinger et al., 2005; Sutter et al., 2007), stature in horses (Makvandi-Nejad et al., 2012), and body temperature maintenance under cold stress in cattle (Igoshin et al., 2019). Randhawa et al. (2016) published a meta-assembly of selection signatures in cattle genome by combining results from various studies. They found that a number of selection hotspots have been identified in European cattle but studies on major cattle groups like Zebu, African and Composite cattle have been few. They also observed that most of the selection signals were unique for each breeds while some were shared across breeds. The most prominent peaks were observed in genes of known major effects like coat color, polled locus and muscle hypertrophy.

Various methods have been proposed to identify genomic signatures of selection which can be broadly classified into two main categories: within population measures for a single population (e.g., runs of homozygosity and integrated haplotype score) or between population measures that compare two or more populations (e.g., fixation index and cross-population extended haplotype homozygosity). Each of these test statistics explore unique facets of the genomic architecture of populations but they are not necessarily consistent with each other. Inconsistencies between selection sweeps are observed not only due to the inherent differences in statistical methodologies but also due to differential sensitivity to sampling, demographic history and linkage disequilibrium between loci (Ewing and Jensen, 2016). Therefore, some studies take a more conservative approach and only focus on the regions that are common across different measures, albeit at the risk of not identifying a proportion of the relevant signals in the process. An alternative approach is to consider the selection signals from different methods as complimentary to each other (Ma et al., 2015) and combine them to get a composite score (Randhawa et al., 2014). Various methods to combine individual signals have already been proposed in the literature (Grossman et al., 2010; Utsunomiya et al., 2013b; Randhawa et al., 2014; Ma et al., 2015). Initial approaches to combine the signals did not account for the covariance structure between signals but Ma et al. (2015) suggested a new approach to calculate a decorrelated composite of multiple signals (DCMS) that adjusted for correlations between signals and was more powerful to detect selected regions in the genome.

This study focused on the identification of signatures of selection in Angus and Hanwoo cattle. Both are beef cattle breeds, but they have been subjected to entirely different selection pressures and have

TABLE 1 Summary of results from runs of homozygosity analysis for Hanwoo and Angus cattle.

| Parameter | Hanwoo | Angus |
|---|---|---|
| Total SNPs | 13241550 | 10057633 |
| Total animals | 10437 | 13202 |
| Percentage of animals having ROH | 99.7 | 99.9 |
| Total number of ROH regions | 129778 | 1169509 |
| Mean number of ROH per animal | 12.5 | 88.7 |
| SD of number of ROH per animal | 8.4 | 18.501 |
| Minimum number of ROH per animal | 1 | 1 |
| Maximum number of ROH per animal | 440 | 270 |
| Mean length of ROH regions in KB | 3024 | 2381.845 |
| SD of length of ROH regions | 5089.125 | 2965.864 |
| Median length of ROH regions | 1384 | 1565 |
| Maximum length of ROH | 123720 | 120023 |
| Minimum length of ROH | 1000 | 1000 |
| No of ROH per animal | 12.46 | 88.67 |
| ROH 1–5 mb | 113515 (87.5%) | 1087433 (92.9%) |
| ROH 5–10 mb | 8369 (6.5%) | 56661 (4.9%) |
| ROH 10–15 mb | 3616 (2.8%) | 13443 (1.1%) |
| ROH 15–20 mb | 1718 (1.3%) | 5491 (0.4%) |
| ROH 20–25 mb | 1017 (0.8%) | 2828 (0.2%) |
| ROH 25–30 mb | 614 (0.5%) | 1527 (0.1%) |
| ROH >30 mb | 929 (0.7%) | 2126 (0.2%) |

different genetic population structures, body characteristics, domestication history, beef quality and breeding program objectives. Hanwoo are Korean taurine cattle, more related to Asian taurine cattle like Japanese Wagyu than to western taurine cattle breeds (Angus, Hereford, etc.) (Lee et al., 2014). Hanwoo have a smaller stature than Angus, but its beef is popular for its juiciness, high levels of marbling, and unique flavor (Cho et al., 2010); which, similarly to Wagyu, attracts a market premium. Hanwoo was historically a draft breed kept by small holder farmers which accounted for more than 99% of the farms in Korea until 1985. Hanwoo steers are typically kept up to 30–32 months of age to improve the marbling score. In 1960s, various breed improvement initiatives were taken in Korea. The recent advances in management of beef production have also led to an increase in the size of beef operations in Korea. Currently, the selection index of the *Korean Proven Bulls* program is mainly driven by 4 traits–marbling score (MS), carcass weight (CWT), eye muscle area (EMA) and back fat thickness (BF). Consequently, Hanwoo have shown considerable improvement in beef quality. Angus, on the other hand, are European taurine cattle that originated from Scotland. Angus have been intensively selected for growth, stature and feed intake in the 20th century and have become the most common beef cattle in the world. Angus are characterized by their high muscularity, fast growth rate, medium height, and moderate levels of intramuscular fat (Albertí et al., 2008). In contrast to Hanwoo, different selection indices are used in Angus cattle breeding programs worldwide depending on the type of beef production operation and its breeding objectives. The average age at slaughter varies between 12 and 20 months depending on whether the calves are weaned and sent directly to a feeding facility to be finished for slaughter or they are grown on grass pastures at first, followed by a high-energy diet for a short period of time (100–120 days) before slaughter. Therefore, due to stark differences in evolutionary origin, artificial selection, farming systems, and body characteristics, differences in genomic landscape between them may point to genetic basis of adaptive traits and meat production.

The objectives of this study were to identify genome wide signals of selection in Angus and Hanwoo beef cattle using imputed whole genome sequence (WGS) data. We used imputed whole genome sequence data for this analysis to get a higher resolution of selected genomic regions. We also combined individual selection measures to obtain a decorrelated composite of multiple signals (DCMS) for identification of selected genomic regions. These signatures of selection were then mapped to the ARS-UCD1.2 reference assembly to identify candidate genes located in these regions. We also highlight important genes related to meat production and quality.

# 2 Materials and methods

## 2.1 Genotype data

Imputed whole genome genotypes of 10,437 Hanwoo animals (13,241,550 SNPs) and 13,202 Angus animals (10,057,633 SNPs) were utilized for this analysis. Respectively, the Hanwoo and Angus data consisted of 9,160 and 11,632 animals genotyped on 50k arrays (Illumina Bovine SNP50 BeadChip; Illumina, San Diego, CA), 1,704 and 1,236 animals genotyped on 700k arrays (777k SNP, Illumina Bovine HD Beadchip, Illumina, San Diego, CA), and 203 and 334 reference animals with whole genome sequence (WGS) data. All Hanwoo animals originated from commercial farms in Korea while Angus data was collected from commercial farms primarily in the US. Sequence analysis was performed using integrated variant discovery pipeline (https://github.com/rodrigopsav/IVDP) to call variants. The key steps in the pipeline include read trimming and adapter removal by trimmomatic, read alignment to ARS-UCD1.2 *Bos taurus* assembly using bwa-mem2, duplicated read marking by sambamba-markdup, base quality recalibration using GATK BaseRecalibratorSpark and ApplyBQSRSpark and variant calling using GATK HaplotypeCaller. Sequenced animals were used a reference to impute genotype data of their respective breeds. Eagle software version 2.3.2 and Minimac3 was used for phasing and imputation respectively. Details on quality control, WGS pipeline and imputation accuracies for Hanwoo were previously reported in Nawaz et al., 2022. Finally, Imputed whole genome data was subset down to the *6,873,624* SNPs that were common between the two breeds to calculate across population measures of selection.

## 2.2 Analysis

We performed principal component analysis on the combined dataset containing all Angus and Hanwoo animals using plink 1.9 to evaluate population structure in the data. Various selection signals were calculated as explained below.

### 2.2.1 Within population measures

Runs of homozygosity (ROH) are defined as long continuous homozygous genomic regions that are assumed to be identical DNA segments inherited by descent from a common ancestor, and that serve as an indicator of genomic autozygosity, consanguinity, selection, and population size reduction. ROH detection was done using the *homozyg* function in plink using the default parameters except for the number of SNPs in a scanning window (*homozyg-window-snp*) which was increased to 100 instead of default 50 SNPs because of the high density of SNPs in the sequence data. Default values were used for all the other required parameters in *homozyg* function.

To identify ROH islands, we calculated the autozygosity of each SNP by taking the proportion of individuals in which a SNP was identified within a ROH region.

Integrated haplotype score (iHS) aims to identify genomic regions that were under recent positive selection based on the relationship between an allele's frequency and the extent of linkage disequilibrium around it. iHS was calculated (Voight et al., 2006) based on extended haplotype homozygosity (EHH) values (Sabeti et al., 2002) calculated using the program *hapbin* (Maclean et al., 2015). Due to the high dimensionality of our data and computational limitations of the software, the analysis was performed by dividing both Hanwoo and Angus datasets into seven and 14 bins containing 1491 and 943 animals per bin, respectively. The correlation of iHS between sample bins ranged from 0.86 to 0.93. Final values of iHS were calculated by taking the average of iHS values from the data bins. Absolute values of iHS were smoothed out in windows of 1,001 SNPs to identify regions under recent positive selection.

### 2.2.2 Across population measures

Fixation Index ($F_{ST}$) is a measure of population differentiation. It represents the proportion of total genetic variance that exists within a sub population. Allele frequencies of Angus and Hanwoo datasets were calculated using *freq* function in *plink*. Average of Angus and Hanwoo allele frequencies were used as the baseline allele frequency ($p$) and genetic variances ($p^*(1-p)$). Finally, $F_{ST}$ was calculated for each SNP by taking squared deviation of allele frequency in a breed from the baseline frequency divided by allelic variance (Weir and Cockerham, 1984):

$$F_{ST} = \frac{\sigma^2}{p(1-p)}$$

To identify prominent genomic regions, $F_{ST}$ was smoothed in windows of 1,001 SNPs using *runmed* function in base R.

Across population extended haplotype homozygosity (XPEHH) (Sabeti et al., 2007) is another population differentiation-based test that is used to detect selective sweeps in which selected regions are close to fixation in one population but remain polymorphic in another population. For XPEHH, we compared the two breeds under study (Angus and Hanwoo) directly against each other to identify regions that were differentially selected between populations. We used the *hapbin* software (Maclean et al., 2015) to perform this analysis with the *xpehh* function.

### 2.2.3 Decorrelated composite of multiple signals (DCMS)

In order to combine the several test statistics, we used the method suggested by Ma et al. (Ma et al., 2015) that takes into account correlations between signals to calculate a decorrelated composite of multiple signals (DCMS) based on their $p$ values. Firstly, fractional ranks of autozyosity and absolute values of ROH, iHS, FST and XPEHH were used to calculate their $p$ values using *stat_to_p-value* function in R package *MINOTAUR* (with parameters *two.tailed = FALSE, right.tailed = TRUE*). Then, a pairwise correlation matrix was created between absolute values of the signals. This matrix was used as an input to DCMS function in MINOTAUR to calculate raw DCMS scores as follows (Ma et al., 2015):

$$DCMS = \sum_{t=1}^{n} \frac{\left(\log \frac{1-p_{lt}}{p_{lt}}\right)}{\sum_{i=1}^{n} |r_{it}|}$$

$p_{lt}$ was the $p$-value of individual selection measures and $r_{it}$ was the Pearson correlation between two measures. Finally, $p$ values of raw *DCMS* scores were calculated by *pnorm* function using empirical mean and standard deviation. Multiple test correction was done by calculating false positive rate (*FDR*) using *p.adjust* function in base R with *method= 'BH'*. SNPs having adjusted $p$-values ($q$) less than 0.05 were deemed to be significant. Adjacent significant SNPs (located less than 1 MB apart) were combined to identify regions under selection by a custom script in R.

### 2.2.4 Functional annotation of signatures of selection

A *Bos taurus* gene annotation dataset which included positional information for all known bovine genes (n = 27,900) mapped to the latest bovine assembly (*ARS_UCD1.2*) was downloaded from ensemble with *BIOMART*. Significant genomic regions were mapped to genes using the *GenomicRanges* package in R (Lawrence et al., 2013).

# 3 Results

The observed heterozygosity in Angus and Hanwoo cattle was 0.30 and 0.31 respectively. Principal component analysis revealed that Angus and Hanwoo animals clearly clustered separately from each other in tight clusters (Figure 1). The first principal component separated the two populations and accounted for 65.1% of genomic variation in the dataset. The second principal component captured variation in Angus animals which accounted for only 5.4% of the total genomic variation in the dataset. These results indicate that majority of the genomic variation in the dataset can be explained by the differences in genomic architecture of the two breeds.

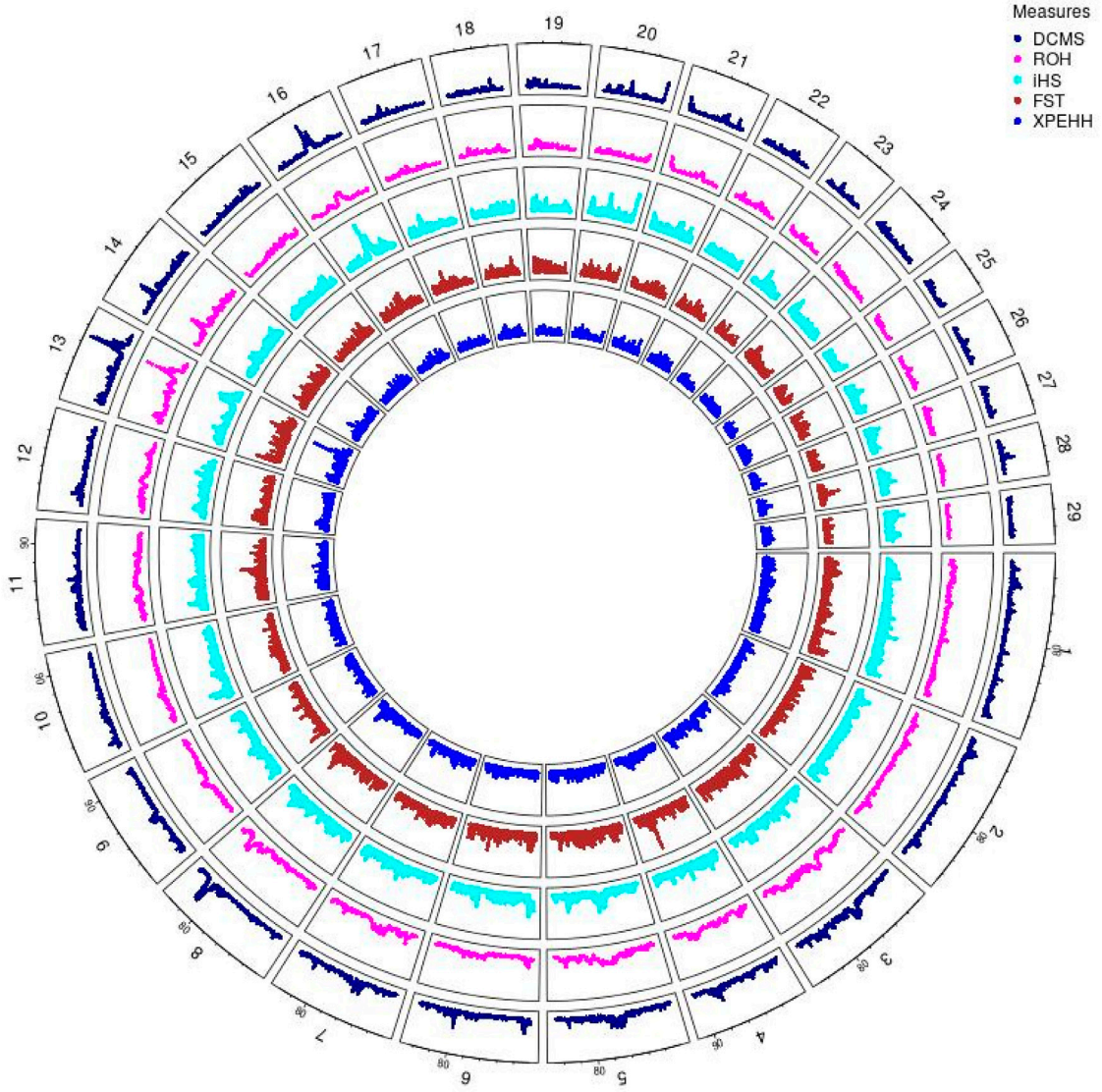



FIGURE 2
Circos plot of *p*-values for genome wide signatures of selection in Angus cattle.

## 3.1 Within population measures

ROH: The mean number of ROH detected per animal was higher in Angus (88.7 ± 18.50) as compared to Hanwoo (12.5 ± 8.4) (Table 1). The median length of ROH regions was also higher in Angus (1,565 BP) as compared to Hanwoo (1,384 BP). However, the proportion of ROH regions longer than 5 MB was higher in Hanwoo (12.5%) than Angus (7.1%). Therefore, Hanwoo had fewer ROH regions, but they were longer than in Angus suggesting a comparatively more recent selection in Hanwoo. Mean genome wide autozygosity was higher in Angus (0.08) as compared to Hanwoo (0.01). The highest peak for Hanwoo was observed on CHR 7 (BP 50280340) and smaller peaks were observed on CHR 2, 12, 23, 24 and 29. In Angus, the strongest signal was observed on CHR 13 (BP 63,854,457). Other significant peaks were also identified on CHR 8 and 14.

iHS: Genome wide distribution of absolute iHS values was similar in Hanwoo and Angus with a mean of 0.31 and 0.30 respectively. Absolute value of iHS indicated genomic regions with unusually long haplotypes on chromosomes 1, 5, 6, 8, 10, 11, 13, 16, 17, 20, 23, 24 and 29 in Angus harboring 13,009 significant SNPs. The strongest signal was detected on CHR 16 (rs208273139) at 40,588,657 BP. In Hanwoo, the strongest signal was observed on chromosome 2 at (rs207720085) 82,874,034 BP. Other peaks were observed on chromosomes 1, 2, 3, 5, 6, 7, 8, 9, 14, 17, 20, 25 and 26 harboring 13,030 significant SNPs. Correlation between iHS values of Angus and Hanwoo was 0.016 indicating differences in the regions of selection sweeps between the two breeds.

We also observed that ROH and iHS were significantly correlated (R = 0.252, 95% confidence interval 0.251–0.253) in Hanwoo and Angus (R = 0.286, 95% interval 0.286–0.287).

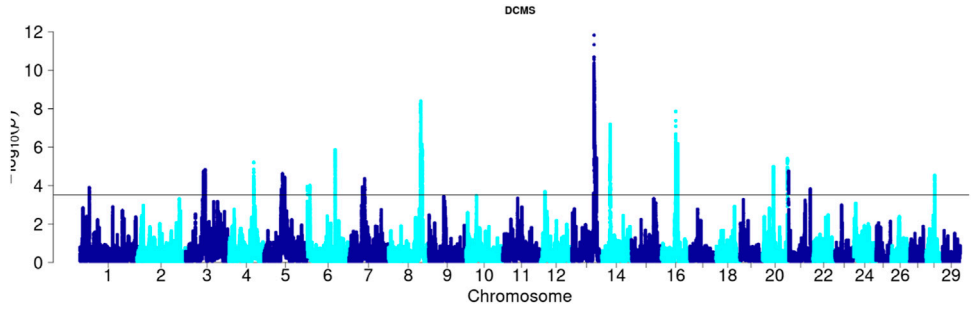



FIGURE 3
Manhattan plot of DCMS *p*-values in Angus cattle. Horizontal black line indicates the significance cut off (0.05 FDR).

## 3.2 Across population measures

Fixation index $F_{ST}$: SNPs with an FST value in the top 0.2% were identified on 18 out of 29 autosomes indicating widespread allele frequency differences between breeds. CHR 4 contained the highest number of significant SNPs (n = 1,577) followed by CHR 8 (n = 1,464) and CHR 5 (n = 1,256). The most significant SNP (rs209900249) was observed on CHR 4 position 69,682,473. Other prominent FST hotspots were observed on CHR 1, 2, 3, 6, 7, 9, 10, 13, 14, 16, 18, 20, 21, 28, and 29.

Across population extended haplotype homozygosity (XPEHH): 13,004 SNPs with top 0.2% XPEHH values were located on CHR 3 (n = 2,216), 8 (n = 4,826), 13 (n = 4,115) and 14 (n = 1,847). The most significant peak was observed on CHR 13 at position 62,594,885 (rs207508467).

We also observed that the two measures of across population measures were significantly correlated, Pearson correlation R = 0.2956 and a 95% confidence interval 0.295–0.296.

## 3.3 Decorrelated composite of multiple signals (DCMS)

Angus: A total of 39,898 SNPs were identified with significant *p*-values. Genic SNPs accounted for 27.49% of all the significant SNPs. 27 significant genomic regions were identified using the DCMS adjusted *p*-value (q value) cutoff of 0.05. The mean length of selected regions was 931.613 Kb (±1,255.33) while their total length was 25.153 Mb. The significant genic regions mapped to CHR 1,3, 4, 5, 6, 7, 8, 12, 13, 14, 16, 20, 21, and 28 (Figures 2, 3) that harbor 360 genes (Table 2). The most significant genomic selection signal was observed on CHR 13 where 91 genes were found spread across 3 distinct regions.

Some of the notable genes identified in significant genomic regions were associated with body size and stature (PLAG1, CHCHD7, RPS20, LYN), growth and feed intake (TMEM68, TGS1, LYN, XKR4), growth differentiation factor (GDF5), feed efficiency (OR6C76, PIK3CD), embryonic growth and reproductive development (NMNAT1), immunity related to tropical adaptation (SLC25A33, SPSB1), immune response and immune regulation (PIK3CD), pigmentation and adaptation to environment (ASIP). A complete list of all the regions and genes identified is shown in Table 2.

Hanwoo: A total of 10,162 SNPs were found in significant hotspots of selection using FDR cut off value of 0.05 on adjusted DCMS *p* values (q value). Out of these only 2,095 (20.6%) SNPs were located in genes. Significant SNPs were used to identify 17 significant genomic regions. The mean length of the selected regions was 306.27 kb (± 337.43) while their total length was 5.21 Mb. Significant genomic regions mapped to CHR 2, 4, 5, 6, 7, 8, 9, 10, 13, 17, 20, and 24 (Figures 4, 5) which harbor 59 genes (Table 3).

The most significant genomic region was on CHR 2 between BP 81860076 and 82963443 BP where only 1 gene was identified (ENSBTAG00000048361). The greatest number of SNPs mapped to a gene on CHR 17 that plays important role in immunity (LRBA). An important region on CHR *24 (BP 43384983–44317964)* was identified that contained genes (*e.g., MC2R*) regulating fat deposition and meat quality. Other genes identified were previously associated with important roles in brain development (CPLANE1), developmental regulation (NIPBL), breakdown of amino acids (BCKDHB), olfactory reception (OR6F1). A complete list of all the regions and genes identified has been provided in Table 3. Interestingly, none of the significantly selected regions were common between Hanwoo and Angus.

# 4 Discussion

The main aim of this study was to identify genomic regions under selective pressure in Angus and Hanwoo cattle utilizing imputed whole genome information. We first identified individual selection signals by four distinct methods primarily based on allele frequency and haplotype patterns. We combined individual signals to identify strong signals of selection. Finally, we identified various positional candidate genes related to beef production and quality. Overall, we observed more genomic regions and genes under selective pressure in Angus than in Hanwoo with a limited overlap of selected regions or genes between the breeds, which is consistent with large differences in breed origin, environmental habitats, divergent selection histories, breeding program objectives and ultimately, the phenotypic differences between the breeds.

TABLE 2 Genomic regions under selection in Angus cattle identified by DCMS q values ≤ 0.05 and genes identified in those regions.

| Start | End | CHR | No of genes | Genes |
|---|---|---|---|---|
| 26683714 | 26745320 | 1 | 0 | None |
| 50460996 | 53793201 | 3 | 35 | MTF2 DIPK1A RPL5 SNORA66 SNORD21 U6 EVI5 ENSBTAG00000055274<br>5S_rRNA GFI1 RPAP2 GLMN C3H1orf146 BTBD8 ENSBTAG00000040248<br>EPHX4 BRDT ENSBTAG00000054884 ENSBTAG00000047443 TGFBR3<br>CDC7 HFM1 ENSBTAG00000054082 ENSBTAG00000046077 ENSBTAG00000055150<br>ZNF644 bta-mir-2285b-2 BARHL2 ZNF326 LRRC8D<br>bta-mir-2285k-5 ENSBTAG00000050182 LRRC8C LRRC8B ENSBTAG00000038625 |
| 55238880 | 55468638 | 3 | 3 | PKN2 ENSBTAG00000051499 ENSBTAG00000051844 |
| 69475528 | 69894433 | 4 | 6 | 7SK SNX10 CBX3 HNRNPA2B1 NFE2L3 MIR148A |
| 48288996 | 48752060 | 5 | 4 | MSRB3 LEMD3 WIF1 U6 |
| 52011261 | 52117527 | 5 | 1 | TAFA2 |
| 53692271 | 54424033 | 5 | 3 | SLC16A7 ENSBTAG00000055198 ENSBTAG00000053531 |
| 58055499 | 59307814 | 5 | 42 | U6 ENSBTAG00000047825 ENSBTAG00000052093<br>ENSBTAG00000049329 ENSBTAG00000051156<br>ENSBTAG00000046778 ENSBTAG00000048295<br>ENSBTAG00000054507 ENSBTAG00000050480<br>ENSBTAG00000051165 ENSBTAG00000051462<br>ENSBTAG00000049219 ENSBTAG00000051274<br>ENSBTAG00000048779 OR6C76 OR6C75 ENSBTAG00000049581<br>ENSBTAG00000049184 ENSBTAG00000048408<br>ENSBTAG00000024691 ENSBTAG00000051265<br>ENSBTAG00000050381 ENSBTAG00000049213<br>ENSBTAG00000054097 ENSBTAG00000049016<br>ENSBTAG00000045922 ENSBTAG00000048168<br>ENSBTAG00000053702 ENSBTAG00000054733 ENSBTAG00000049913 ENSBTAG00000051198<br>ENSBTAG00000002913 ENSBTAG00000051990 ENSBTAG00000048864 ENSBTAG00000046446<br>ENSBTAG00000049753 ENSBTAG00000054193 OR10A7 ENSBTAG00000049751<br>ENSBTAG00000053229 ENSBTAG00000053772 ENSBTAG00000037629 |
| 1155763 | 1320935 | 6 | 1 | ENSBTAG00000051456 |
| 8724140 | 8824920 | 6 | 0 | None |
| 78535547 | 78939028 | 6 | 0 | None |
| 38009575 | 38227430 | 7 | 8 | FAF2 RNF44 CDHR2 GPRIN1 SNCB EIF4E1B TSPAN17 UNC5A |
| 44026848 | 44466715 | 7 | 22 | ENSBTAG00000012150 MEX3D MBD3 UQCR11 TCF3 ONECUT3 ATP8B3 REXO1<br>KLF16 ABHD17A ENSBTAG00000050118 SCAMP4 CSNK1G2 bta-mir-6120 BTBD2<br>SOWAHA SHROOM1 GDF9 UQCRQ LEAP2 AFF4 U6 |
| 89565309 | 94976398 | 8 | 42 | 5S_rRNA ENSBTAG00000052296 NXNL2 SPIN1 ENSBTAG00000051928 ENSBTAG00000054632<br>CDK20 FBXW12 ENSBTAG00000021235 MSANTD3 TMEFF1 CAVIN4 PLPPR1 5S_rRNA ENSBTAG00000025760<br>MRPL50 ZNF189 ALDOB PGAP4 RNF20 GRIN3A ENSBTAG00000050971 ENSBTAG00000030953 |

(Continued on following page)





10.3389/fgene.2024.1368710

TABLE 2 (*Continued*) Genomic regions under selection in Angus cattle identified by DCMS q values ≤ 0.05 and genes identified in those regions.

| Start | End | CHR | No of genes | Genes |
|---|---|---|---|---|
| | | | | CYLC2 U6 SMC2 ENSBTAG00000047350 ENSBTAG00000013445 ENSBTAG00000050829 ENSBTAG00000052864 ENSBTAG00000053491 ENSBTAG00000016173 ENSBTAG00000050464 ENSBTAG00000000145 ENSBTAG00000052019 OR13C3 ENSBTAG00000049256 OR13C8 ENSBTAG00000048409 NIPSNAP3A ABCA1 SLC44A1 |
| 12701319 | 12716083 | 12 | 1 | TNFSF11 |
| 61100092 | 61130961 | 13 | 12 | ENSBTAG00000052743 DEFB121 DEFB122A DEFB122 DEFB123 DEFB124 REM1 HM13 bta-mir-12010 ID1 COX4I2 BCL2L1 |
| 62482399 | 65519468 | 13 | 71 | BPIFB4 BPIFA2A ENSBTAG00000031375 BPIFA2C ENSBTAG00000011704 BPIFA2B ENSBTAG00000031373 BPIFA3 BPIFA1 BPIFB1 BPIFB5 CDK5RAP1 ENSBTAG00000031354 SNTA1 ENSBTAG00000010131 ENSBTAG00000053051 ENSBTAG00000053797 NECAB3 C13H20orf144 E2F1 PXMP4 ZNF341 CHMP4B RALY EIF2S2 ASIP AHCY ENSBTAG00000050108 ENSBTAG00000046623 ITCH DYNLRB1 MAP1LC3A PIGU TP53INP2 NCOA6 GGT7 ACSS2 GSS MYH7B bta-mir-499 TRPC4AP EDEM2 PROCR MMP24 EIF6 FAM83C UQCC1 ENSBTAG00000053266 GDF5 ENSBTAG00000052250 CEP250 ENSBTAG00000030976 ERGIC3 ENSBTAG00000053187 SPAG4 CPNE1 RBM12 NFS1 ROMO1 RBM39 ENSBTAG00000053775 PHF20 5S_rRNA SCAND1 CNBD2 ENSBTAG00000052997 ENSBTAG00000053403 EPB41L1 ENSBTAG00000050801 AAR2 DLGAP4 |
| 67831405 | 69506639 | 13 | 8 | FAM83D ENSBTAG00000044690 DHX35 U6 ENSBTAG00000049087 ENSBTAG00000050378 ENSBTAG00000048871 MAFB |
| 22710076 | 24757731 | 14 | 24 | XKR4 TMEM68 TGS1 LYN RPS20 ENSBTAG00000045097 U1 MOS PLAG1 CHCHD7 ENSBTAG00000054153 SDR16C5 SDR16C6 PENK U6 BPNT2 FAM110B ENSBTAG00000047136 ENSBTAG00000051748 UBXN2B CYP7A1 U1 SDCBP NSMAF |
| 40378981 | 41205611 | 16 | 9 | TNFSF18 ENSBTAG00000052047 ENSBTAG00000053302 TNFSF4 ENSBTAG00000020550 AADACL4 DHRS3 VPS13D SNORA59A |
| 42527352 | 44175641 | 16 | 32 | MTOR ANGPTL7 EXOSC10 SRM MASP2 TARDBP CASZ1 PEX14 DFFA ENSBTAG00000045105 CORT CENPS PGD ENSBTAG00000048790 ENSBTAG00000048747 ENSBTAG00000054239 U6 UBE4B RBP7 NMNAT1 LZIC CTNNBIP1 CLSTN1 PIK3CD U6 5S_rRNA U6 TMEM201 SLC25A33 ENSBTAG00000049485 SPSB1 H6PD |
| 45628162 | 46172668 | 16 | 3 | ENSBTAG00000048839 ENSBTAG00000051176 CAMTA1 |
| 31142802 | 31450979 | 20 | 11 | ENSBTAG00000033187 NNT PAIP1 ENSBTAG00000049623 C20H5orf34 TMEM267 CCL28 HMGCS1 ENSBTAG00000048672 NIM1K ENSBTAG00000042376 |
| 69997413 | 70839780 | 20 | 8 | ENSBTAG00000050065 IRX2 U6 ENSBTAG00000054006 5S_rRNA IRX4 NDUFS6 ENSBTAG00000050317 |
| 2650859 | 3126131 | 21 | 2 | ATP10A U6 |
| 62909681 | 63080951 | 21 | 3 | 5S_rRNA ENSBTAG00000049199 ENSBTAG00000052737 |
| 25323298 | 25520626 | 28 | 9 | DDX21 KIFBP U6 SRGN ENSBTAG00000042264 ENSBTAG00000051145 VPS26A SUPV3L1 HKDC1 |





10.3389/fgene.2024.1368710

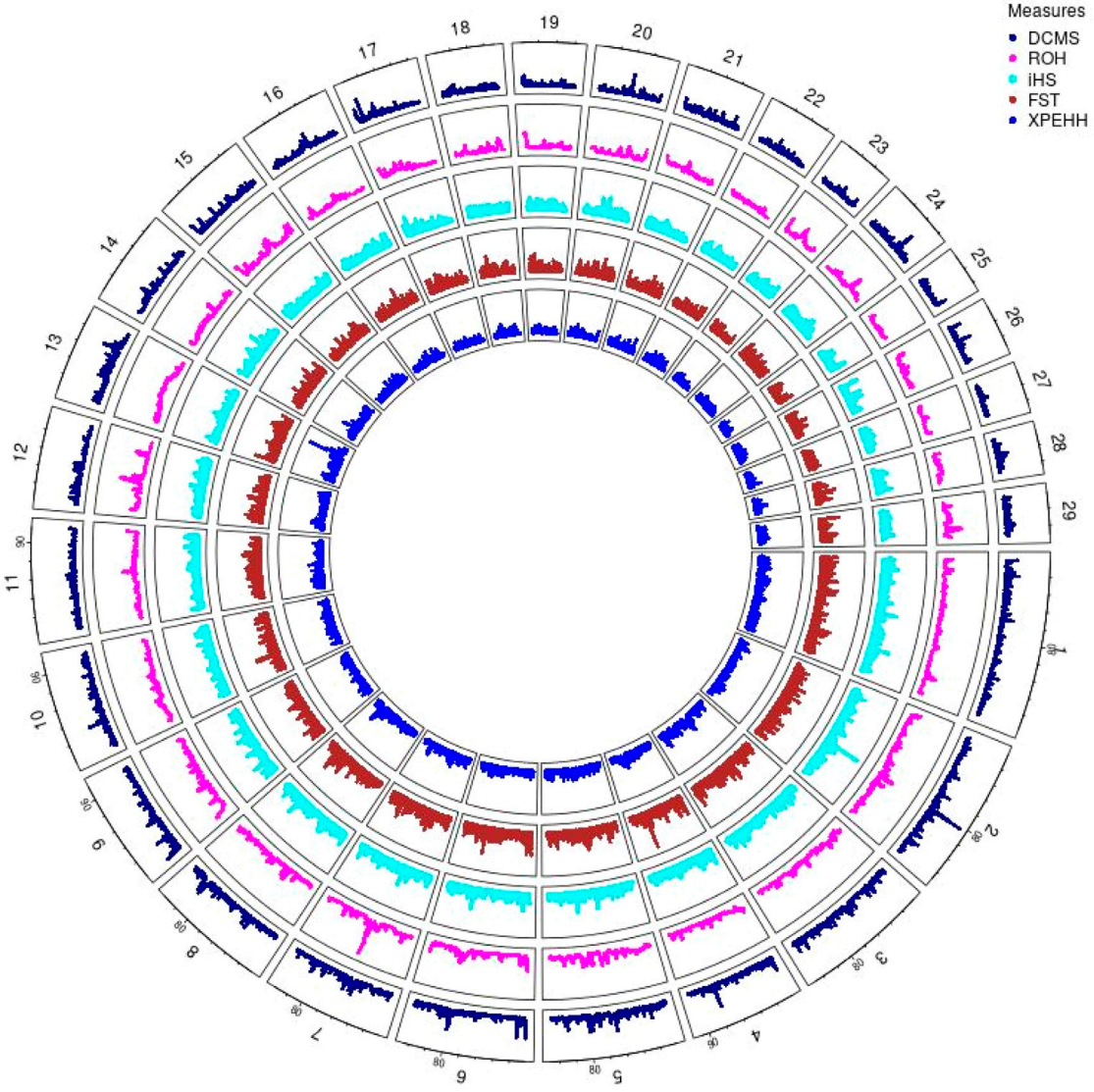



FIGURE 4
Circos plot of *p*-values for genome wide signatures of selection in Hanwoo cattle.

Genes identified within selected genomic regions in Angus included previously known regulators of growth, body size, feed intake, reproductive performance, and immunity. For example, *PLAG1* regulates cell proliferation and its association with carcass weight and stature has been reported in several cattle breeds (Utsunomiya et al., 2013a; Takasuga, 2016; Fink et al., 2017). Similarly, *LYN*, another regulator of cell proliferation and *RPS20*, a catalyst of protein synthesis, have been associated with body weight and preweaning daily gain in Nellore (Utsunomiya et al., 2013a; Fink et al., 2017). *CHCHD7* was previously reported as significantly associated with height in Jersey and Holstein (Utsunomiya et al., 2013a; Fink et al., 2017) and with carcass weight in Wagyu cattle (Nishimura et al., 2012). Both *PLAG1* and *RPS20* have also been associated with fetal growth and calving ease (Takasuga, 2016). Several olfactory receptors were also found in significant genomic regions (e.g., *OR6C76, OR6C75, OR10A7, OR13C3, OR13C8*). The olfactory transduction pathway has been associated with feed intake as it affects the perception of odor and in turn influences food preference and consumption (Abo-Ismail et al., 2010). Olfactory receptor loci have also been identified in other selective sweep studies in cattle and there are indications of recent duplication events (Ramey et al., 2013); which suggests that olfactory receptors may be under strong selection. *TMEM68* (a cyltransferase involved in glycerolipid metabolism) and *XKR4* have been associated with growth and feed intake in Nellore (Terakado et al., 2018). *XKR4* has also been associated with subcutaneous fat in indicine and composite cattle (Porto Neto et al., 2012). TGS1 (trimethylguanosine synthase 1) has pleitropic effects in growth traits and feed efficiency (Terakado et al., 2018; Ghoreishifar et al., 2020). GDF5 (growth differentiation factor) is critical for normal skeletal development. Loss of GDF5 function results in developmental delay and a shortened appendicular skeleton (Buxton et al., 2001). PIK3CD (a component of the phosphatidylinositol-3-kinase pathway) is involved in lymphocyte

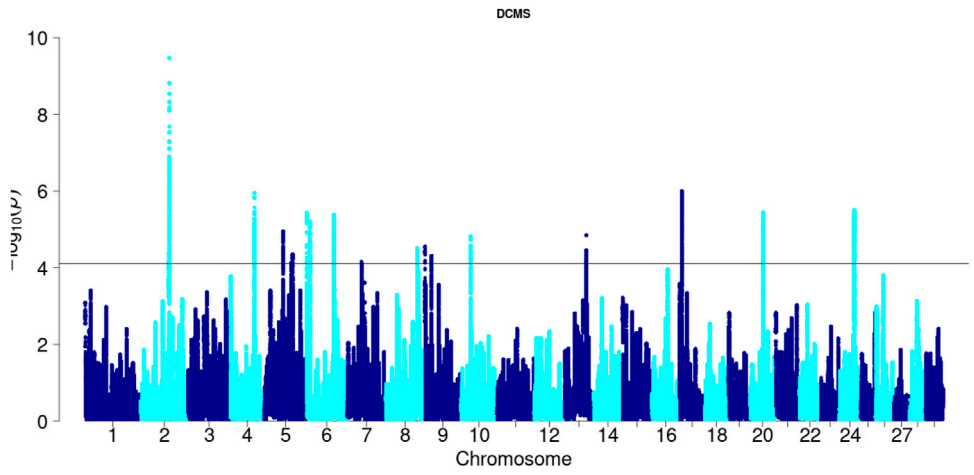



FIGURE 5
Manhattan plot of DCMS *p*-values in Hanwoo cattle. Horizontal black line indicates the significance cut off (0.05 FDR).

signaling. Mutations in PIK3CD causes immune dysregulation and disease pathogenesis (Tangye et al., 2019). SPSB1 (splA/ryanodine receptor domain and SOCS box containing 1) is an important component of mammalian innate immune system regulation that recognizes foreign molecules derived from pathogens (Lewis et al., 2011). We also identified solute carrier genes (SLC44A1, SLC16A7, SLC25A33) which belong to a major class of transport proteins in the cell membrane and play an important role in response to metabolic states and environmental conditions (Pizzagalli et al., 2021). Various solute carrier genes were also identified in another study directly comparing zebu and taurine cattle using differential allele frequency and haplotype diversity methods (Chan et al., 2010). This strongly suggests their role in adaptation to tropical environments. ASIP (Agouti signaling protein) is a well-known gene associated with coat color pigmentation and environmental adaptation in several species (Bertolini et al., 2018).

Considering the breed's innate characteristics and the high focus of the Hanwoo breeding program to select for increased marbling, it was reasonable to expect that some genomic regions under selection would be related to marbling score. An important region on CHR 24 was identified which contained ENSBTAG00000046153, MC2R and SETBP1 genes. The same region was also identified by composite signal in a multi breed study within a Hanwoo-specific signal (Gutiérrez-Gil et al., 2015). MC2R (adrencorticotropin receptor) and MC5R (melanocortin 5 receptor) genes belong to a family of melanocortin receptors (reviewed by Switonski et al. (2013)) that are involved in fatty acid and lipid metabolism pathways and reproduction. These genes have been previously located within a QTL region for marbling and backfat thickness and meat quality in pigs (Kováčik et al., 2012; Switonski et al., 2013). MC5R is a functional candidate for fatness in domestic animals and obesity in humans (Switonski et al., 2013) because it regulates interleukin 6 (IL6) (Jun et al., 2010) and downregulates leptin secretion (Hoggard et al., 2004) respectively resulting in increased fat deposition and increased feed intake. Based on these findings, we conclude that this selected region on CHR 24 is an important functional region for meat quality and should be further investigated in future studies in Hanwoo and/or other beef cattle.

Although it is common to focus on the genes identified in selected genomic regions, it should also be considered that much of the phenotypic diversity originates from differential regulation of gene expression by regulatory elements like promoters, enhancers, silencers, etc. (van Laere et al., 2003; Salinas et al., 2016). In this study 29.67% and 27.3% of the significant SNPs found in Angus and Hanwoo were annotated to gene coding regions, while the majority of the significant variants were located elsewhere. Similarly, Vernot et al. (2012) reported that the number of regulatory variants under selection far exceeded the number of variants in protein coding regions although their effect sizes may be small. Therefore, apart from the genes highlighted above, there may also be important regulatory elements within these significant genomic regions that play an important role in determining the phenotypic diversity of these breeds.

Detection of signatures of selection in populations can be challenging as it may be confounded with various other events in the population's history that can lead to false positive results, e.g., population bottlenecks, migration, and genetic drift. Ascertainment bias is also a common problem in SNP data (Vitti et al., 2013). This study utilized whole genome sequence information from thousands of animals which should, to some extent, mitigate these issues. However, our study did not account for variation in the rate of recombination which may mimic the characteristics of selection signals (Haasl and Payseur, 2016). We also did not focus on other types of structural variants under selection such as copy number variants and tandem repeats which can play important biological roles. Moreover, the cutoff values used to initially filter the raw selection sweep signals across methods is largely arbitrary. For example, studies analyzing genotype data tend to adopt more liberal cutoffs of top 1% or 5% while those based on sequence data typically use a more conservative cutoff value such as the top 0.1% or 0.01%. For discovery of important QTLs or therapeutic targets, these thresholds may have downstream implications. In this study we first used 0.02% as a threshold of significance for individual selection signals just to highlight the peak genomic

TABLE 3 Genomic regions under selection in Hanwoo cattle identified by DCMS q values ≤ 0.05 and genes identified in those regions.

| Start | End | Chr | No of genes | Genes |
|---|---|---|---|---|
| 81860076 | 82963443 | 2 | 1 | ENSBTAG00000048361 |
| 69605816 | 69985802 | 4 | 5 | SNX10 CBX3 HNRNPA2B1 NFE2L3 MIR148A |
| 53668560 | 53873255 | 5 | 3 | SLC16A7 ENSBTAG00000055198 ENSBTAG00000053531 |
| 77883587 | 77916175 | 5 | 1 | RESF1 |
| 80625444 | 80681767 | 5 | 0 | None |
| 1092999 | 1344929 | 6 | 1 | ENSBTAG00000051456 |
| 9935965 | 10067004 | 6 | 0 | None |
| 78629307 | 79704653 | 6 | 3 | ENSBTAG00000054580 5S_rRNA TECRL |
| 41006911 | 41016811 | 7 | 9 | ENSBTAG00000039484 OR2G2 ENSBTAG00000030735 OR2G3 ENSBTAG00000039804 ENSBTAG00000052311 ENSBTAG00000054452 OR6F1 ENSBTAG00000045691 |
| 90427692 | 90752832 | 8 | 4 | TMEFF1 CAVIN4 PLPPR1 5S_rRNA |
| 1692932 | 1844248 | 9 | 0 | None |
| 20087216 | 20220978 | 9 | 1 | BCKDHB |
| 27861580 | 27896822 | 10 | 10 | ENSBTAG00000050516 ENSBTAG00000038188 ENSBTAG00000013255 ENSBTAG00000053839 ENSBTAG00000047465 ENSBTAG00000051986 ENSBTAG00000053279 ENSBTAG00000003549 OR4F15 ENSBTAG00000052056 |
| 62709174 | 62826868 | 13 | 10 | ENSBTAG00000011704 BPIFA2B ENSBTAG00000031373 BPIFA3 BPIFA1 BPIFB1 BPIFB5 CDK5RAP1 ENSBTAG00000031354 SNTA1 |
| 7070755 | 7207418 | 17 | 2 | LRBA MAB21L2 |
| 37138810 | 37552209 | 20 | 5 | bta-mir-2360 CPLANE1 NIPBL ENSBTAG00000050782 SLC1A3 |
| 43661886 | 44310133 | 24 | 4 | MC2R ENSBTAG00000048673 ENSBTAG00000046153 SETBP1 |

regions. We acknowledge this choice is subjective and these peaks were not used for any downstream analysis. Importantly, for the DCMS *p*-values we adopted a more conservative approach and used FDR cutoff of 0.05 which is widely used and acceptable in animal breeding and genetics. Candidate gene search was only performed for SNPs that passed the FDR cut-off based on the DCMS *p*-values. Theoretically, this approach should control the false positive rate in this study.

Signatures of selection can serve as a complementary method to genome-wide association studies for identification of functional variants in the genome and to provide new insights into the underlying biology of traits important for agricultural production. Since detecting selected genomic regions does not require phenotypic data, these studies can be particularly useful to identify genes for traits that are difficult or at time impossible to measure, for example, adaption to extreme environment and disease resistance. Significant genomic regions in this study may be used to select SNPs in future and test for their predictive ability. However, SNPs located in conserved genomic regions may have lower frequencies making it difficult to estimate their effects correctly and thus using them for prediction. These challenges may be overcome by overlapping results from various selection sweeps as well as GWAS particularly for traits that are known to be regulated by large effect loci. Finally, future projects comparing Hanwoo and Angus against indicine cattle breeds may also reveal candidate genes related to environmental adaptation.

## 5 Conclusion

To date, this is the largest signatures of selection study in Angus and Hanwoo beef cattle, both in terms of the density of SNPs and the number of animals per breed. We detected more selected genomic regions in Angus than in Hanwoo and the total length of genomic regions with evidence of selection was also higher in Angus. Moreover, we observed that the signatures of selection in Hanwoo and Angus are unique markedly reflecting differences in their selection history, genomic architecture and breed characteristics. More specifically, in Angus, we identified genes associated with growth, body size, feed intake, reproductive development and immunity, while in Hanwoo important genes associated with immunity, fat deposition, cholesterol metabolism, neuronal development and meat quality were identified. Future studies may help independently validate key functional genes regulating traits associated with these breeds.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Parts of the data that support the findings of this study were available from the Rural Development Administration, Republic of Korea and American Angus Association. Restrictions apply to the availability of these data,

which were used under license for this study. Requests to access these datasets should be directed to gondroce@msu.edu.

## Ethics statement

Ethical approval was not required for the study involving animals in accordance with the local legislation and institutional requirements because this study was conducted on data from commercial animals.

## Author contributions

MN: Writing–original draft, Conceptualization, Formal Analysis, Methodology, Visualization. RS: Methodology, Software, Writing–review and editing. DL: Funding acquisition, Project administration, Resources, Writing–review and editing. SL: Funding acquisition, Project administration, Resources, Supervision, Writing–review and editing. CC: Conceptualization, Funding acquisition, Resources, Supervision, Writing–review and editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Next-Generation BioGreen21 Program, by the Rural Development Administration, Republic of Korea, and by the National Institute of Food and Agriculture (AFRI Projects No. 2019-67015-29323 and 2021-67015-33411). MN was supported by an internship from Angus Genetics Inc.

## 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.2024.1368710/full#supplementary-material

## References

Abo-Ismail, M. K., Voort, G. V., Squires, J. J., Swanson, K. C., Mandell, I. B., Liao, X., et al. (2010). Factors affecting beef cattle producer perspectives on feed efficiency. *J. Anim. Sci.* 88, 3749–3758. doi:10.2527/jas.2010-2907

Albertí, P., Panea, B., Sañudo, C., Olleta, J. L., Ripoll, G., Ertbjerg, P., et al. (2008). Live weight, body size and carcass characteristics of young bulls of fifteen European breeds. *Livest. Sci.* 114, 19–30. doi:10.1016/j.livsci.2007.04.010

Bertolini, F., Servin, B., Talenti, A., Rochat, E., Kim, E. S., Oget, C., et al. (2018). Signatures of selection and environmental adaptation across the goat genome post-domestication 06 Biological Sciences 0604 Genetics. *Genet. Sel. Evol.* 50, 1–24. doi:10.1186/s12711-018-0421-y

Buxton, P., Edwards, C., Archer, C. W., and Francis-West, P. (2001). Growth/differentiation factor-5 (GDF-5) and skeletal development. *J. Bone Jt. Surg.* 83-A Suppl 1, S23–30.

Chan, E. K. F., Nagaraj, S. H., and Reverter, A. (2010). The evolution of tropical adaptation: comparing taurine and zebu cattle. *Anim. Genet.* 41, 467–477. doi:10.1111/j.1365-2052.2010.02053.x

Cho, S. H., Kim, J., Park, B. Y., Seong, P. N., Kang, G. H., Kim, J. H., et al. (2010). Assessment of meat quality properties and development of a palatability prediction model for Korean Hanwoo steer beef. *Meat Sci.* 86, 236–242. doi:10.1016/j.meatsci.2010.05.011

Ewing, G. B., and Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* 25, 135–141. doi:10.1111/mec.13390

Fink, T., Tiplady, K., Lopdell, T., Johnson, T., Snell, R. G., Spelman, R. J., et al. (2017). Functional confirmation of PLAG1 as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. *Sci. Rep.* 7, 44793–44798. doi:10.1038/srep44793

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, 52–15. doi:10.1186/s12711-020-00571-5

Grossman, S. R., Shlyakhter, 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 (327), 883–886. doi:10.1126/science.1183863

Gutiérrez-Gil, B., Arranz, J. J., and Wiener, P. (2015). An interpretive review of selective sweep studies in *Bos taurus* cattle populations: identification of unique and shared selection signals across breeds. *Front. Genet.* 6, 167. doi:10.3389/fgene.2015.00167

Haasl, R. J., and Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* 25, 5–23. doi:10.1111/mec.13339

Hoggard, N., Hunter, L., Duncan, J. S., and Rayner, D. V. (2004). Regulation of adipose tissue leptin secretion by alpha-melanocyte-stimulating hormone and agouti-related protein: further evidence of an interaction between leptin and the melanocortin signalling system. *J. Mol. Endocrinol.* 32, 145–153. doi:10.1677/jme.0.0320145

Igoshin, A. V., Yurchenko, A. A., Belonogova, N. M., Petrovsky, D. V., Aitnazarov, R. B., Soloshenko, V. A., et al. (2019). Genome-wide association study and scan for signatures of selection point to candidate genes for body temperature maintenance under the cold stress in Siberian cattle populations. *BMC Genet.* 20, 26. doi:10.1186/s12863-019-0725-0

Jun, D. J., Na, K. Y., Kim, W., Kwak, D., Kwon, E. J., Yoon, J. H., et al. (2010). Melanocortins induce interleukin 6 gene expression and secretion through melanocortin receptors 2 and 5 in 3T3-L1 adipocytes. *J. Mol. Endocrinol.* 44, 225–236. doi:10.1677/JME-09-0161

Kováčik, A., Bulla, J., Trakovická, A., Žitný, J., and Rafayová, A. (2012). The effect of the porcine melanocortin-5 receptor (Mc5R) gene associated with feed intake, carcass and physico-chemical characteristics. *J. Microbiol.* 1, 498–506.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, 1–10. doi:10.1371/journal.pcbi.1003118

Lee, S.-H., Park, B.-H., Sharma, A., Dang, C.-G., Lee, S.-S., Choi, T.-J., et al. (2014). Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. *J. Anim. Sci. Technol.* 56, 2. doi:10.1186/2055-0391-56-2

Lewis, R. S., Kolesnik, T. B., Kuang, Z., D'Cruz, A. A., Blewitt, M. E., Masters, S. L., et al. (2011). TLR regulation of SPSB1 controls inducible nitric oxide synthase induction. *J. Immunol.* 187, 3798–3805. doi:10.4049/jimmunol.1002993

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. *Hered. (Edinb)* 115, 426–436. doi:10.1038/hdy.2015.42

Maclean, C. A., Chue Hong, N. P., and Prendergast, J. G. D. (2015). Hapbin: an efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Mol. Biol. Evol.* 32, 3027–3029. doi:10.1093/molbev/msv172

Makvandi-Nejad, S., Hoffman, G. E., Allen, J. J., Chu, E., Gu, E., Chandler, A. M., et al. (2012). Four loci explain 83% of size variation in the horse. *PLoS One* 7, 1–6. doi:10.1371/journal.pone.0039929

Nawaz, M. Y., Bernardes, P. A., Savegnago, R. P., Lim, D., Lee, S. H., and Gondro, C. (2022). Evaluation of whole-genome sequence imputation strategies in Korean Hanwoo cattle. *Animals* 12, 2265. doi:10.3390/ani12172265

Nishimura, S., Watanabe, T., Mizoshita, K., Tatsuda, K., Fujita, T., Watanabe, N., et al. (2012). Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *BMC Genet.* 13, 40. doi:10.1186/1471-2156-13-40

Pizzagalli, M. D., Bensimon, A., and Superti-Furga, G. (2021). A guide to plasma membrane solute carrier proteins. *FEBS J.* 288, 2784–2835. doi:10.1111/febs.15531

Pollinger, J. P., Bustamante, C. D., Fledel-Alon, A., Schmutz, S., Gray, M. M., and Wayne, R. K. (2005). Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* 15, 1809–1819. doi:10.1101/gr.4374505

Porto Neto, L. R., Bunch, R. J., Harrison, B. E., and Barendse, W. (2012). Variation in the XKR4 gene was significantly associated with subcutaneous rump fat thickness in indicine and composite cattle. *Anim. Genet.* 43, 785–789. doi:10.1111/j.1365-2052.2012.02330.x

Ramey, H. R., Decker, J. E., McKay, S. D., Rolf, M. M., Schnabel, R. D., and Taylor, J. F. (2013). Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics* 14, 382. doi:10.1186/1471-2164-14-382

Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 15, 34–19. doi:10.1186/1471-2156-15-34

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–e0153030. doi:10.1371/journal.pone.0153013

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi:10.1038/nature01140

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164), 913–918. doi:10.1038/nature06250

Salinas, F., De Boer, C. G., Abarca, V., García, V., Cuevas, M., Araos, S., et al. (2016). Natural variation in non-coding regions underlying phenotypic diversity in budding yeast. *Sci. Rep.* 6, 21849–21913. doi:10.1038/srep21849

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., et al. (2007). A single IGF1 allele is a major determinant of small size in dogs. *Science* 316, 112–115. doi:10.1126/science.1137045

Switonski, M., Mankowska, M., and Salamon, S. (2013). Family of melanocortin receptor (MCR) genes in mammals-mutations, polymorphisms and phenotypic effects. *J. Appl. Genet.* 54, 461–472. doi:10.1007/s13353-013-0163-z

Takasuga, A. (2016). PLAG1 and NCAPG-LCORL in livestock. *Animal Sci. J.* 87, 159–167. doi:10.1111/asj.12417

Tangye, S. G., Bier, J., Lau, A., Nguyen, T., Uzel, G., and Deenick, E. K. (2019). Immune dysregulation and disease pathogenesis due to activating mutations in PIK3CD—the goldilocks' effect. *J. Clin. Immunol.* 39, 148–158. doi:10.1007/s10875-019-00612-9

Terakado, A. P. N., Costa, R. B., De Camargo, G. M. F., Irano, N., Bresolin, T., Takada, L., et al. (2018). Genome-wide association study for growth traits in Nelore cattle. *Animal* 12, 1358–1362. doi:10.1017/S1751731117003068

Utsunomiya, Y. T., do Carmo, A. S., Carvalheiro, R., Neves, H. H. R., Matos, M. C., Zavarez, L. B., et al. (2013a). Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. *BMC Genet.* 14, 52. doi:10.1186/1471-2156-14-52

Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., et al. (2013b). Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One* 8, 1–11. doi:10.1371/journal.pone.0064280

Van Laere, A. S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 425, 832–836. doi:10.1038/nature02064

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., et al. (2012). Personal and population genomics of human regulatory variation. *Genome Res.* 22, 1689–1697. doi:10.1101/gr.134890.111

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi:10.1146/annurev-genet-111212-133526

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–e0458. doi:10.1371/journal.pbio.0040072

Weir, B. S., and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evol. (N Y)* 38, 1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x

# Frontiers in Genetics

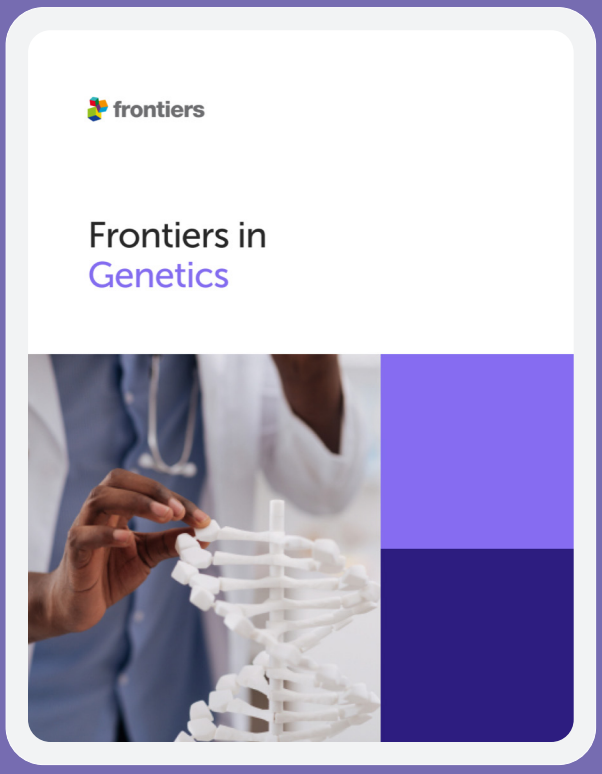
**Highlights genetic and genomic inquiry relating to all domains of life**

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