frontiers | Research Topics

# Lactation genomics and phenomics in farm animals: Where are we at?

**Edited by**
Xiao-Lin Wu, Luiz Brito, Asha Marie Miles, Yunxia Zhao,
Zhihua Jiang, Xiangdong Ding, Shu-Hong Zhao and Bjørg Heringstad

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

ISSN 1664-8714
ISBN 978-2-8325-2316-2
DOI 10.3389/978-2-8325-2316-2

## 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

# Lactation genomics and phenomics in farm animals: Where are we at?

**Topic editors**
Xiao-Lin Wu — Council on Dairy Cattle Breeding, United States
Luiz Brito — Purdue University, United States
Asha Marie Miles — United States Department of Agriculture (USDA), United States
Yunxia Zhao — Huazhong Agricultural University, China
Zhihua Jiang — Washington State University, United States
Xiangdong Ding — China Agricultural University, China
Shu-Hong Zhao — Huazhong Agricultural University, China
Bjørg Heringstad — Norwegian University of Life Sciences, Norway

**Citation**
Wu, X.-L., Brito, L., Miles, A. M., Zhao, Y., Jiang, Z., Ding, X., Zhao, S.-H., Heringstad, B., eds. (2023). *Lactation genomics and phenomics in farm animals: Where are we at?* Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-2316-2

*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*

# Table of contents

TYPE Editorial
PUBLISHED 18 April 2023
DOI 10.3389/fgene.2023.1173595

# Editorial: Lactation genomics and phenomics in farm animals: Where are we at?

Xiao-Lin Wu[1,2]*, Xiangdong Ding[3], Yunxia Zhao[4], Asha M. Miles[5], Luiz F. Brito[6], Bjorg Heringstad[7], Shuhong Zhao[4] and Zhihua Jiang[8]

[1]Council on Dairy Cattle Breeding, Bowie, MD, United States, [2]Department of Animal Sciences, University of Wisconsin, Madison, WI, United States, [3]College of Animal Science and Technology, China Agricultural University, Beijing, China, [4]College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, China, [5]Animal Genomics and Improvement Laboratory, USDA, Agricultural Research Service, Beltsville, MD, United States, [6]Department of Animal Sciences, Purdue University, West Lafayette, IN, United States, [7]Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Aas, Norway, [8]Department of Animal Science, Washington State University, Pullman, WA, United States

OPEN ACCESS

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

*CORRESPONDENCE
Xiao-Lin Wu,
nick.wu@uscdcb.com

RECEIVED 24 February 2023
ACCEPTED 11 April 2023
PUBLISHED 18 April 2023

CITATION
Wu X-L, Ding X, Zhao Y, Miles AM, Brito LF, Heringstad B, Zhao S and Jiang Z (2023), Editorial: Lactation genomics and phenomics in farm animals: Where are we at?
*Front. Genet.* 14:1173595.
doi: 10.3389/fgene.2023.1173595

KEYWORDS

milk, genomic prediction, GWAS, high-throughput phenotyping, SNP

Editorial on the Research Topic
Lactation genomics and phenomics in farm animals: where are we at?

Lactation is a crucial process for dairy animals, as it provides the primary source of nutrition for their offspring, and a balanced source of nutrients for human consumption. The studies of lactation genomics involve investigating the genome's structure, function, evolution, and regulation that underly lactation biology. Over the past two decades, genomics has revolutionized dairy cattle breeding worldwide, leading to significantly reduced generation intervals and increased genetic gain by year (Wiggans et al., 2017). Dairy farmers can now use genomic bulls for more accurate selection instead of waiting for progeny testing results. However, the success of genomic selection depends on comprehensive phenotyping, i.e., phenomics. The latter involves generating high-dimensional and close-to-biology phenotypic data on an animal-wide scale (Houle et al., 2010). This approach breaks down composite traits into more direct indicators of ultimate breeding goals that can be easily measured in large-scale farming applications (Brito et al., 2020). Livestock breeders have used complex selection indices to combine many traits into a single performance measurement for decades. Now, there is a renewed interest in collecting high-throughput data on individual animals driven by various research initiatives and promising technologies for massive, low-cost, and accurate phenotypes (Cole et al., 2020). The latest 10-year blueprint for animal genomics research, led by the U.S. Department of Agriculture, emphasizes the need to close the genome-to-phenome gaps (Rexroad et al., 2019).

This Research Topic comprises nine papers on a wide range of topics related to lactation genomics and phenomics. The accuracy of milk records is fundamental to farm management decisions and genomic predictions. Since the 1960s, cost-effective milk recording routines have been adapted to supplement the standard supervised twice-daily monthly testing scheme, assuming equal morning (AM) and evening (PM) milking intervals. However, in reality, the AM and PM milking intervals can vary considerably. Wu et al. characterized and compared additive (ACF) and multiplicative (MCF) correction factors. ACFs provide additive adjustments beyond twice AM or PM yields, while MCFs represent daily to partial milk yield ratios, although their mathematical forms and statistical interpretations vary. Overall, the MCF and linear regression models outperformed the ACF models. An exponential regression model was proposed, analogous to an exponential growth function with the yield from single milking as the initial state and the rate of change tuned by a linear function of milking interval (Wu et al., 2022). This model provided the most accurate estimates of test-day milk yields. However, discretizing milking intervals into large categories was a major concern because it led to substantial accuracy loss, regardless of the model used.

Genome-wide association studies (GWAS) and differential gene expression profiling remain the main tools for discovering genes that determine and regulate milk production and other relevant traits. Wang et al. identified seven significant SNPs associated with multiple traits that link to candidate genes with known functionalities in fat metabolism or mammary gland development. Pan et al. estimated genetic parameters and reported associations for milk production traits and somatic cell scores that varied across different lactation stages of the Shanghai Holstein population. Lin et al. showed that the long-chain acyl-CoA synthetase 1 (ACSL1) gene, which plays a vital role in fatty acids metabolism and is highly expressed in the lactating mammary gland epithelial cells of lactating animals, was associated with milk production performance in buffalos. Through a comparative analysis of expression patterns in non-lactation and lactation mammary glands of goats, sheep, and cows, Zhang et al. postulated that two ACS genes, ACSS2 and ACSF3, could participate in the formation mechanisms of the goat milk flavor. Using differential gene expression analysis in the mammary glands, Sadovnikova et al. reported a direct relationship between the response to dexamethasone, an exogenous glucocorticoid, and the concurrent suppression of milk yield due to the reduced synthesis of $\alpha$-lactalbumin and lactose by the mammary epithelium. Farhadian et al. reported potential lactation- and breed-specific SNPs in the regions of QTL and candidate genes associated with milk production using a transcriptome approach.

Genomic selection in indigenous breeds or minor breeds is often limited by size of available training populations. Therefore, combining phenotypic, pedigree, and genomic data from genetically related populations can be a feasible strategy to overcome this limitation. Teissier et al. evaluated the genetic connectedness and population structure of dairy goats from four countries and found that international genomic evaluations are feasible, especially for French and Italian goats. Using whole-genome sequence (WGS) data can enhance understanding of the genomic background of economic traits such as milk production in farm animals and the genomic predictions of genomic breeding values (Meuwissen et al., 2021). Jiang et al. demonstrated that genotype imputation from SNP arrays to WGS data was a cost-effective approach to obtain high-density genotypes for GWAS and genomic predictions, subject to model parameter optimization.

Still, the coverage of the nine topics on lactation genomics and phenomics is very limited. Appealing subjects yet not addressed include dairy breeding focusing on adaptation and environmental resilience, genomics solutions to metabolic and nutritional problems related to milk production, genomic mating toward sustainable dairy breeding, and integration of multi-omics to understand the biological mechanisms underlying lactation physiology better, to name a few. In advanced countries, high-throughput phenotyping is a reality in dairy farming. For example, large dairy farms or operations have adopted automatic milking systems capable of massive recording of phenotypes (Pedrosa et al., 2023). Precision livestock farming tools are being developed and used to collect detailed, in-depth, and high-through measurements about animal productivity, health, environmental efficiency, and welfare and their environments in or near real-time. However, these new phenotyping technologies are proprietary when offered to dairy producers, and they lack independent and unbiased validation (Cole et al., 2020). Developing non-invasive techniques for measuring phenotypic traits can improve animal welfare and reduce the costs and time associated with phenotyping, such as 3D imaging and infrared spectroscopy. New phenotypes are emerging, though their roles in genetic improvement programs remain unclear. Looking forward, we expect that integrating lactation phenomics and genomics will continue to provide a more comprehensive understanding of lactation biology and aid in developing better tools for dairy management and genetic improvement.

## Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## 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.

## References

Brito, L. F., Oliveira, H. R., McConn, B. R., Schinckel, A. P., Arrazola, A., Marchant-Forde, J. N., et al. (2020). Large-scale phenotyping of livestock welfare in commercial production systems: A new frontier in animal breeding. *Front. Genet.* 11, 793. doi:10.3389/fgene.2020.00793

Cole, J. B., Eaglen, S. A. E., Maltecca, C., Mulder, H. A., and Pryce, J. E. (2020). The future of phenomics in dairy cattle breeding. *Anim. Front.* 10 (2), 37–44. doi:10.1093/af/vfaa007

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: The next challenge. *Nat. Rev. Genet.* 11, 855–866. doi:10.1038/nrg2897

Meuwissen, T., van den Berg, I., and Goddard, M. (2021). On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. *Genet. Sel. Evol.* 53, 19. doi:10.1186/s12711-021-00607-4

Pedrosa, V. B., Boerman, J. P., Gloria, L. S., Chen, S. Y., Montes, M. E., Doucette, J. S., et al. (2023). Genomic-based genetic parameters for milkability traits derived from automatic milking systems in North American Holstein cattle. *J. Dairy Sci.* 106, 2613–2629. doi:10.3168/jds.2022-22515

Rexroad, R., Vallet, J., Matukumalli, L. K., Reecy, J., Bickhart, D., Blackburn, H., et al. (2019). Genome to phenome: Improving animal health, production, and well-being – a new USDA blueprint for animal genome research 2018–2027. *Front. Genet.* 10, 327. doi:10.3389/fgene.2019.00327

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: The USDA experience. *Annu. Rev. Anim. Biosci.* 5, 309–327. doi:10.1146/annurev-animal-021815-111422

Wu, X.-L., Wiggans, G. R., Norman, H. D., Miles, A. M., Van Tassell, C. P., Baldwin VI, R. L., et al. (2022). Daily milk yield correction factors: What are they? *JDS Commun.* 4 (1), 40–45. doi:10.3168/jdsc.2022-0230

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 06 June 2022
doi: 10.3389/fgene.2022.896910

# Novel Insight Into the Role of *ACSL1* Gene in Milk Production Traits in Buffalo

*Yuxin Lin[1,2,3], Hui Sun[1], Aftab Shaukat[1], Tingxian Deng[4], Hamdy Abdel-Shafy[5], Zhaoxuan Che[1], Yang Zhou[1], Changmin Hu[1], Huazhao Li[1], Qipeng Wu[1], Liguo Yang[1,6] and Guohua Hua[1,2,3,6]**

*[1]Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, China, [2]Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Shenzhen, China, [3]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, [4]Guangxi Key Laboratory of Buffalo Genetice, Breeding and Reproduxtion, Guangxi Buffalo Research Institute, Chinese Academy of Agricultural Sciences, Guangxi, China, [5]Department of Animal Production, Faculty of Agriculture, Cairo University, Giza, Egypt, [6]National Center for International Research on Animal Genetics, Breeding and Reproduction (NCIRAGBR); Frontiers Science Center for Animal Breeding and Sustainable Production; Key Laboratory of Smart Farming for Agricultural Animals, Huazhong Agricultural University, Wuhan, China*

OPEN ACCESS

***Edited by:***
*Xiangdong Ding,*
*China Agricultural University, China*

***Reviewed by:***
*Mao Yongjiang,*
*Yangzhou University, China*
*Xue Gao,*
*Institute of Animal Sciences (CAAS), China*

****Correspondence:***
*Guohua Hua*
*huaguohua09@gmail.com*

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

***Received:*** *15 March 2022*
***Accepted:*** *27 April 2022*
***Published:*** *06 June 2022*

***Citation:***
*Lin Y, Sun H, Shaukat A, Deng T, Abdel-Shafy H, Che Z, Zhou Y, Hu C, Li H, Wu Q, Yang L and Hua G (2022) Novel Insight Into the Role of ACSL1 Gene in Milk Production Traits in Buffalo.*
*Front. Genet. 13:896910.*
*doi: 10.3389/fgene.2022.896910*

Understanding the genetic mechanisms underlying milk production traits contribute to improving the production potential of dairy animals. Long-chain acyl-CoA synthetase 1 (ACSL1) plays a key role in fatty acid metabolism and was highly expressed in the lactating mammary gland epithelial cells (MGECs). The objectives of the present study were to detect the polymorphisms within *ACSL1* in Mediterranean buffalo, the genetic effects of these mutations on milk production traits, and understand the gene regulatory effects on MGECs. A total of twelve SNPs were identified by sequencing, including nine SNPs in the intronic region and three in the exonic region. Association analysis showed that nine SNPs were associated with one or more traits. Two haplotype blocks were identified, and among these haplotypes, the individuals carrying the H2H2 haplotype in block 1 and H5H1 in block 2 were superior to those of other haplotypes in milk production traits. Immunohistological staining of *ACSL1* in buffalo mammary gland tissue indicated its expression and localization in MGECs. Knockdown of *ACSL1* inhibited cell growth, diminished MGEC lipid synthesis and triglyceride secretion, and downregulated *CCND1*, *PPARγ*, and *FABP3* expression. The overexpression of *ACSL1* promoted cell growth, enhanced the triglyceride secretion, and upregulated *CCND1*, *PPARγ*, *SREBP1*, and *FABP3*. *ACSL1* was also involved in milk protein regulation as indicated by the decreased or increased β-casein concentration and *CSN3* expression in the knockdown or overexpression group, respectively. In summary, our present study depicted that *ACSL1* mutations were associated with buffalo milk production performance. This may be related to its positive regulation roles on MGEC growth, milk fat, and milk protein synthesis. The current study showed the potential of the *ACSL1* gene as a

**Abbreviations:** ACSL1, long-chain acyl-CoA synthetases; GS, genomic selection; LD, linkage disequilibrium; GWAS, genome-wide association studies; PM, peak milk yield; MY, milk yield; FY, milk fat yield; FP, milk fat percentage; PM, milk protein yield; PP, milk protein percentage; MGECs, mammary gland epithelial cells; TG, triglyceride; PIC, polymorphism information content; HWE, Hardy–Weinberg equilibrium.

candidate for milk production traits and provides a new understanding of the physiological mechanisms underlying milk production regulation.

**Keywords: buffalo, milk production traits, ACSL1, genetic mutation, mammary gland epithelial cells**

## INTRODUCTION

Water buffalo (*Bubalus bubalis*) is the second-largest milk producer contributing more than 15% of the world's total milk production. Buffalo is considered the most promising species in developing countries due to its high adaptability to local environmental conditions along with their significant contribution to milk and meat production (Du et al., 2019). Buffalo milk has a higher nutritional value in fat, protein, and iron and less cholesterol content compared to that of dairy cow milk (Barłowska et al., 2011). However, the low milk yield limited buffalo industry progress. Therefore, improving the buffalo milk yield while maintaining its high milk quality is the major challenge for modern buffalo breeding.

Genomic information can be directly utilized through genomic selection (GS) without the knowledge about the biological function of the genetic markers used for prediction, where GS mainly depends on the linkage disequilibrium (LD) among genetic markers and loci associated with the trait variation to create the prediction equation (Goddard et al., 2011; Wang et al., 2019). Recently, it has been reported that incorporating prior physiological knowledge and pre-selected genetic variants into GS increased the accuracy of prediction (Hayes and Daetwyler, 2019). Identification of these genetic loci can be achieved by genome-wide association studies (GWASs) and/or candidate gene approaches. In this issue, several published reports were performed to detect the genomic loci associated with milk production traits in Brazilian, Chinese, Egyptian, Iranian, Italian, and Philippine buffalo (Abdel-Shafy et al., 2020). However, none of these loci was overlapped among different populations and validated, which indicates a modest effect of each SNP and complexity of milk production traits. In this case, candidate gene approaches would be required to accurately identify the genetic markers and causative mutations associated with the relevant trait (Wilkening et al., 2009).

One of the promising candidate genes affecting milk production traits and mammary gland development is the long-chain acyl-CoA synthetases (ACSL), which have been previously detected in dairy cattle (Liang Y. et al., 2020). *ACSL* isoforms (*ACSL1*, *ACSL3*, *ACSL4*, *ACSL5*, and *ACSL6*) differ in various tissues, suggesting that each isoform may have a unique role in a specific tissue (Li et al., 2010). ACSL1 is predominantly expressed in the mammary gland epithelial cells (MGECs) in dairy cattle and is consistently upregulated before the peak of lactation (Bionaz and Loor, 2008). ACSL1 is the most prevalent long-chain acyl-CoA synthetase subtype in major metabolic tissues, which catalyzes fatty acids (FAs) to form acyl-CoA *via* an ATP-dependent process before entry into different intracellular metabolic pathways (Shi et al., 2017). Afterward, it becomes oxidized to provide acylated proteins and complex lipids such as triacylglycerol, phospholipids, and cholesterol esters (Jiang et al., 2019). Considering that triglycerides constitute over 98% of the milk fat composition, it is reasonable to propose that *ACSL1* might regulate triglyceride synthesis and related functions in MGECs.

Therefore, the objectives of this study were to identify genetic mutations of the *ACSL1* gene in buffalo and detect the association between these genetic markers and milk production traits in the tested populations. Furthermore, we tried to explore the regulatory roe of *ASCL1* on MGEC proliferation, lipid distribution, triglycerides, and β-casein synthesis.

## MATERIALS AND METHODS

### Samples and Phenotypes

A total of 331 buffalo blood DNA samples and relevant milk production records were derived from our previous studies (Deng et al., 2018; Li et al., 2018; Ye et al., 2021). In those previous studies, the Ethical Animal Care and Use Committee of Federico II University of Naples (Italy) approved the experimental design and animal treatment (Deng et al., 2018; Li J. et al., 2020). In addition, all purebred Mediterranean buffalos were selected from four herds in the southern part of Italy, and milk production traits were provided by the Italian Buffalo Breeders Association (ANASB) and the Italian Agricultural Research Council (CAR). Milk production traits were peak milk yield, total milk yield, milk fat yield, milk fat percentage, milk protein yield, and milk protein percentage. All the milk production records were adjusted to 270 days in milk as previously described (Liu J. et al., 2020). The Nanodrop 2000 spectrophotometer (Thermo-Fisher Scientific, Wilmington, DE) and 1.5% agarose gel was used to determine the concentration and quality of extracted DNA.

### SNP Identification and Genotyping

Fifty buffalo samples were randomly selected to identify the variants of the *ACSL1* gene by pooled DNA sequencing. According to the buffalo *ACSL1* genomic sequence (GenBank accession number NC_059157.1); promoter, 3′UTR, 5′UTR, and all the exon sequences were used for selective amplification by the polymerase chain reaction (PCR) (**Table 1**). The PCR products were detected by agarose gel electrophoresis and sequenced by BGI Biotechnology (Co., Ltd., Shenzhen, China). DNAstar 7.1 software (Co., Inc. Madison, Wisconsin, United States) was used to identify mutations in the sequence. Genotyping was performed by matrix-assisted laser desorption by Compass Biotechnology (Co., Ltd., Beijing, China).

### Linkage Disequilibrium and Association Analysis

Allelic frequencies, genotypic frequencies, polymorphism information content (PIC), and Hardy–Weinberg equilibrium (HWE) were calculated for each locus using PowerMarker Version 3.25. Phased genotypes were partitioned into

**TABLE 1 |** Primer information for the buffalo *ACSL1* gene.

| Primer | Sequence (5′-3′) | Region | Start and end position | Product length [bp] |
|---|---|---|---|---|
| 1 | F: GTGGTTGAAGGTGGAAGACACGA<br>R: GGTCCCCGATGCTATTTAAGGG | promoter | −330–90 | 339 |
| 2 | F: CTCCTAGGCTGCAGCGAGTGGCTGGA<br>R: TGGCCGGCAGGGTAGCCTTAGATC | 5'UTR | 12–250 | 239 |
| 3 | F: GTTTGTCACAGCATCCCTCCT<br>R: AATTTGGGGATGAGCCTCTGC | exon 1 | 22,036–22,823 | 788 |
| 4 | F: GTGGAGGATTTATGTCAGACGC<br>R: AAGTTACAGGAGGAGATAGGGAG | exon 2 | 35,590–35,967 | 378 |
| 5 | F: AGCAAAACTCAGACCCAAACC<br>R: AGCACCCCTCAAGACAGAAAG | exon 3 | 36,387–37,069 | 683 |
| 6 | F: TTTGTGTCCCTGGATTGCTTT<br>R: ATTCTCTGTGCTTTGGTTGCC | exon 4 | 39,708–40,127 | 420 |
| 7 | F: GGCAAGTGTTTTGTTCATTAGG<br>R: GTGTTCAGGGAAGGGGGCAGGG | exon 5 and exon 6 | 42,551–43,165 | 615 |
| 8 | F: CTTGGAATCAGTCCTGTTTC<br>R: GTCTTAGAGGGTGCGTGTAG | exon 7 and exon 8 | 45,470–46,200 | 731 |
| 9 | F: AAAGACATCAGCCCTGGGATTT<br>R: TTGGGGATCAGGTCCATAGTG | exon 9 | 46,378–47,145 | 768 |
| 10 | F: TGCTCTGAAATAAATGGAGAAT<br>R: TGCAAGCGGTAAAAATGAAATG | exon 10 | 49,297–49,530 | 234 |
| 11 | F: ACCGCAACTAGAGAAAAGCC<br>R: ATTGTCAAGGTGAGAAAACG | exon 11 | 51,069–52,013 | 945 |
| 12 | F: AGTGGGGTTGTTTCCTCTTT<br>R: TCTCGCTGACCTTCTCTTTTA | exon 12 | 53,396–54,116 | 721 |
| 13 | F: CGGAACCAAACCCGTCAGGTGT<br>R: CCGAAGAAAAGAAGGGGCACAT | exon 13 | 54,519–54,860 | 342 |
| 14 | F: TTGTGGTATTGTCTTCTGTGTG<br>R: CTCTGAACCTAGTATAAGGGGC | exon 14 | 55,444–55,891 | 448 |
| 15 | F: TTGTTGGAGATCAAAGCAATCT<br>R: CATGCCCCCACCCCCTGAGACT | exon 15 | 58,616–58,930 | 315 |
| 16 | F: ATAGAACTGACCCCAGCCCT<br>R: TCAAACCAGAAGCAGCAACC | exon 16 | 59,516–60,154 | 639 |
| 17 | F: CTCATCCCTTCTCTGTCTCACT<br>R: CCTGGACGTCTTATAATATTGT | exon 17 | 61,116–61,562 | 447 |
| 18 | F: TATCTATCCCATTATTTGCG<br>R: GACAGAATCAGGACCACAGC | exon 18 and exon 19 | 63,025–63,816 | 792 |
| 19 | F: CTCCCTACCCTATGTTGAGATG<br>R: GGTGGCTGTAAGGCAGTGTTCC | exon 20 | 63,870–64,229 | 360 |
| 20 | F: TTCGGAATTATTTCAGGTCACAGA<br>R: GCAACTGGAAGTGGCGGGAT | 3'UTR | 63,845–65,550 | 1706 |

*F: forward primer. R: reverse primer.*

haplotype blocks using Haploview version 4.2 (Broad Institute, Cambridge, MA, United States). Haploview 4.2 was also used to estimate the LD of all SNPs (Li et al., 2017; Ye et al., 2021). The haplotype structure of each buffalo was inferred by the software Phase 2.1 (Stephens et al., 2001).

The associations between *ACSL1* polymorphisms and milk production traits (peak milk yield, 270 days milk yield, milk fat yield, milk fat percentage, milk protein yield, and milk protein percentage) were analyzed using the custom-made R script (R Core Team., Vienna, Austria), with the following mixed linear model described by Pauciullo et al. (2012):

$$Y_{ijklmn} = \mu + G_i + P_j + S_k + F_l + a_{m(i)} + e_{ijklmn},$$

where $Y_{ijklmn}$ = phenotype observations; μ = overall mean, $G_i$ = the fixed-effect of the i$^{th}$ genotype or haplotype combination; $P_j$ = the fixed-effect of the $j^{th}$ parity (1–7); $S_k$ = the fixed-effect of the k$^{th}$ season (spring is from March to May, summer is from June to August, autumn is from September to November, and winter is from December to January and February of the following year); $F_l$ = the fixed-effect of the l$^{th}$ farm (four different farms); $a_{m(i)}$ = the random effects of the $m^{th}$ individual buffalo nested within *ACSL1* genotype or haplotype combination i$^{th}$; and $e_{ijklmn}$ = the random residual. The covariance matrices of random effects of buffalo and residual were assumed to be diagonal $\boldsymbol{I}\sigma_c^2$ and $\boldsymbol{I}\sigma_e^2$, respectively. The least-square means with standard error for multiple comparisons between different genotypes and haplotypes were performed using Bonferroni correction for multiple F-testing.

## Cell Culture and Transfection

The mammary epithelial cell line (MAC-T) was obtained from Bogoo Biotechnology (Co.,Ltd., Shanghai, China). The MAC-T cells were cultured in DMEM/F12 medium (HyClone, United States) containing 10% fetal bovine serum (Gibco, Gaithersburg, MD, United States) and 1% penicillin–streptomycin (HyClone, United States) in a 37°C incubator with 5% $CO_2$. MAC-T cells were cultured into six-well plates overnight. Then, MAC-T cells were transfected with *ACSL1* siRNA and NC or pcDNA3.1-*ACSL1* and pcDNA3.1 for 48 or 72 h, using jetPRIME transfection reagent (Polyplus-transfection, FRANCE) following the manufacture's instruction. The six-well plates were then placed in the incubator and replaced with the fresh complete medium after 6 h.

## Quantitative Real-Time PCR Assay

Total RNA was isolated using TRIzol reagent following the manufacturer's instructions. The isolated RNA was quantified spectrophotometrically at 260/280 nm and cDNA was synthesized. The qRT-PCR was conducted using Hieff qPCR SYBR Green Master Mix (Yeasen Biotech Co., Ltd, Shanghai, China) to determine the mRNA expression of the target genes using gene-specific primers (**Supplementary Table S1**). The *GAPDH* was used as a reference gene, and relative expression was measured using the $2^{-\Delta\Delta Ct}$ method.

## Western Blot

The protein was extracted from MAC-T cells, and bicinchoninic acid (BCA) assay was performed to determine the protein concentration. After protein denaturation, 15 μg sample was loaded on SDS-PAGE (10%). Subsequently, the protein was transferred to the PVDF membrane and 5% non-fat milk was used for blocking. The membranes were probed overnight by primary antibodies at 4°C and consequently incubated with secondary antibody (1:5000) for 1.5 h at room temperature. The protein expression was quantified using ImageJ software (National Institutes of Health, Bethesda, MD, United States).

## Cell Viability and Cell Counting Assays

The cells were harvested 72 h after transfection of siRNA or plasmid DNA for the determination of cell viability or cell numbers. The CCK-8 reagent (Dojindo, Japan) was supplemented in each well of the experimental group according to the manufacturer's instructions, and the cells were incubated for 1 h at 37°C. Next, the cell viability was determined at a wavelength of 450 nm.

## Cell Cycle and Apoptosis Assay

Cell cycle and apoptosis assay was performed based on a protocol established in our laboratory (Wang et al., 2020). Briefly, the cells were pretreated and the cell cycle was detected by using the cell cycle detection kit (KeyGEN BioTECH, Jiangsu, China) according to the manufacturer's instructions. Apoptosis was analyzed by using the Annexin V-FITC/PI Apoptosis Detection Kit (DOJINDO, Japan). A flow cytometer (BD FACSCalibur, America) was used to detect cell proportions.

## Triglyceride Content Detection and Bodipy Staining

A triglyceride enzyme assay kit (Jiancheng Bioengineering Institute, Nanjing, China) was used to determine the triglyceride contents in the MAC-T cell lysate. The cells were transfected with *ACSL1* siRNA or overexpression plasmid for 72 h, and 100 μl cell lysate was mixed with the working solution. The absorbance was measured at 510 nm using a microplate reader (PerkinElmer Enspire, China).

MAC-T cells were fixed at room temperature using paraformaldehyde (4%) for 15 min, washed twice with PBS, and followed by bodipy staining at room temperature for 10 min and DAPI staining for 5 min in the dark. Then, an anti-fluorescence quenching agent was added, and pictures were taken with an inverted fluorescence microscope (AXIO OBSERVER, ZEISS). Finally, the fluorescence intensity of lipid droplets was measured by ImageJ.

## Enzyme-Linked Immunosorbent Assay

A commercial β-casein kit (Mlbio, Shanghai, China) was used to detect the concentration of β-casein. MAC-T cells were transfected with *ACSL1* siRNA or overexpression plasmid for 72 h and cell culture medium was collected. A total of 50 μl culture medium was used to detect the β-casein secretion level. Absorbance was measured at 450 nm using a microplate reader (PerkinElmer Enspire, China).

## Statistical Analyses for Gene Expression

The statistical analyses of gene functional studies were conducted with SPSS 19.0 software (SPSS Inc., Chicago, IL, United States) and graphing with Graphpad Prism v5.0 (GraphPad Software, Inc., La Jolla, CA, United States). The results are expressed as means ± standard error of the mean (Mean ± SEM). Significant differences between the two groups were compared using Student's *t*-test, and comparisons among multiple groups were performed with a one-way analysis of variance followed by Dunnett's test. *P-value* < 0.05 was considered statistically significant. All experiments were conducted at least three times.

# RESULTS

## *ACSL1* Gene SNP Screening and Genotyping

Buffalo DNA pool sequencing data identified a total of twelve potential SNPs in the tested samples (**Supplementary Figure S1**). Among these twelve SNPs, three were located in the exonic region, and the remaining nine were in the intronic region. The SNP at g.517571A >G, g.524019A >G, g.529284A >G, g.530394C >G, and g.534640A >G was located within intron 9, intron 11, intron 15, intron 16, and intron 20, respectively. The SNPs at g.519961C >T and g.522165C >T were located within intron 10, and the SNPs at g.531913A >C, g.532009C >T, and g.532389A >C were located within intron 17 (**Supplementary Table S2**). The SNPs at g.492696A >G (exon1), g.492756A >G (exon1), and g.531913A >C (exon17) were all synonymous mutations (**Supplementary Table S2**).

Detected SNPs were genotyped by MALDI-TOF-MS in 331 Mediterranean buffalo samples (**Supplementary Figure S2**). Genetic analysis for the tested samples showed that allele frequencies of all SNPs were higher than 15%, and the genotype frequencies of AA in g.492696A >G and g.531913A >C, GG in g.492756A >G and g.534640A >G, TT in g.532009C

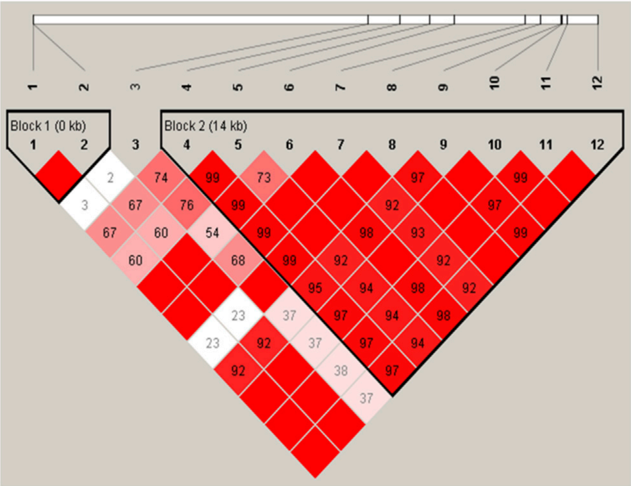


**FIGURE 1 |** Linkage disequilibrium of the twelve SNPs was detected in the *ACSL1* gene in buffalo. The red squares represent high pairwise linkage disequilibrium, coloring down to white squares of low pairwise linkage disequilibrium, and the linkage disequilibrium is shown as D′.

>T, and CC in g.532389A >C were lower than 10%, and the frequencies of other genotypes were all higher than 10% (**Supplementary Table S2**). In addition, all the identified SNPs were in accordance with Hardy–Weinberg equilibrium ($\chi^2$ test, $P > 0.01$) and were moderately polymorphic ($0.25 <$ PIC $<0.50$) (**Supplementary Table S2**).

## Association Study of *ACSL1* Genotypes With Milk Production Traits

The association analysis between the twelve detected SNPs and six milk production traits (peak milk yield, 270 days milk yield, milk fat yield, milk fat percentage, milk protein yield, and milk protein percentage) was conducted. The results showed that nine SNPs were associated with at least one of the milk production traits ($p < 0.05$ or $p < 0.01$) (**Supplementary Table S3**).

The SNPs at g.492696A > G and g.492756A > G loci were significantly associated with milk protein percentage (PP), and buffaloes with mutant type GG at g.492696A > G and g.492756A >G showed the lowest (4.52% ± 0.07%) and highest (4.66% ± 0.09%) PP than those with the A-allele ($p < 0.05$). The SNP at g.517571A >G loci was significantly associated with 270 days of milk yield (MY), and buffaloes with mutant type GG had significantly lower MY (2684.28 ± 89.64 kg) than those with A-allele ($p < 0.001$). The SNP at g.522165C >T loci was significantly associated with MY, milk fat yield (FY), and milk protein yield (PY), where buffaloes with the wild type CC showed the lowest MY(2726.10 ± 87.59 kg), FY(215.38 ± 8.12 kg), and PY(123.93 ± 4.07 kg) compared to the other genotypes ($P < 0.05$ or $P < 0.01$). The SNPs at g.529284A >G loci were significantly associated with FY, where buffaloes with the mutant type GG showed the lowest FY (214.58 ± 8.21 kg) compared to the other genotypes ($p < 0.05$). The SNP at g.531913A >C locus was associated with peak milk yield (PM), MY, and PY, and the heterozygous buffaloes had significantly lower PM (14.82 ± 0.41 kg), MY(2743.01 ± 86.22 kg) and PY(124.68 ± 4.00 kg) than those with homozygous genotypes ($p < 0.05$). The SNPs at g.532009C >T, g.532389A >C and g.534640A >G loci were significantly associated with PM. The buffaloes with the heterozygotic type TC, CA, and GA at g.532009C >T, g.532389A >C, and g.534640A >G, respectively, showed the lowest PM compared to the other genotypes ($p < 0.01$) (**Supplementary Table S3**).

## Linkage Disequilibrium and Haplotypes Analysis

We further performed LD and haplotype analysis for the twelve detected SNPs, and two haplotype blocks were identified. The LD plot showed that two SNPs were in complete LD resided in haplotype block 1 (D’ = 1), and the remaining nine were in strong LD resided in haplotype block 2 (D’ > 0.8) (**Figure 1**).

We identified three major haplotype combinations (haplotype pairs) (H1H1, H1H2, and H2H2) in block 1 (**Table 2**), and generated six major haplotype combinations (H1H1, H2H1, H3H1, H2H3, H4H1, and H5H1) in block 2 (**Table 3**). All of them accounted for a higher frequency of over 5% in the studied subjects. These were selected to perform haplotype-based association analysis. The results showed that all nine haplotype combinations were highly associated with milk protein percentage ($p < 0.05$) (**Tables 2, 3**). Moreover, the individuals with H2H2 haplotype combination in block 1 (**Table 2**) and H5H1 in block 2 (**Table 3**) obtained a higher milk protein percentage than other individuals ($p < 0.05$).

**TABLE 2 |** Assocation analysis between haplotypes of block 1 in the *ACSL1* gene and milk production traits.

| block1 | Frequency (no.) | Sequence | Peak milk yield (kg) | 270 days milk yield (kg) | Milk fat yield (kg) | Milk fat percentage (%) | Milk protein yield (kg) | Milk protein percentage (%) |
|---|---|---|---|---|---|---|---|---|
| H1H1 | 0.58 (191) | GA/GA | 14.71 ± 0.10 | 2814.27 ± 22.22 | 231.37 ± 2.05 | 8.23 ± 0.04 | 129.74 ± 0.96 | 4.63 ± 0.01$^{b}$ |
| H1H2 | 0.35 (116) | GA/AG | 14.62 ± 0.15 | 2810.01 ± 28.99 | 228.17 ± 2.63 | 8.13 ± 0.05 | 130.98 ± 1.33 | 4.67 ± 0.02$^{ab}$ |
| H2H2 | 0.06 (20) | AG/AG | 14.70 ± 0.29 | 2725.47 ± 63.67 | 221.94 ± 5.80 | 8.17 ± 0.13 | 128.10 ± 2.72 | 4.72 ± 0.04$^{a}$ |
| P value | | | 0.9993 | 0.5661 | 0.3419 | 0.3555 | 0.4354 | 0.0149 |

*The values of milk production traits in each genotype are represented as mean ± SE.,. The values with different superscripts within the same column differed significantly at* $p < 0.05$.

**TABLE 3 |** Assocation analysis between haplotypes of block 2 in the *ACSL1* gene and milk production traits.

| block2 | Frequency (no.) | Sequence | Peak milk yield (kg) | 270 days milk yield (kg) | Milk fat yield (kg) | Milk fat percentage (%) | Milk protein yield (kg) | Milk protein percentage (%) |
|---|---|---|---|---|---|---|---|---|
| H1H1 | 0.2030 (67) | CTGAGCCAA/ CTGAGCCAA | 14.89 ± 0.20 | 2802.20 ± 39.06 | 224.72 ± 3.66 | 8.03 ± 0.08 | 131.05 ± 1.78 | 4.69 ± 0.02$^{abc}$ |
| H2H1 | 0.1636 (54) | TCAGCCCAA/ CTGAGCCAA | 14.98 ± 0.21 | 2869.15 ± 38.36 | 234.24 ± 3.42 | 8.19 ± 0.08 | 131.96 ± 1.67 | 4.61 ± 0.02$^{c}$ |
| H3H1 | 0.1515 (50) | TCAGCATCG/ CTGAGCCAA | 14.38 ± 0.19 | 2796.19 ± 42.92 | 233.22 ± 3.98 | 8.34 ± 0.08 | 129.50 ± 1.92 | 4.64 ± 0.02$^{bc}$ |
| H2H3 | 0.0758 (25) | TCAGCCCAA/ TCAGCATCG | 14.40 ± 0.26 | 2650.07 ± 56.45 | 221.82 ± 5.71 | 8.33 ± 0.11 | 123.55 ± 2.57 | 4.67 ± 0.02$^{abc}$ |
| H4H1 | 0.0697 (23) | TCGACCCAA/ CTGAGCCAA | 14.54 ± 0.36 | 2745.29 ± 60.81 | 229.84 ± 5.77 | 8.38 ± 0.10 | 129.79 ± 2.74 | 4.74 ± 0.04$^{ab}$ |
| H5H1 | 0.0515 (17) | CTGACCCAA/ CTGAGCCAA | 14.49 ± 0.31 | 2674.64 ± 68.12 | 219.56 ± 6.94 | 8.20 ± 0.14 | 128.5 ± 3.41 | 4.81 ± 0.04$^{a}$ |
| P value | | | 0.1446 | 0.0849 | 0.2029 | 0.2611 | 0.1299 | 0.0364 |

*The values of milk production traits in each genotypes are represented as mean ± SE.,. The values with different superscripts within the same column differed significantly at* p *< 0.05.*

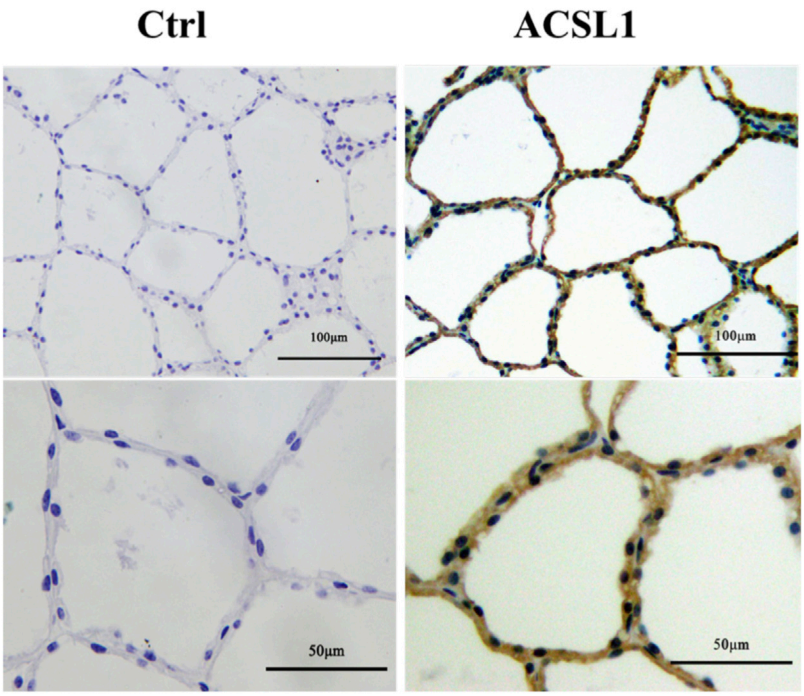


**FIGURE 2 |** Expression and localization of ACSL1 in buffalo mammary gland. Immunohistochemistry staining of ACSL1 in buffalo mammary gland tissue. The brown color indicated ACSL1 immuno signal, and the nuclei were counterstained in blue. Scale bar: 100 μm (upper panel, 20 times magnification) and 50 μm (lower panel, 40 times magnification).

Thus, H2H2 was regarded as a dominant haplotype pair in block 1 and H5H1 was the dominant haplotype pair in block 2 for increasing milk protein percentage.

## Expression and Localization of *ACSL1* in Buffalo Mammary Gland

To identify ACSL1 localization in buffalo mammary gland tissue, we performed immunohistological staining. The result showed that the mammary glands had closely arranged epithelial cells and showed numerous acinar cavities (**Figure 2**). Furthermore, the buffalo mammary epithelial cells displayed specific immunolabeling for *ACSL1*, of which cytoplasm was intensely labeled (**Figure 2**).

## *ACSL1* Regulates the Mammary Epithelial Cell Growth

The potential effects of *ACSL1* on cellular functions were investigated in an *in vitro* model using MAC-T mammary epithelial cells. The cells were transfected with siRNA, and

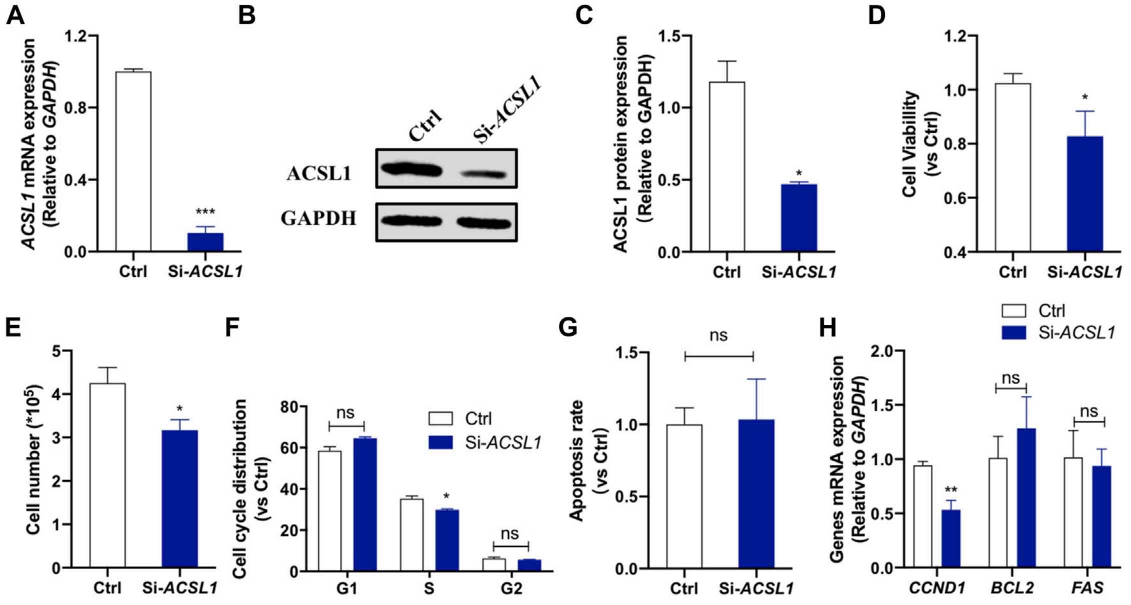


**FIGURE 3 |** *ACSL1* interference inhibited mammary epithelial cell growth. **(A)** MAC-T cells were transfected with si-*ACSL1*, and fluorescence quantitative qPCR was used to detect the *ACSL1* mRNA levels; **(B,C)** Western Blot detected the protein expression of *ACSL1*; **(D)** CCK-8 assays were applied to check the cell viability after *ACSL1* knockdown; **(E)** Living cell number in control (Ctrl) and *ACSL1* knockdown (Si-*ACSL1*) groups; **(F)** Flow cytometry was used to detect cell cycle progression; **(G)** Quantification of apoptosis by flow cytometry; **(H)** mRNA expression of cell cycle and cell apoptosis–related genes. *GAPDH* was used as the inner control; *$p$ < 0.05, **$p$ < 0.01, and ***$p$ < 0.001, ns: nonsignificant difference.*ACSL1* overexpression was then performed to confirm its regulatory role on mammary epithelial cell growth by using *ACSL1*-overexpressing plasmid (pcDNA3.1-*ACSL1*). Transfection of *ACSL1*-overexpressing plasmid significantly increased *ACSL1* mRNA (**Figure 4A**) and protein abundance (**Figures 4B,C**). The CCK-8 assay showed that *ACSL1* overexpression resulted in a significant promotion in cell viability ($p$ < 0.01) (**Figure 4D**). The cell counting test showed that MAC-T cells were significantly increased after *ACSL1* overexpression ($p$ < 0.01) (**Figure 4E**). Then, flow cytometric analysis demonstrated a significant difference in cell cycle distribution in *ACSL1* overexpression cells ($p$ > 0.05) (**Figure 4F**). The cell apoptosis rate showed no significant differences between the control and *ACSL1* overexpression cells (**Figure 4G**). The overexpression of *ACSL1* upregulated *CCND1* expression ($p$ < 0.01), while that of *BCL2* and *FAS* remained unchanged (**Figure 4H**).

knockdown was confirmed by qRT-PCR and Western-blot. The results showed that RNA interference downregulated *ACSL1* mRNA expression by 90% ($p$ < 0.001) (**Figure 3A**) and protein expression by 51% compared to the control group ($p$ < 0.05) (**Figures 3B,C**). To confirm the *ACSL1* role in cell proliferation, we performed CCK-8 assays to examine the effect of *ACSL1* on the viability of MAC-T cells. The results demonstrated that cell viability was significantly reduced by *ACSL1* knockdown ($p$ < 0.05) (**Figure 3D**). In addition, the cell counts were measured with an automatic cell counter, and the results revealed a significant decrease in the number of cells in *ACSL1* knockdown cells (**Figure 3E**). We next examined cell cycle and apoptosis using flow cytometry. *ACSL1* knockdown resulted in a severe S-phase arrest ($p$ < 0.05) (**Figure 3F**), whereas knockdown of *ACSL1* had no major impact on cell apoptosis progression ($p$ > 0.05) (**Figure 3G**). Consistently, *ACSL1* knockdown inhibits the cell cycle–related gene (*CCND1*) expression ($p$ < 0.01), without changing *Bcl2* and *FAS* expression (**Figure 3H**).

## *ACSL1* Regulate Mammary Epithelial Cell Lipogenesis

To make a thorough exploration of *ACSL1* function in mammary epithelial cells, we detected the effect of *ACSL1* on milk fat synthesis. The BODIPY staining of neutral lipid accumulation confirms the reduction of lipid droplets in *ACSL1* knockdown cells ($p$ < 0.001) (**Figures 5A,B**). The secretory effect of *ACSL1* on the level of triglyceride (TG), a major lipid milk fat, was examined. The results showed that *ACSL1* knockdown reduced the secretion of triglycerides ($p$ < 0.05) (**Figure 5C**) and the overexpression of *ACSL1* led to an increase of triglyceride content of 38% over the control group ($p$ < 0.01) (**Figure 5D**) in MAC-T cells. We next examined the expression of genes associated with lipid anabolism. The results showed that *ACSL1* knockdown decreased *FABP3* ($p$ < 0.01) and *PPARγ* ($p$ < 0.05) expression but did not alter the mRNA level of *SREBP1* and *AGPAT6* ($p$ > 0.05) (**Figure 5E**). In contrast, the overexpression of *ACSL1* promotes the expression of *FABP3*, *PPARγ*, *SREBP1*, and *AGPAT6* ($p$ < 0.05) (**Figure 5F**).

## *ACSL1* Regulated Mammary Epithelial Cell Casein Synthesis

Our previous association study revealed that *ACSL1* mutation affected the milk protein percentage (**Supplementary Tables S4, S5**). We further performed the ELISA assay to detect the β-casein (a major lactoprotein) levels in the culture medium and κ-casein (*CSN3*) expression of mammary gland epithelial cells after *ACSL1* silencing or overexpression. Here, we showed that *ACSL1*

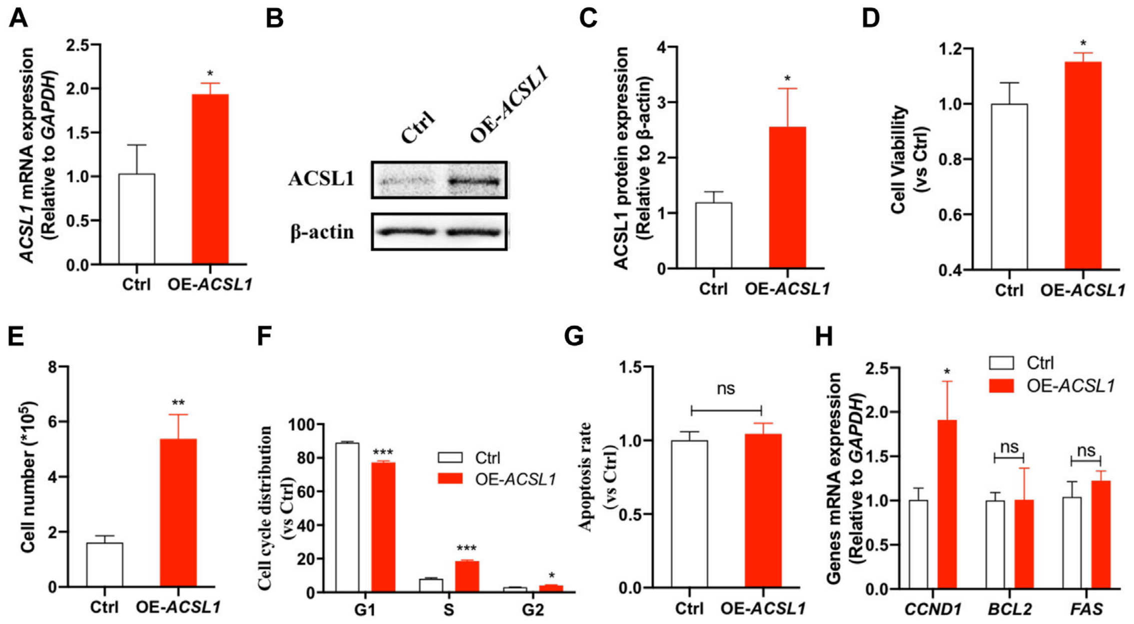


**FIGURE 4 |** *ACSL1* overexpression promotes cell growth. **(A)** MAC-T cells were transfected with pcDNA3.1-*ACSL1* for 48°h, and fluorescence quantitative qPCR was used to detect the changes of *ACSL1* mRNA levels; **(B,C)** Western blot detects the protein expression of ACSL1; **(D)** CCK-8 assays were applied to check the cell viability after *ACSL1* overexpression; **(E)** Living cell number in control (Ctrl) and *ACSL1* overexpression (OE-*ACSL1*) groups; **(F)** Flow cytometry was applied to cell cycle progression; **(G)** Quantification of apoptosis by flow cytometry; **(H)** mRNA expression of cell cycle and cell apoptosis–related genes. *GAPDH* was used as inner control; $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$, ns: nonsignificant difference.

knockdown significantly attenuated β-casein production and downregulated κ-casein (*CSN3*) expression (**Figures 6A,B**, $P < 0.05$). The overexpression of *ACSL1* significantly increased β-casein production and *CSN3* expression ($p < 0.05$ or $p < 0.01$) ().

## DISCUSSION

Milk production traits are complex in nature, where several genes are involved in their regulation along with different environmental factors. Nowadays, the selection of superior animals for increasing the frequency of desired alleles with a positive effect on a given trait focuses on genetic improvement in livestock. Moreover, identifying the single-nucleotide polymorphisms (SNPs) for milk production traits is currently being essential to increase the accuracy of prediction for animal genetic merit, which is useful for genetic improvement of production traits in livestock (Jiang et al., 2019). Importantly, the *ACSL1* is deregulated in many tumors, leading to abnormal lipid synthesis and extracellular lipid uptake that promotes uncontrolled cancer cell proliferation (Rossi Sebastiano and Konstantinidou, 2019). Recently, a lot of evidence indicated that mutation in the *ACSL1* might affect the production performance. Quantitative trait loci analysis demonstrated that *ACSL1* is a candidate gene for the location and function of the fatty acid composition of bovine skeletal muscle (Widmann et al., 2011). Polymorphism analysis of *ACSL1* suggested an association between genotype and backfat thickness. Manichaikul et al. (2016) demonstrated that three SNPs located in the intronic regions of *ACSL1* are associated with the level of glucose in fasting or diabetes. However, *ACSL1* genetic data regarding milk production are very preliminary, especially in buffalo. In this context, Liang S. S. et al. (2020) have observed high expression levels of the *ACSL1* mRNA in the mammary tissue of lactating buffalo, suggesting that *ACSL1* may be related to lactation performance of buffalo. In our study, the presence of ACSL1 protein was detected in mammary epithelial cells by immunohistochemistry, and it was mainly in the cytoplasm, which was consistent with the findings of Wang et al. (2017) in the study on the relationship between *ACSL1* and human breast cancer.

The haplotype analysis of the *ACSL1* promoter region in *Bos grunniens* has a significant correlation with the milk protein percentage and milk fat percentage. In the present study, identified *ACSL1* polymorphisms (g.531913A >C, g.532009C >T, g.532389A >C, and g.534640A >G) were associated with peak milk yield (g.517571A >G, g.522165C >T, and g.531913A >C), with 270 days milk yield (g.522165C >T and g.529284A >G), with milk fat yield (g.5522165C >T and g.531913A >C), with milk protein yield (g.492696A >G and g.492756A >G), and with milk protein percentage. Li et al. (2012) identified four SNPs in the pig *ACSL1*, and the mutations of exon were all synonymous. In our study, the SNPs of g.492696A >G, g.492756A >G, and g.531913A >C located in the exon were synonymous substitutions. Apparently, the synonymous mutations do not alter the amino acid encoded by the affected codon due to the degeneracy of the genetic code but change the DNA and RNA sequence (Sharma et al., 2019). Nonetheless, recent studies suggested their

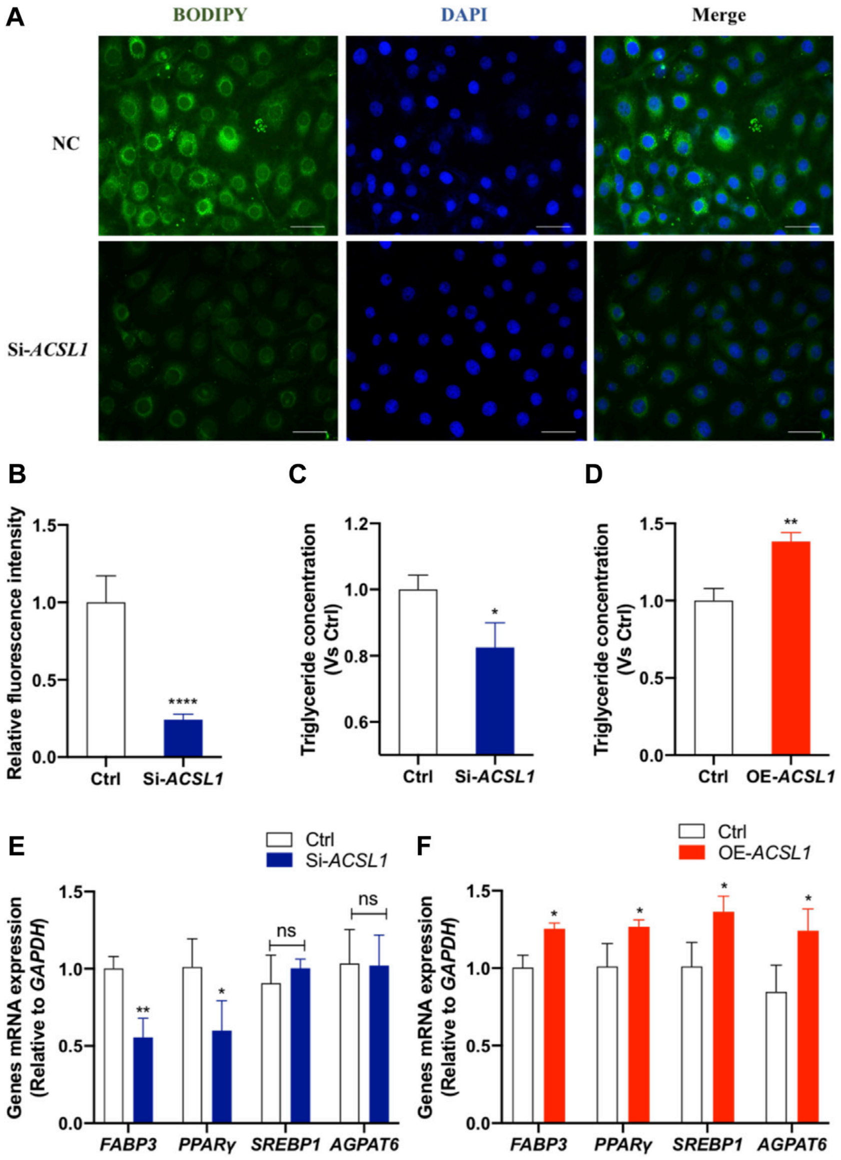


**FIGURE 5 |** *ACSL1* regulated lipogenesis and triglyceride synthesis in MAC-T cells. **(A)** MAC-T cells were transfected either with *ACSL1* siRNA or negative control for 72 h. Bodipy staining (green) was used to indicate the lipid distribution, and nuclei were stained by DAPI (blue). Scale bar: 20 μm; **(B)** Quantification of *BODIPY* + fluorescent signal density; **(C,D)** Triglyceride concentration was detected in the cell lysate. Triglyceride concentration was normalized by control (Ctrl); **(E,F)** mRNA expression of lipid metabolism–related genes after *ACSL1* knockdown or overexpression, and *GAPDH* was used as the inner control. $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.0001$, ns: nonsignificant difference.

significant impact on splicing, RNA stability, RNA folding, translation, or co-translational protein folding. In addition, many studies have revealed that synonymous mutations play a role in a variety of human diseases and can be linked to a patient's clinical outcome or responsiveness to treatment (Schutz et al., 2013). Thus, the expression of *ACSL1* may be affected by SNP g.492696A >G (exon1), g.492756A >G (exon1), and g.531913A >C (exon17), which has an influence on milk fat metabolism and ultimately affects some buffalo milk production traits. The remaining nine SNPs found in the intronic region were non-functional SNPs and did not lead to alterations in amino acids. Nevertheless, an increasing amount of evidence reveals that noncoding regions in the genome cause abnormal splicing of gene transcripts. Similarly, Rose (2008) investigated that one is often overlooked. Still, many genes with an intact promoter were essentially not expressed at all without an intron, while many genes with an intact promoter were essentially not expressed at all without an intron. Hence, the SNP of g.517571A >G, g.519961C >T, g.522165C >T, g.524019A >G, g.529284A >G, g.530394C >G, g.532009C >T, g.532389A >C, and g.534640A >G may affect the milk producing traits by affecting ACSL1 protein formation or linkage with other marker loci associated with milk-production

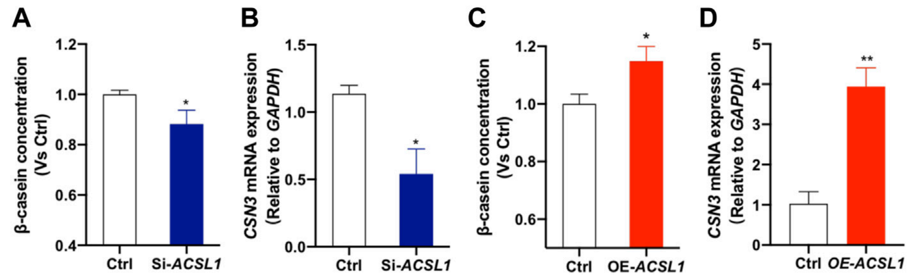


**FIGURE 6 |** *ACSL1* regulated β-casein synthesis and κ-casein expression in MAC-T cells. **(A,C)** Content of β-casein in the MAC-T cell culture supernatant was determined by ELISA, and β-casein concentration was normalized by control (Ctrl); **(B–D)** mRNA expression of κ-casein (*CSN3*) after *ACSL1* knockdown or overexpression, and *GAPDH* was used as the inner control. *$p < 0.05$; **$p < 0.01$.

traits. As a matter of fact, compared to individual SNPs, LD and haplotypes had more genetic information. Testing multiple SNPs simultaneously can capture the underlying architecture of complex quantitative traits better (Abdel-Shafy et al., 2014). For this purpose, Li et al. (2019) claimed that H2H3 and H2H2 in Chinese Holstein cow *FBP2* were the dominant haplotype combinations, improving milk yield, milk fat, and milk protein. Our study indicated that *ACSL1* functional diplotypes (H1H1, H1H2, and H2H2) in block 1, comprising haplotypes from two detected SNPs (g.492696A >G and g.492756A >G), and (H12H12, H1H12, H4H12, H1H4, and H5H12) in block 2, comprising haplotypes from nine detected SNPs (g.519961C >T, g.522165C >T, g.524019A >G, g.529284A >G, g.530394C >G, g.531913A >C, g.532009C >T, g.532389A >C, and g.534640A >G), were associated with milk protein percentage. Under selection, haplotype-based approaches have further advantages, suggesting that H5H1 diplotypes in block 2 were selected during artificial selection. This research is the first study to examine *ACSL1* polymorphisms associated with buffalo milk production traits to the best of our knowledge. The exploration of *ACSL1* genetic variants can provide added value to buffalo molecular breeding.

The quantity and activity of mammary epithelial cells are known to be linked with lactation and play a key role in the growth of mammary glands (Boutinaud et al., 2004). According to the NCBI buffalo and dairy cow *ACSL1* genomic sequences (Gene ID: 102414095 and Gene ID: 537161, respectively), buffalo *ACSL1* showed close homology (98%) to dairy cow *ACSL1* sequences. Therefore, we explored the *ACSL1* regulation on mammary epithelial cell growth. In all types of cell cultures, the measurement of cell viability is crucial and is often used to determine cell proliferation within a cell population. Chen et al. (2016) found that *ACSL1* knockdown inhibited breast cancer cell proliferation. In the present study, *ACSL1* knockdown reduced cell viability, and *ACSL1* overexpression significantly increased cell viability, indicating that the effect of *ACSL1* on cell viability was consistent with the aforementioned research results. The majority of the mammary epithelial cells are secretory cells that undergo functional differentiation to generate milk during pregnancy (Qiu et al., 2019). The number of mammary epithelial cells affect milk yield during lactation. The decline in milk production after peak lactation is accompanied by a gradual reduction in the number of mammary epithelial cells (Boutinaud et al., 2004). In our study, *ACSL1* overexpression significantly increased the cell population, while its depletion downregulated the cell population. These findings were consistent with those of the cell viability analysis, which further supported that *ACSL1* might promote mammary epithelial cell proliferation.

As for cell cycle analysis, Ma et al. (2021) found that *ACSL1* knockdown blocked the cell cycle and stopped prostate cancer cells from proliferating and migrating. Similarly, our results showed that *ACSL1* knockdown resulted in the G1/S-phase arrest and affected the DNA synthesis in mammary epithelial cells, while *ACSL1* overexpression increased the S-phase rate. The cell cycle is a complex process tightly regulated by the cyclins and their catalytic moieties. It has been shown that recombinant complexes of *CDK4* or *CDK6* and *CCND1* are necessary for the G1/S transition (Ma et al., 2015). In our study, we found that *ACSL1* positively regulated *CCND1* expression, which further supported that *ACSL1* regulates the cell cycle and affects cell proliferation. Actually, following a cyclic pattern of lactation–involution–lactation, mammary epithelial cells experience multiplication, differentiation, apoptosis, and regeneration (Monks et al., 2008). As for apoptosis analysis, Zhao et al. (2019) concluded that a lack of *ACSL1* causes a generalized impairment in muscle fuel metabolism, which leads to an increase in protein catabolism, resulting in myocyte apoptosis. In addition, inhibition of *ACSL1* during fatty acid loading results in macrophage apoptosis *via* the accumulation of free fatty acids (Pan et al., 2007). Together, these results suggested that *ACSL1* regulated mammary epithelial cell growth and may pose a positive role in bovine milk yield.

The lipids are a major energetic constituent of milk, and the principal lipids of milk are triacylglycerides, representing up to 98% of the total lipids (Liu Z. et al., 2020). Several studies have confirmed that the triglyceride increased after *ACSL1* overexpression or decreased after *ACSL1* knockdown (Li T. et al., 2020). As expected, in our study, *ACSL1* overexpression increased triglyceride in MAC-T cells, and *ACSL1* knockdown decreased triglyceride levels. Lipid droplets, which promote coordination and communication between diverse organelles and serve as key hubs of cellular metabolism, are the most common storage form for neutral lipids. Lian et al. (2016) indicated that bta-miR-181a negatively regulated *ACSL1* and then proved that *ACSL1* positively regulated lipid droplet and triglyceride synthesis. Zhao et al. (2020) also reported that *ACSL1* overexpression resulted in lipid droplet aggregation. Coincidentally, in this study, *ACSL1* knockdown inhibits the accumulation of lipid droplets, which was consistent with the abovementioned results. PPARγ can maintain mature adipocytes and promote adipogenesis. PPARγ activation increased triglyceride content and elevated the number and size of lipid droplets in the mouse liver (Liu et al., 2021). Zhao et al. (2020) found that *ACSL1* overexpression significantly increased *PPARγ* expression and triglyceride secretion, while significantly decreasing FA oxidation–related gene *CPT1A* expression. The results of the present study were consistent with these findings as *ACSL1* positively regulates *PPARγ* expression and triglyceride secretion. However, other studies had dictated that *PPAR* is involved in the β-oxidation of fatty acids in the liver. Li T. et al. (2020) indicated that *ACSL1* negatively regulated *PPARγ* in human liver cells, and *ACSL1* overexpression reduces fatty acid β-oxidation *via* the PPAR*γ* pathway, resulting in a rise in triglyceride levels. In contrast to adipocytes and liver cells, *ACSL1*-deficient macrophages do not reduce β-oxidation (Rubinow et al., 2013). These findings suggested that the function of *ACSL1* may differ in cells. The changes in the upstream signaling cascade and transcriptional networks that regulate *ACSL1* expression, in particular, may have an impact on the entry of fatty acyl-CoAs into several metabolic processes. In addition, *SREBP1* is the key positive regulator in milk fat synthesis of dairy cow mammary epithelial cells (He et al., 2020). *AGPAT6* is highly expressed in mammary epithelium tissue, which is crucial for producing milk fat. *FABP3* upregulated the expression of *SREBP1* and *PPARγ* to increase lipid droplet accumulation. Bionaz and Loor (2008) reported that *ACSL1*, *FABP3*, and *AGPAT6* coordinate and regulate the channeling of fatty acids toward copious milk fat synthesis in bovine mammary glands. In our study, *ACSL1* positively regulated *SREBP1*, *FABP3*, *PPARγ*, and *AGPAT6* mRNA expression. Therefore, it is suggested that the lipogenesis process was regulated by *ACSL1* in MAC-T cells.

Caseins are an important group of proteins in milk that accounted for approximately 80% of milk proteins and are secreted by the mammary epithelial cells (Cavaletto et al., 2008). There are four types of casein: $\alpha_{s1}$-casein, $\alpha_{s2}$-casein, β-casein, and κ-casein, all of which possess different structures and functionality, and both $\alpha_{s1}$-casein and β-casein are major caseins (Cosenza et al., 2021). Wang et al. (2014) found that *Pten* downregulates dairy cow mammary epithelial cell secretion of β-casein. The present study found that *ACSL1* knockdown resulted in a significant reduction in β-casein content, and *ACSL1* overexpression significantly increased β-casein secretion. *ACSL1* polymorphisms were significantly associated with milk protein yield and milk protein percentage. Accordingly, *ACSL1* may affect milk protein synthesis and lactation in MAC-T cells.

## CONCLUSION

In conclusion, it is demonstrated that twelve SNPs regulate *ACSL1* in buffalo. Four SNPs were significantly associated with peak milk yield; three SNPs were significantly associated with 270 days milk yield; two SNPs were significantly associated with 270 days milk fat yield; two SNPs were significantly associated with 270 days milk protein yield, and two SNPs were significantly associated with milk protein percentage. Three diplotypes in block 1 and six diplotypes in block 2 were associated with protein percentage, and H5H1 in block 2 was the dominant diplotype. Furthermore, *ACSL1* positively regulated the cell growth, triglyceride and casein synthesis, and related gene expressions such as *CCND1* and *PPARγ*. These findings provide evidence that the buffalo *ACSL1* gene may be a potential candidate gene for marker-assisted selection in the buffalo breeding program.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The name of the repository and link to the data can be found below: Harvard Dataverse; https://doi.org/10.7910/DVN/KBVOI6.

## AUTHOR CONTRIBUTIONS

YL conducted all the experiments and wrote the original draft. HS designed a partial experiment. TD is in charge of data collection. AS and HA contributed to manuscript writing and revising. CH provided partial materials. ZC, HL, and QW were responsible for statistical analysis. LY provided equipment and technical guidance. GH managed the project and provided financial support.

## FUNDING

This research is financially supported by the Natural Science Foundation of China (No. 32172732), the Fundamental Research Funds for the Central Universities (No. 2662022DKYJ002), the HZAU-AGIS Cooperation Fund (No. SZYJY2022025) and the China Agriculture Research System of MOF and MARA.

## ACKNOWLEDGMENTS

We appreciate Dr. Giuseppe Campanile's group who facilitated the sample collection.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.896910/full#supplementary-material

## REFERENCES

Abdel-Shafy, H., Bortfeldt, R. H., Tetens, J., and Brockmann, G. A. (2014). Single Nucleotide Polymorphism and Haplotype Effects Associated with Somatic Cell Score in German Holstein Cattle. *Genet. Sel. Evol.* 46, 35. doi:10.1186/1297-9686-46-35

Abdel-Shafy, H., Awad, M. A. A., El-Regalaty, H., El-Assal, S. E., and Abou-Bakr, S. (2020). Prospecting Genomic Regions Associated with Milk Production Traits in Egyptian Buffalo. *J. Dairy Res.* 87, 389–396. doi:10.1017/S0022029920000953

Barłowska, J., Szwajkowska, M., Litwińczuk, Z., and Król, J. (2011). Nutritional Value and Technological Suitability of Milk from Various Animal Species Used for Dairy Production. *Compr. Rev. Food Sci. Food Saf.* 10, 291–302. doi:10.1111/j.1541-4337.2011.00163.x

Bionaz, M., and Loor, J. J. (2008). ACSL1, AGPAT6, FABP3, LPIN1, and SLC27A6 Are the Most Abundant Isoforms in Bovine Mammary Tissue and Their Expression Is Affected by Stage of Lactation. *J. Nutr.* 138 (6), 1019–1024. doi:10.1093/jn/138.6.1019

Boutinaud, M., Guinard-Flament, J., and HélèneJammes, H. (2004). The Number and Activity of Mammary Epithelial Cells, Determining Factors for Milk Production. *Reprod. Nutr. Dev.* 44 (5), 499–508. doi:10.1051/rnd:2004054

Cavaletto, M., Giuffrida, M. G., and Conti, A. (2008). Milk Fat Globule Membrane Components-Aa Proteomic Approach. *Adv. Exp. Med. Biol.* 606, 129–141. doi:10.1007/978-0-387-74087-4_4

Chen, W.-C., Wang, C.-Y., Hung, Y.-H., Weng, T.-Y., Yen, M.-C., and Lai, M.-D. (2016). Systematic Analysis of Gene Expression Alterations and Clinical Outcomes for Long-Chain Acyl-Coenzyme A Synthetase Family in Cancer. *PLoS One* 11 (5), e0155660. doi:10.1371/journal.pone.0155660

Cosenza, G., Gallo, D., Auzino, B., Gaspa, G., and Pauciullo, A. (2021). Complete CSN1S2 Characterization, Novel Allele Identification and Association with Milk Fatty Acid Composition in River Buffalo. *Front. Genet.* 11, 622494. doi:10.3389/fgene.2020.622494

Deng, T., Liang, A., Liu, J., Hua, G., Ye, T., Liu, S., et al. (2018). Genome-Wide SNP Data Revealed the Extent of Linkage Disequilibrium, Persistence of Phase and Effective Population Size in Purebred and Crossbred Buffalo Populations. *Front. Genet.* 9, 688. doi:10.3389/fgene.2018.00688

Du, C., Deng, T., Zhou, Y., Ye, T., Zhou, Z., Zhang, S., et al. (2019). Systematic Analyses for Candidate Genes of Milk Production Traits in Water Buffalo (Bubalus Bubalis). *Anim. Genet.* 50 (3), 207–216. doi:10.1111/age.12739

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011). Using the Genomic Relationship Matrix to Predict the Accuracy of Genomic Selection. *J. Anim. Breed. Genet.* 128 (6), 409–421. doi:10.1111/j.1439-0388.2011.00964.x

Hayes, B. J., and Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* 7, 89–102. doi:10.1146/annurev-animal-020518-115024

He, Q., Luo, J., Wu, J., Yao, W., Li, Z., Wang, H., et al. (2020). FoxO1 Knockdown Promotes Fatty Acid Synthesis via Modulating SREBP1 Activities in the Dairy Goat Mammary Epithelial Cells. *J. Agric. Food Chem.* 68 (43), 12067–12078. doi:10.1021/acs.jafc.0c05237

Jiang, Y., Xie, M., Fan, W., Xue, J., Zhou, Z., Tang, J., et al. (2019). Transcriptome Analysis Reveals Differential Expression of Genes Regulating Hepatic Triglyceride Metabolism in Pekin Ducks during Dietary Threonine Deficiency. *Front. Genet.* 10, 710. doi:10.3389/fgene.2019.00710

Li, L. O., Klett, E. L., and Coleman, R. A. (2010). Acyl-CoA Synthesis, Lipid Metabolism and Lipotoxicity. *Biochim. Biophys. Acta (BBA) - Mol. Cell Biol. Lipids* 1801 (3), 246–251. doi:10.1016/j.bbalip.2009.09.024

Li, Q., Tao, Z., Shi, L., Ban, D., Zhang, B., Yang, Y., et al. (2012). Expression and Genome Polymorphism of ACSL1 Gene in Different Pig Breeds. *Mol. Biol. Rep.* 39 (9), 8787–8792. doi:10.1007/s11033-012-1741-6

Li, J., Liang, A., Li, Z., Du, C., Hua, G., Salzano, A., et al. (2017). An Association Analysis between PRL Genotype and Milk Production Traits in Italian Mediterranean River Buffalo. *J. Dairy Res.* 84 (4), 430–433. doi:10.1017/s0022029917000693

Li, J., Liu, S., Li, Z., Zhang, S., Hua, G., Salzano, A., et al. (2018). DGAT1 Polymorphism in Riverine Buffalo, Swamp Buffalo and Crossbred Buffalo. *J. Dairy Res.* 85 (4), 412–415. doi:10.1017/s0022029918000468

Li, Q., Li, Y., Shi, L. J., Gao, Y. X., Li, Q. F., Sun, D. X., et al. (2019). Analysis of Genetic Effects of FBP2 Gene on Milk Production Traits in Chinese Holstein (*Bos taurus*). *Chin. J. Agric. Biotechol.* 27 (09), 1582–1595. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=NYSB201909007&DbName=CJFQ2019

Li, J., Liu, J., Liu, S., Campanile, G., Salzano, A., Gasparrini, B., et al. (2020). Genome-wide Association Study for Buffalo Mammary Gland Morphology. *J. Dairy Res.* 87 (1), 27–31. doi:10.1017/s0022029919000967

Li, T., Li, X., Meng, H., Chen, L., and Meng, F. (2020). ACSL1 Affects Triglyceride Levels through the PPARγ Pathway. *Int. J. Med. Sci.* 17 (6), 720–727. doi:10.7150/ijms.42248

Lian, S., Guo, J. R., Nan, X. M., Ma, L., Loor, J. J., and Bu, D. P. (2016). MicroRNA Bta-miR-181a Regulates the Biosynthesis of Bovine Milk Fat by Targeting ACSL1. *J. Dairy Sci.* 99 (5), 3916–3924. doi:10.3168/jds.2015-10484

Liang, S. S., Pang, C. Y., Deng, T. X., Ma, X. Y., Lu, X. Y., and Liang, X. W. (2020). Expression of ACSL1 and its Effect on Expression Involved in Fatty Acid Metabolism in Buffalo. *Chin. J. Anim. Sci.* 56 (05), 41–45.

Liang, Y., Gao, Q., Zhang, Q., Arbab, A. A. I., Li, M., Yang, Z., et al. (2020). Polymorphisms of the ACSL1 Gene Influence Milk Production Traits and Somatic Cell Score in Chinese Holstein Cows. *Animals* 10 (12), 2282. doi:10.3390/ani10122282

Liu, R., Liu, X., Bai, X., Xiao, C., and Dong, Y. (2021). A Study of the Regulatory Mechanism of the CB1/PPARγ2/PLIN1/HSL Pathway for Fat Metabolism in Cattle. *Front. Genet.* 12, 631187. doi:10.3389/fgene.2021.631187

Liu, J., Wang, Z., Li, J., Li, H., and Yang, L. (2020). Genome-wide Identification of Diacylglycerol Acyltransferases (DGAT) Family Genes Influencing Milk Production in Buffalo. *BMC Genet.* 21 (1), 26. doi:10.1186/s12863-020-0832-y

Liu, Z., Li, C., Pryce, J., and Rochfort, S. (2020). Comprehensive Characterization of Bovine Milk Lipids: Phospholipids, Sphingolipids, Glycolipids, and Ceramides. *J. Agric. Food Chem.* 68 (24), 6726–6738. doi:10.1021/acs.jafc.0c01604

Ma, Y., Kanakousaki, K., and Buttitta, L. (2015). How the Cell Cycle Impacts Chromatin Architecture and Influences Cell Fate. *Front. Genet.* 6, 19. doi:10.3389/fgene.2015.00019

Ma, Y., Zha, J., Yang, X., Li, Q., Zhang, Q., Yin, A., et al. (2021). Long-chain Fatty Acyl-CoA Synthetase 1 Promotes Prostate Cancer Progression by Elevation of Lipogenesis and Fatty Acid Beta-Oxidation. *Oncogene* 40 (10), 1806–1820. doi:10.1038/s41388-021-01667-y

Manichaikul, A., Wang, X.-Q., Zhao, W., Wojczynski, M. K., Siebenthall, K., Stamatoyannopoulos, J. A., et al. (2016). Genetic Association of Long-Chain Acyl-CoA Synthetase 1 Variants with Fasting Glucose, Diabetes, and Subclinical Atherosclerosis. *J. lipid Res.* 57 (3), 433–442. doi:10.1194/jlr.m064592

Monks, J., C., Smith-Steinhart, E. R., Kruk, V. A., Fadok, and P. M., Henson. (2008). Epithelial cells remove apoptotic epithelial cells during post-lactation involution of the mouse mammary gland. *Biol Reprod* 78(4): 586–594. doi:10.1095/biolreprod.107.065045

Pan, M.-H., Chang, Y.-H., Badmaev, V., Nagabhushanam, K., and Ho, C.-T. (2007). Pterostilbene Induces Apoptosis and Cell Cycle Arrest in Human Gastric Carcinoma Cells. *J. Agric. Food Chem.* 55 (19), 7777–7785. doi:10.1021/jf071520h

Pauciullo, A., Cosenza, G., Steri, R., Coletta, A., Jemma, L., Feligini, M., et al. (2012). An Association Analysis between OXT Genotype and Milk Yield and Flow in Italian Mediterranean River Buffalo. *J. Dairy Res.* 79 (2), 150–156. doi:10.1017/s0022029911000914

Qiu, Y., Qu, B., Zhen, Z., Yuan, X., Zhang, L., and Zhang, M. (2019). Leucine Promotes Milk Synthesis in Bovine Mammary Epithelial Cells via the PI3K-DDX59 Signaling. *J. Agric. Food Chem.* 67 (32), 8884–8895. doi:10.1021/acs.jafc.9b03574

Rose, A. B. (2008). Intron-mediated Regulation of Gene Expression. *Curr. Top. Microbiol. Immunol.* 326, 277–290. doi:10.1007/978-3-540-76776-3_15

Rossi Sebastiano, M., and Konstantinidou, G. (2019). Targeting Long Chain Acyl-CoA Synthetases for Cancer Therapy. *Ijms* 20 (15), 3624. doi:10.3390/ijms20153624

Rubinow, K. B., Wall, V. Z., Nelson, J., Mar, D., Bomsztyk, K., Askari, B., et al. (2013). Acyl-CoA Synthetase 1 Is Induced by Gram-Negative Bacteria and Lipopolysaccharide and Is Required for Phospholipid Turnover in Stimulated Macrophages. *J. Biol. Chem.* 288 (14), 9957–9970. doi:10.1074/jbc.m113.458372

Schutz, F. A., Pomerantz, M. M., Gray, K. P., Atkins, M. B., Rosenberg, J. E., Hirsch, M. S., et al. (2013). Single Nucleotide Polymorphisms and Risk of Recurrence of

Renal-Cell Carcinoma: a Cohort Study. *Lancet Oncol.* 14 (1), 81–87. doi:10.1016/s1470-2045(12)70517-x

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, H., Zhang, T., Li, C., Wang, J., Huang, J., and Li, Z. (2017). trans-10,cis-12-Conjugated Linoleic Acid Affects Expression of Lipogenic Genes in Mammary Glands of Lactating Dairy Goats. *J. Agric. Food Chem.* 65 (43), 9460–9467. doi:10.1021/acs.jafc.7b02377

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *Am. J. Hum. Genet.* 68 (4), 978–989. doi:10.1086/319501

Wang, Z., Hou, X., Qu, B., Wang, J., Gao, X., and Li, Q. (2014). Pten Regulates Development and Lactation in the Mammary Glands of Dairy Cows. *PLoS One* 9 (7), e102118. doi:10.1371/journal.pone.0102118

Wang, Y., Cai, X., Zhang, S., Cui, M., Liu, F., Sun, B., et al. (2017). HBXIP Up-Regulates ACSL1 through Activating Transcriptional Factor Sp1 in Breast Cancer. *Biochem. Biophys. Res. Commun.* 484 (3), 565–571. doi:10.1016/j.bbrc.2017.01.126

Wang, R. J., Gao, X. F., Yang, J., and Kong, X. R. (2019). Genome-Wide Association Study to Identify Favorable SNP Allelic Variations and Candidate Genes that Control the Timing of Spring Bud Flush of Tea (Camellia Sinensis) Using SLAF-Seq. *J. Agric. Food Chem.* 67 (37), 10380–10391. doi:10.1021/acs.jafc.9b03330

Wang, C., Lv, X., He, C., Davis, J. S., Wang, C., and Hua, G. (2020). Four and a Half LIM Domains 2 (FHL2) Contribute to the Epithelial Ovarian Cancer Carcinogenesis. *Int. J. Mol. Sci.* 21 (20), 7751. doi:10.3390/ijms21207751

Widmann, P., Nuernberg, K., Kuehn, C., and Weikard, R. (2011). Association of an ACSL1 Gene Variant with Polyunsaturated Fatty Acids in Bovine Skeletal Muscle. *BMC Genet.* 12, 96. doi:10.1186/1471-2156-12-96

Wilkening, S., Chen, B., Bermejo, J. L., and Canzian, F. (2009). Is There Still a Need for Candidate Gene Approaches in the Era of Genome-wide Association Studies? *Genomics* 93 (5), 415–419. doi:10.1016/j.ygeno.2008.12.011

Ye, T., Deng, T., Hosseini, S. M., Raza, S. H. A., Du, C., Chen, C., et al. (2021). Association Analysis between FASN Genotype and Milk Traits in Mediterranean Buffalo and its Expression Among Different Buffalo Tissues. *Trop. Anim. Health Prod.* 53 (3), 366. doi:10.1007/s11250-021-02713-3

Zhao, L., Pascual, F., Bacudio, L., Suchanek, A. L., Young, P. A., Li, L. O., et al. (2019). Defective Fatty Acid Oxidation in Mice with Muscle-specific Acyl-CoA Synthetase 1 Deficiency Increases Amino Acid Use and Impairs Muscle Function. *J. Biol. Chem.* 294 (22), 8819–8833. doi:10.1074/jbc.ra118.006790

Zhao, Z., Abbas Raza, S. H., Tian, H., Shi, B., Luo, Y., Wang, J., et al. (2020). Effects of Overexpression of ACSL1 Gene on the Synthesis of Unsaturated Fatty Acids in Adipocytes of Bovine. *Arch. Biochem. Biophys.* 695, 108648. doi:10.1016/j.abb.2020.108648

**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.

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fgene.2022.862838

# Genetic Characterization and Population Connectedness of North American and European Dairy Goats

**Marc Teissier**[1]*, **Luiz F. Brito**[2,3], **Flavio S. Schenkel**[3], **Guido Bruni**[4], **Pancrazio Fresi**[5], **Beat Bapst**[6], **Christèle Robert-Granie**[1] **and Hélène Larroque**[1]

[1]GenPhySE, Université de Toulouse, Toulouse, France, [2]Department of Animal Sciences, Purdue University, West Lafayette, IN, United States, [3]Department of Animal Biosciences, Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada, [4]ARAL, Crema, Italy, [5]AssoNaPa, Roma, Italy, [6]Qualitas AG, Zug, Switzerland

**Edited by:**
*Jesús Fernández,*
*Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain*

**Reviewed by:**
*George R. Wiggans,*
*Council on Dairy Cattle Breeding, United States*
*Manuel Ramón*

***Correspondence:**
*Marc Teissier*
*marc.teissier@inrae.fr*

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

**Received:** 26 January 2022
**Accepted:** 03 May 2022
**Published:** 17 June 2022

**Citation:**
*Teissier M, Brito LF, Schenkel FS, Bruni G, Fresi P, Bapst B, Robert-Granie C and Larroque H (2022) Genetic Characterization and Population Connectedness of North American and European Dairy Goats. Front. Genet. 13:862838. doi: 10.3389/fgene.2022.862838*

Genomic prediction of breeding values is routinely performed in several livestock breeding programs around the world, but the size of the training populations and the genetic structure of populations evaluated have, in many instances, limited the increase in the accuracy of genomic estimated breeding values. Combining phenotypic, pedigree, and genomic data from genetically related populations can be a feasible strategy to overcome this limitation. However, the success of across-population genetic evaluations depends on the pedigree connectedness and genetic relationship among individuals from different populations. In this context, this study aimed to evaluate the genetic connectedness and population structure of Alpine and Saanen dairy goats from four countries involved in the European project SMARTER (SMAll RuminanTs Breeding for Efficiency and Resilience), including Canada, France, Italy, and Switzerland. These analyses are paramount for assessing the potential feasibility of an across-country genomic evaluation in dairy goats. Approximately, 9,855 genotyped individuals (with 51% French genotyped animals) and 6,435,189 animals included in the pedigree files were available across all four populations. The pedigree analyses indicated that the exchange of breeding animals was mainly unilateral with flows from France to the other three countries. Italy has also imported breeding animals from Switzerland. Principal component analyses (PCAs), genetic admixture analysis, and consistency of the gametic phase revealed that French and Italian populations are more genetically related than the other dairy goat population pairs. Canadian dairy goats showed the largest within-breed heterogeneity and genetic differences with the European populations. The genetic diversity and population connectedness between the studied populations indicated that an international genomic evaluation may be more feasible, especially for French and Italian goats. Further studies will investigate the accuracy of genomic breeding values when combining the datasets from these four populations.

**Keywords: dairy goats, genetic diversity, population structure, small ruminants, Alpine goats, Saanen goats**

# 1 INTRODUCTION

Genomic prediction of breeding values is routinely performed in several livestock species, including dairy and beef cattle, dairy sheep, and dairy goats (Boichard et al., 2012; Carillier et al., 2013; Baloche et al., 2014; Ibanez-Escriche and Simianer, 2016; Rupp et al., 2016). Genomic selection has become possible due to the availability of a large enough training population (individuals with both genotypes and phenotypes for the traits of interest) genotyped for thousands of genomic markers. However, the success of these genomic predictions depends on population-specific parameters, including the effective population size, level of linkage disequilibrium (LD), genetic relationship between the training and target populations, pedigree connectedness, and trait heritability (Misztal et al., 2020; van den Berg et al., 2020; VanRaden, 2020). For instance, a single-nucleotide polymorphism (SNP) chip panel of enough SNP density is required to capture the LD between quantitative trait loci (QTL) and surrounding markers and thus accurately estimates the SNP effects (de Roos et al., 2008; Lund et al., 2011). The size of the training populations and the pedigree connectedness also play a major role in the accuracy of genomic predictions (Lund et al., 2011; VanRaden, 2020), and lower-heritability traits require an even larger training population (Pszczola et al., 2012).

Combining data from genetically related populations can be an efficient strategy for enlarging training populations for genomic predictions (Berry et al., 2014; Cardoso et al., 2021). For instance, this has been performed in European dairy cattle populations through the Eurogenomics Consortium (www.eurogenomics.com/), which maintains a training population of ~40,000 genotyped bulls and provides genomic estimated breeding values (GEBVs) for 11 countries. More recently, international genomic evaluations have also been implemented in beef cattle populations (Bonifazi et al., 2020). In general, the method chosen to conduct these analyses is the multi-trait single-step genomic best linear unbiased prediction [ssGBLUP (Bonifazi et al., 2020)], in which the same trait measured across countries is considered different, but genetically correlated, traits. International genomic evaluations have been successfully implemented in international beef and dairy cattle populations. However, the success of across-population genomic evaluations requires a close collaboration between the partners and close population structure and genetic connectedness among the involved populations. For instance, the level of genetic connectedness (as a consequence of the exchange of genetic material) between the different populations needs to be sufficient to obtain accurate genomic prediction (Weigel et al., 2000; Fouilloux et al., 2006). Furthermore, combining data from several populations is only feasible if they are genetically related (Lund et al., 2014; Rezende et al., 2020). However, recent studies in Norwegian and New Zealand sheep with similar development history, but reduced recent exchange of genetic material, have reported that collaborative genomic analyses could still be feasible (Oliveira et al., 2020; Oliveira et al., 2022).

Currently, genomic evaluations have been implemented in dairy goats in France (Carillier et al., 2013) and tested in Canada (Massender et al., 2022) for both Alpine and Saanen breeds. GEBVs are more accurate than pedigree-based EBVs (Carillier et al., 2013; Carillier et al., 2014; Massender et al., 2022), but the observed gains in accuracy are still lower than dairy cattle. This is likely due to specific population characteristics such as the smaller size of the training populations and higher genetic diversity in dairy goats (Carillier et al., 2013; Brito et al., 2015). Combining data from different countries could contribute to improving the accuracy of genomic predictions by increasing the size of the training populations for economically important traits. Furthermore, across-country genomic predictions could be even more beneficial to countries that do not currently carry out genomic evaluations, such as Italy and Switzerland. Therefore, there is a need to assess the genetic connectedness and population structure of dairy goats from France, Italy, Canada, and Switzerland to evaluate the feasibility of an across-country genomic evaluation. In this context, the main objectives of this study were 1) to investigate the historical exchanges of genetic material between these four countries based on pedigree recording (genetic connectedness) and 2) to evaluate the genomic relatedness of these four populations based on genome-wide levels of LD, consistency of the gametic phase across population pairs, principal component analysis (PCA), and population admixture analyses. These analyses are paramount for assessing the potential feasibility of an across-country genomic evaluation in dairy goats.

# 2 MATERIALS AND METHODS

## 2.1 Pedigree and Genomic Datasets

This study was carried out in the framework of the "practical selection tools to benefit from international harmonisation and cooperation" work package of the SMARTER project (www.smarterproject.eu/). Four countries (Canada, France, Italy, and Switzerland) have shared 9,941 raw genotypes and pedigree information from Alpine and Saanen dairy goat populations. The animal identification (ID) was standardized in each country partner and was formed based on four components: three letters indicating the breed of the animal (ALP for Alpine and SAA for Saanen) + three letters indicating the country of origin (CAN, FRA, ITA, and CHE representing Canada, France, Italy, and Switzerland, respectively) + one letter indicating the sex of the animal (F for female and M for male) + 16 characters with the animal identifier (including the animal birth country code in two letters and the remaining characters after adding the animal ID completed on the left side by as many 0 as needed). For instance, the final identification of an Alpine female with local ID 5248383, born in France, and raised in Switzerland would be ALPCHEF0000000FR5248383. Imported animals may have multiple identifiers (one from the country of origin and another one in the importing country). Therefore, up to three aliases could be provided by the partners in addition to the ID of the animal. This identification is important to enable tracing the origin of the curated data but also useful for finding the connections between the different pedigrees.

**TABLE 1 |** Number of animals after the quality control, per breed (Alpine and Saanen), included in the pedigree and genotype files shared by each country (Canada, France, Italy, and Switzerland).

| Country | Alpine | | Saanen | | Total pedigree | Total genotype |
|---|---|---|---|---|---|---|
| | Pedigree | Genotypes | Pedigree | Genotypes | | |
| Canada | 56,601 | 793 | 36,741 | 903 | 93,342 | 1,696 |
| France | 3,518,473 | 2,968 | 2,527,443 | 2,009 | 6,045,916 | 4,977 |
| Italy | 107,566 | 1,061 | 131,376 | 338 | 238,942 | 1,399 |
| Switzerland | 28,083 | 1,280 | 28,906 | 503 | 56,989 | 1,783 |
| Total | 3,710,723 | 6,102 | 2,724,466 | 3,753 | 6,435,189 | 9,855 |

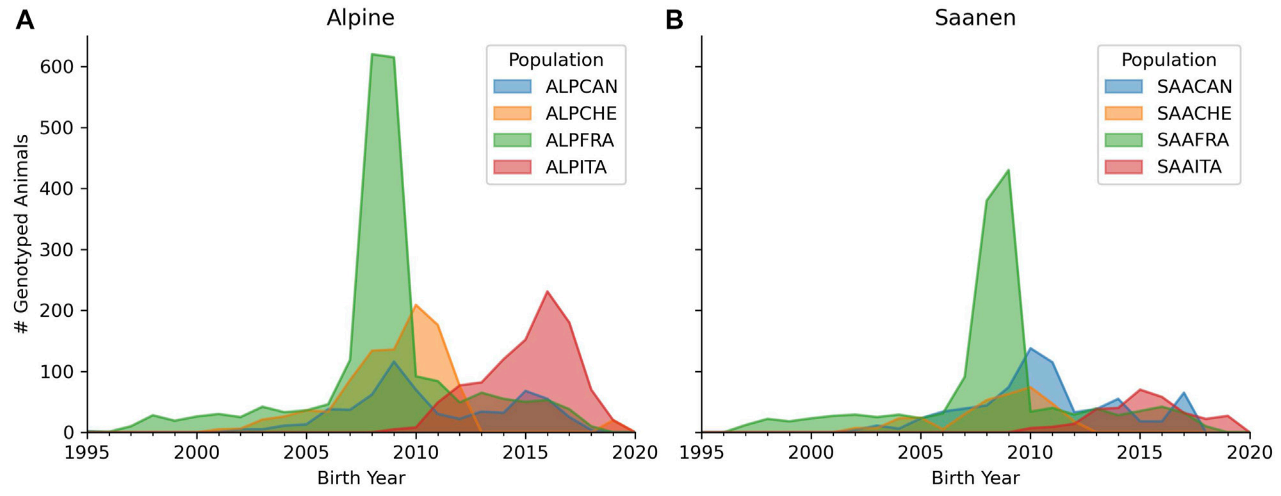


**FIGURE 1 |** Number of genotyped animals according to the birth year for Alpine **(A)** and Saanen **(B)** breeds in each country (France, Canada, Italy, and Switzerland). The legend represents the breed (ALP for Alpine and SAA for Saanen) and country (CAN for Canada, CHE for Switzerland, FRA for France, and ITA for Italy).

Various quality control filters were implemented in these datasets. First, the format of each animal's identification included in the pedigree files was verified for consistency, including checking that all the animals present as sires or dams were also registered as individuals in the pedigree. After removing or correcting these inconsistencies, 6,435,189 animals remained in the pedigree files (**Table 1**). The pedigree file had 86%, 89%, 91%, and 94% females in Canada, Switzerland, France, and Italy, respectively, which were born between 1944 and 2020. The males were born from 1936 to 2020.

All the individuals were genotyped using the same SNP chip panel, i.e., Goat SNP50 BeadChip (Illumina Inc., San Diego, CA, United States). There are currently two versions of this SNP chip panel, but 90% of the genotyping was performed based on the first version that contains 53,347 SNPs. Genotypes were exchanged in the TOP/BOT format and based on the ARS1 reference genome. As more than 90% of the genotyping was carried out based on the first version of the SNP chip panel and all SNPs included in the version 1 ($n$ = 53,347) were also present in version 2, only the SNPs from version 1 were considered for further analyses. Duplicated genotypes were filtered out based on the animal call rate, in which the genotype sample with a higher call rate was kept in the dataset. SNPs with a minor allele frequency (MAF) lower than 0.01 and a call rate lower than 0.90 were filtered out. Furthermore, animals with a sample call rate lower than 0.90 were also removed from the analyses ($n$ = 86). Quality control was performed within the breed and country but also after merging the four datasets. The quality control analyses were performed by PLINK 1.9 software (Purcell et al., 2007). After quality control, 9,855 animals and 50,578 SNPs remained for further analyses (**Table 1**).

**Figure 1** shows a density plot of the birth years of Alpine- and Saanen-genotyped animals in each country. An important point to highlight is that the genotyping activities did not start at the same time across partners. The oldest genotyped animals were born in 1997, 2000, 2001, and 2009 for French, Swiss, Canadian, and Italian goats, respectively.

## 2.2 Pedigree Connectedness

The pedigree connection evaluations were conducted by pairs of countries, comparing a source pedigree and a target pedigree. The goal was to extract animals from the source country in the target pedigree and seek to find them in the source pedigree (for example, French animals from the Swiss pedigree were found in the French pedigree). In total, 12 comparisons were made to find all the connections. These analyses were performed using Python scripts prepared by the authors.

The standardization of animal identification facilitated the extraction of foreign animals present in the other pedigree files. Several strategies were developed to retrieve the pedigrees of these animals. The simplest approach was to compare the identifiers (and aliases) of these animals with the source pedigree (e.g., France). This step was easily automated but not sufficient to find all the pedigree connections. For instance, considering the Swiss dairy goat pedigree, some French animals were registered in Switzerland with only the last digits of the French identifiers. For these animals, we used the fuzzy string-matching approaches (with the fuzzywuzzy library in Python; https://pypi.org/project/fuzzywuzzy/) to find the matches between the two pedigrees. The verification of the proposed animal matches based on this approach was carried out manually. This approach enabled the identification of animals with typos at the time of registration.

## 2.3 Characterization of Genetic Diversity

### 2.3.1 Linkage Disequilibrium

The extent of LD was calculated for each breed both within each country and also with merged datasets. This was determined based on the –$r^2$ option implemented in the PLINK 1.9 software (Purcell et al., 2007). The $r^2$ statistic was calculated as $\frac{(p_{AB}-p_A p_B)^2}{pA(1-p_1)p_b(1-p_B)}$, where $p_A$ and $p_B$ are the respective frequencies of alleles A and B (two different loci), respectively, and $p_{AB}$ is the frequency of the haplotype AB, as proposed by Hill and Robertson (1968). The LD between markers was measured for each pair of SNPs within a chromosome. The distance between two SNPs ranging from 0 to 1 Mb was categorized into 50 classes of 20 kb. The average LD was obtained by calculating the average $r^2$ for each class. In the **Sections 3**, **4**, each class was named based on the median distance in each interval. The LD decay plots were also created for each breed within the country.

### 2.3.2 Consistency of the Gametic Phase

The calculation of consistency of the gametic phase was determined following Oliveira et al. (2020) by first calculating the square roots of the $r^2$ statistic and then adding the sign of the D-value obtained with the dprime-signed option of the PLINK 1.9 software (Purcell et al., 2007). The consistency of the gametic phase was obtained as the Pearson correlation coefficient calculated between the signed-squared-root values of each country pair within the breed and between the signed-squared-root values across the two breeds within the country when grouping the two breeds together. The consistency of the gametic phase was also calculated for nine categories of SNPs according to their distance: (0 kb, 1 kb], (1 kb, 10 kb], (10 kb, 20 kb], (20 kb, 40 kb], (40 kb, 60 kb], (60 kb, 100 kb], (100 kb, 200 kb], (200 kb, 500 kb], and (500 kb, 1,000 kb]. We used the same interval classes as those presented by Mdladla et al. (2016).

### 2.3.3 Inbreeding Estimation

In addition to linkage disequilibrium and consistency of the gametic phase, we investigated inbreeding of genotyped animals. Pedigree-based inbreeding was estimated by inbupgf90 software (Misztal et al., 2002). Genomic inbreeding was estimated in two steps by PLINK 1.9 software (Purcell et al., 2007). The first step was to prune SNPs with the options –indep; then, inbreeding was estimated on the prune dataset with the options –het and –ibc.

### 2.3.4 Genetic Relatedness and Population Structure Analyses

The study of the genetic similarity and structure of the eight populations (two breeds x four countries) was performed based on two methods: principal component analysis (PCA) and genetic admixture analysis. To comply with the data independence assumption for performing PCA, the genotypes were pruned using the default parameter of the option -indep implemented in the PLINK 1.9 software (Purcell et al., 2007). A total of 31,951 SNPs were retained for the PCA analyses. PCA was performed using the -pca option of the PLINK 1.9 software (Purcell et al., 2007). The PCA was applied to the matrix of genomic relationships calculated as in Yang et al. (2011). The same pruned dataset was used to perform the admixture analysis using the Admixture software (Alexander et al., 2009). This software clusters individuals into k predefined groups according to allele frequencies (Oliveira et al., 2020). We tested k values ranging from 2 to 8 as it would be a more representative value of the expected number of subpopulations in our dataset. Only results with a k value equal to 4 will be presented because it yielded the lowest cross-validation error.

# 3 RESULTS

## 3.1 Pedigree Connectedness

**Table 2** describes the animals registered in several pedigrees for pairwise pedigree comparisons based on the animals' country of origin. Some animals could be identified as belonging to a country, but their pedigrees were not found in the country of origin. This scenario corresponds to the row "missing in local pedigree" in **Table 2**. We observed that only French and Swiss animals were found in several pedigree files. French animals were found in all pedigrees (Canada, France, Italy, and Switzerland), indicating that France exported animals to all country partners of the project. However, France did not import any animals from these countries. In contrast, Italian and Canadian animals were not exported to any other country based on the available recording. The pedigree comparison of these two countries shows that they have only French animals in common, which are all found in the French pedigree: 94 Alpine and 41 Saanen (**Table 2**). Italy was the only country that imported animals from both France (9,037 animals) and Switzerland (1,095 animals). In Italy, 1,863 French animals were not found in the French pedigree (859 Alpine and 1,004 Saanen). This number corresponds to 309 for French animals in Switzerland and 495 for Swiss animals in Italy.

Since the majority of animal exchanges occurred between France and the other three countries, we have identified Canadian, Italian, and Swiss animals with French parents to estimate the importance of their descendants in the host country. **Table 3** presents the number of Canadian, Italian, and Swiss animals with a French sire. In total, 17,137 Italian

**TABLE 2 |** Pedigree connectedness for Alpine and Saanen populations between pairs of four countries. Country abbreviations are CAN for Canada, CHE for Switzerland, FRA for France, and ITA for Italy. The native country is provided in each animal's name; it is possible to check if a foreign animal in a pedigree is found in its native pedigree (found in local pedigree) or not (missing in local pedigree).

| **Pairwise pedigree comparisons** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Status** | **Local origin of animal** | **Breed** | **CAN-FRA** | **CAN-ITA** | **CAN-CHE** | **FRA-ITA** | **CHE-FRA** | **CHE-ITA** | **All** |
| Found in local pedigree | CHE | ALP | 0 | 0 | 0 | 0 | 0 | 798 | 798 |
| | CHE | SAA | 0 | 0 | 0 | 0 | 0 | 297 | 297 |
| | FRA | ALP | 119 | 94 | 25 | 4,580 | 187 | 138 | 5,143 |
| | FRA | SAA | 61 | 41 | 9 | 4,457 | 135 | 85 | 4,788 |
| Missing in local pedigree | CHE | ALP | 0 | 0 | 0 | 0 | 0 | 305 | 305 |
| | CHE | SAA | 0 | 0 | 0 | 0 | 0 | 190 | 190 |
| | FRA | ALP | 0 | 0 | 0 | 859 | 215 | 0 | 1,074 |
| | FRA | SAA | 0 | 0 | 0 | 1,004 | 94 | 0 | 1,098 |
| | | All | 180 | 135 | 34 | 10,900 | 631 | 1,813 | 13,693 |

**TABLE 3 |** Number of Canadian (CAN), Italian (ITA), and Swiss (CHE) animals with a French (FRA) sire for the Alpine (ALP) and Saanen (SAA) breeds according to the sex of the animals (M for male and F for female). The proportion of animals relative to the native pedigree is given in parentheses.

| **French sire** | **ALP F (%)** | **ALP M (%)** | **SAA F (%)** | **SAA M (%)** | **Total (%)** |
|---|---|---|---|---|---|
| ITA | 5,396 (5.01) | 1,821 (1.69) | 8,216 (6.25) | 1,704 (1.29) | 17,137 (7.2) |
| CHE | 374 (1.33) | 107 (0.38) | 305 (1.05) | 93 (0.32) | 879 (1.54) |
| CAN | 276 (0.48) | 167 (0.29) | 69 (0.18) | 30 (0.08) | 542 (0.58) |
| Total | 6,046 (0.16) | 2,095 (0.05) | 8,590 (0.31) | 1,827 (0.07) | 18,558 (0.29) |

**TABLE 4 |** Count of genotyped animals recorded in several countries for each national pedigree independently of the origin of the animals. The proportion of these animals compared to their native pedigree is represented in the columns $\%_{\text{native\_pedigree}}$.

| **Native pedigree** | **ALP** | | **SAA** | |
|---|---|---|---|---|
| | **# Animal** | **$\%_{\text{native\_pedigree}}$** | **# Animal** | **$\%_{\text{native\_pedigree}}$** |
| ITA | 449 | 0.66 | 248 | 0.3 |
| CHE | 97 | 0.38 | 15 | 0.06 |
| CAN | 53 | 0.67 | 35 | 0.83 |
| FRA | 388 | 0.02 | 258 | 0.02 |

**TABLE 5 |** Average and standard error of pedigree-based and genomic inbreeding within each country and breed for genotyped animals.

| | **ALP** | | **SAA** | |
|---|---|---|---|---|
| | **Pedigree** | **Genomic** | **Pedigree** | **Genomic** |
| ITA | 0.021 (0.045) | 0.073 (0.049) | 0.036 (0.048) | 0.127 (0.044) |
| CAN | 0.015 (0.032) | 0.116 (0.058) | 0.048 (0.042) | 0.095 (0.078) |
| CHE | 0.019 (0.036) | 0.042 (0.039) | 0.015 (0.027) | 0.078 (0.034) |
| FRA | 0.022 (0.012) | 0.084 (0.034) | 0.025 (0.013) | 0.121 (0.039) |

animals had a French sire, which represented 7.2% of the Italian pedigree. This proportion was lower for the Swiss (1.54%) and Canadian (0.58%) populations. For animals with a French dam, we observed lower numbers: 3,932 (1.6%) Italian animals, 101 (0.1%) Swiss animals, and 1 (0.0%) Canadian animals.

**Table 4** describes the number of animals that are both genotyped and present in at least two countries. There is some overlapping when counting animals across countries because French animals, for example, are present in more than two countries. The animal count ranged between 53 and 449 for Alpine breed and between 15 and 258 for Saanen breed. The number of genotyped animals used in several countries remains limited (less than 1% whatever the country) when compared to the native pedigree.

**Table 5** describes the average pedigree-based and genomic inbreeding observed for genotyped animals. For Alpine breed, the averaged pedigree inbreeding is close for Switzerland (0.019), France (0.022), and Italy (0.021) and lower in Canada (0.015). We observed different trends in Saanen with high inbreeding in Canada (0.048) and then in Italy (0.036), France (0.025), and Switzerland (0.015). The averaged genomic inbreeding is higher than pedigree inbreeding whatever the country for both breeds with differences (genomic – pedigree) from 0.023 (Alpine in Switzerland) to 0.101 (Alpine in Canada).

## 3.2 Linkage Disequilibrium

The average LD calculated in Alpine (A) and Saanen (B), for each country separately and for multiple countries (ALP or SAA) or multiple breeds (All breeds) as a function of the SNP distance, is presented in **Figure 2**. For both Alpine and Saanen, the average LD was higher in Canadian than in the other goat populations. The average LD at 50 kb was 0.17 for Alpine and 0.19 for Saanen. The differences of LD values between Canada and the other

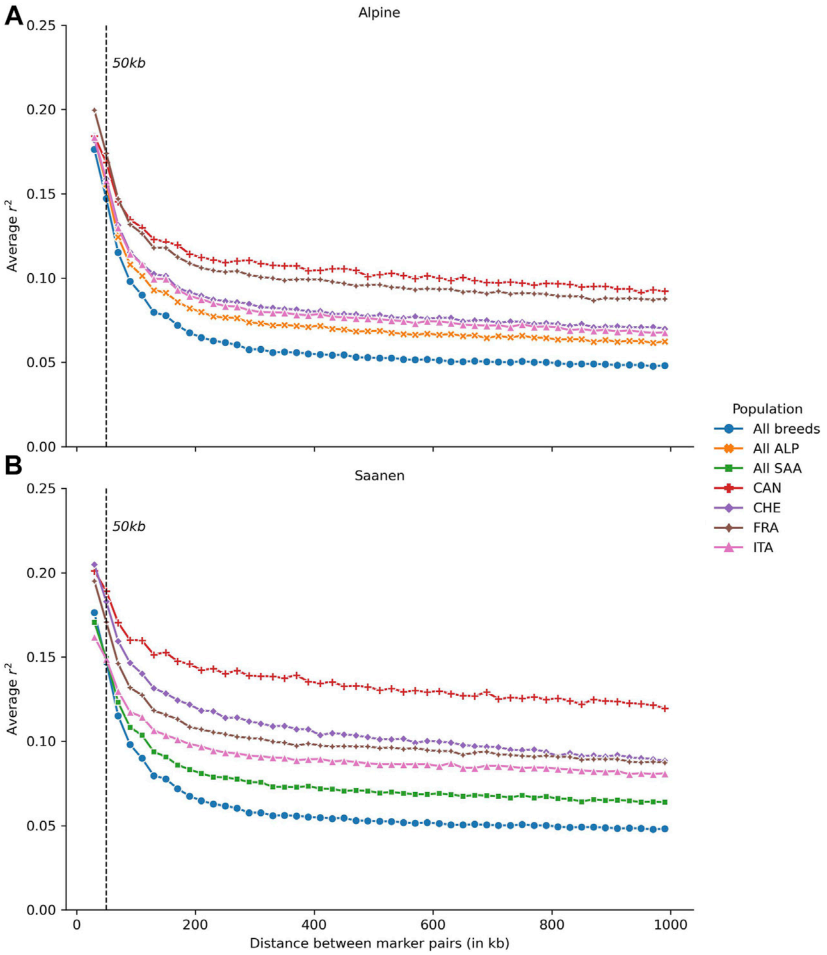


**FIGURE 2 |** Average linkage disequilibrium (LD) in **(A)** Alpine (ALP) and **(B)** Saanen (SAA) breeds, according to the distance between SNPs for each country evaluated: Canada (CAN), Switzerland (CHE), France (FRA), and Italy (ITA) and Saanen from the four countries together (All SAA), Alpine from the four countries together (All ALP), and both Saanen and Alpine goats from the four countries (All animals).

countries were higher for the Saanen breed. For the Alpine breed, the average LD at 50 kb ranged between 0.16 (Italy) and 0.17 (France and Canada). The average LD was quite close between Canada and France, regardless of the distance between SNPs and the $r^2$ values stabilized around 0.10–1 Mb. The average LD for the Swiss and Italian populations was also very similar and stabilized around 0.07–1 Mb.

For the Saanen breed, the range of LD values at 50 kb was wider than in the Alpine breed, with an average LD at 50 kb between 0.15 (Italy) and 0.19 (Canada). Canadian populations had a higher LD than in the other countries, regardless of the distance between SNPs. For short distances, LD values for Canadian and Swiss populations were close (0.18 and 0.19 at 50 kb, respectively) before differentiating for distances greater than 90 kb. The maximum difference was observed at 810 kb with an average LD of 0.09 in Swiss and 0.12 in Canadian goats.

## 3.3 Consistency of the Gametic Phase

The consistency of the gametic phase according to nine classes of distances between SNPs is shown in **Figure 3**. **Figures 3A,B** present the consistency of the gametic phase between pairs of countries within the Alpine (A) and Saanen (B) breeds. **Figure 3C** presents the consistency of the gametic phase between the Alpine and Saanen breeds within each country. Within the Alpine breed (**Figure 3A**), the consistency of the gametic phase values was the highest between France and Italy and ranged from 1 (distance of (0, 1 kb]) to 0.67 (distance (500, 1,000 kb]). The lowest values were obtained when comparing Canadian and European populations (ALPCAN_ALPITA, ALPCHE_ALPCAN, and ALPFRA_ALPCAN). In this case, the consistency was on average 0.97 for a distance of [0, 1 kb] and dropped to 0.11 for a distance of [500, 1,000 kb]. The intermediate consistency of the gametic phase was obtained when comparing Switzerland to France or Italy (ALPCHE_ALPFRA and ALPCHE_ALPITA) with an average consistency of 0.97 for a distance of (0, 1 kb] and a drop to 0.17 for a distance of (500, 1,000 kb].

The trends observed in Alpine were also found in the Saanen breed (**Figure 3B**) but with slightly lower values than in the Alpine breed. Between France and Italy, the consistency of gametic phases varied between 0.99 for a distance of (0, 1 kb]

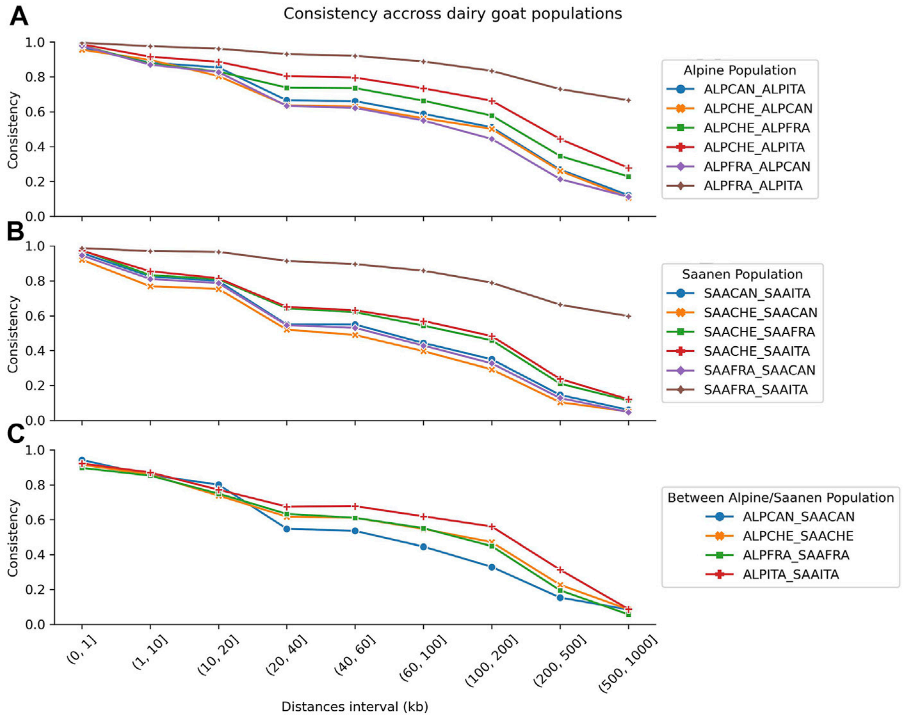


**FIGURE 3 |** Comparison of the consistency of the gametic phase for nine classes of distances between SNPs with comparison between Alpine populations **(A)**, Saanen populations **(B)**, and Alpine and Saanen from the same country **(C)**. Breeds are represented by Alpine (ALP) and Saanen (SAA), while countries are represented by Canada (CAN), Switzerland (CHE), France (FRA), and Italy (ITA).

and 0.60 for a distance of (500, 1,000 kb]. For a distance of (500, 1,000 kb], the consistency of the gametic phase values for all pairs of countries ranged between 0.06 and 0.60, while in Alpine, these values ranged from 0.11 to 0.67.

**Figure 3C** shows the consistency of the gametic phase within country when comparing Alpine and Saanen populations. The consistency of the gametic phase is similar for all countries for short distances with an average consistency of 0.92 for (0 kb, 1 kb], 0.86 for (1 kb, 10 kb], and 0.77 for (10 kb, 20 kb]. Then, the consistency between French and Swiss goat populations is similar across all distance intervals with an average difference of 0.01. The highest differences were observed between Canadian and Italian populations with an average difference of 0.10 across all distance intervals.

## 3.4 Principal Component Analysis

**Figure 4** presents the projection of each individual on the first two principal components of the PCA (PC1 and PC2). The first two components allow separate individuals according to their breed (PC1 3.26%), with the Alpine animals on the left and the Saanen on the right, and according to their country (PC2 2.32%), with the Canadian populations at the bottom and the European populations at the top. The French and Italian populations largely overlap and are indistinguishable for both breeds. The Canadian Saanen population is the most differentiated and does not group with the other Saanen populations. The few individuals present between the Canadian Saanen and the European populations are in fact animals with at least one French parent. The Swiss Saanen population is also more differentiated from the other European Saanen populations than the Alpine. Indeed, for the Alpine, there is an overlap of individuals for France, Italy, and Switzerland, which is not the case in Saanen with a more homogeneous cluster.

## 3.5 Admixture

The breed composition for each animal calculated with the Admixture software is shown in **Figure 5**. This analysis determines, for a given genotype, the proportion originating from each k ancestral cluster. The lowest cross-validation errors were observed when k was equal to 4. It was observed that the French and Italian populations have close and similar genetic background. For Alpine, on average, 0.89 of the genome of French goats and 0.72 of the genome of Italian goats come from the same ancestral cluster (orange color in **Figure 5**). This cluster is present to a lower extent in the Canadian (0.19) and Swiss (0.26) populations. On the other hand, there is very little present in Saanen (less than 0.10 for all populations). A second ancestral cluster (red color in **Figure 5**) is predominant in Saanen for French (0.88) and Italian (0.79) goats. This cluster is present at 0.41 in Switzerland for Saanen but is almost absent in Canadian Saanen (0.06).

The Canadian Saanen population seems to be largely different from the other Saanen groups. Indeed, the main ancestral cluster in Canadian Saanen (blue color in **Figure 5**) covers 0.82 of the genome, while it represents, on average, 0.11 for Swiss animals,

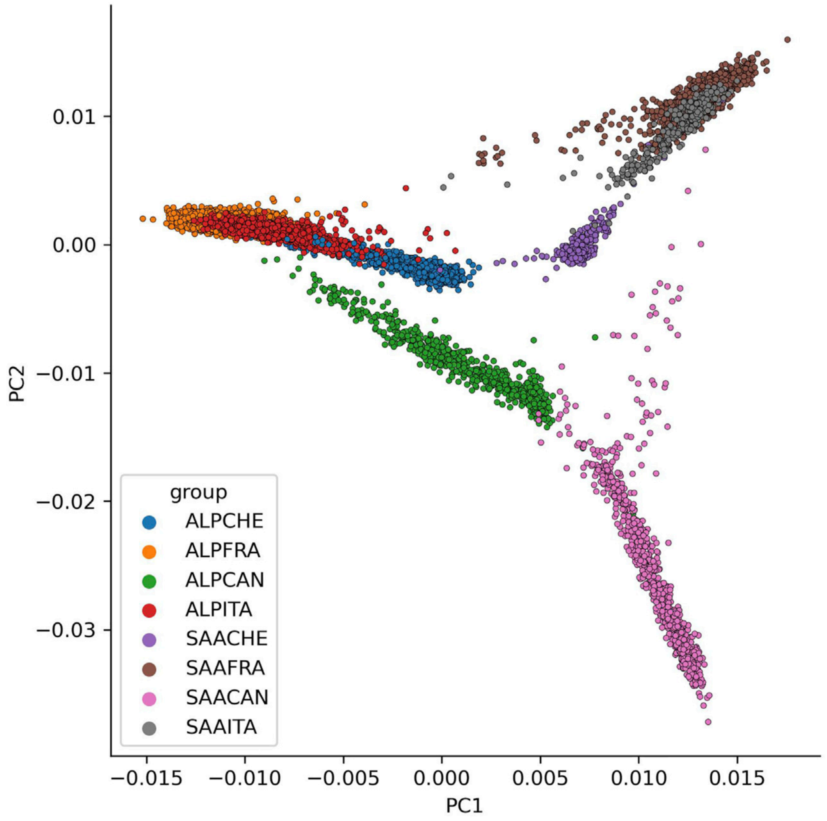


**FIGURE 4 |** Principal component analysis (PCA) with all genotypes for each breed (ALP, Alpine; SAA, Saanen) and country (CAN, Canada; FRA, France; ITA, Italy; CHE, Switzerland) on the two first PCA components (PC1 to PC2).

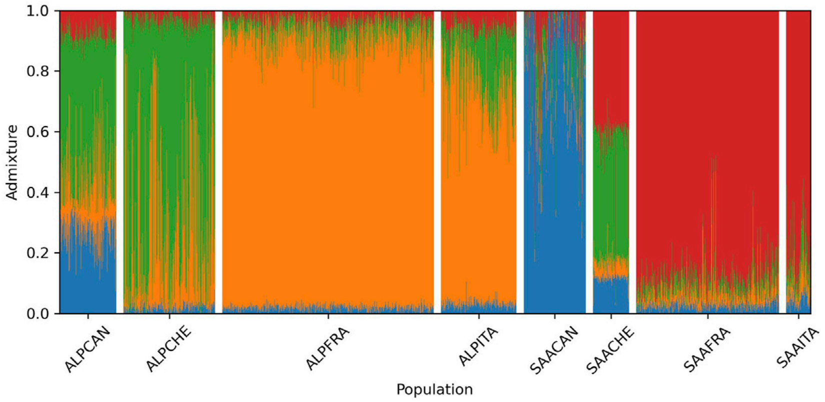


**FIGURE 5 |** Breed composition per animal for each breed-country population estimated by the Admixture software when considering k = 4 (ALP, Alpine; SAA, Saanen; CAN, Canada; FRA, France; ITA, Italy; CHE, Switzerland).

0.02 for French, and 0.04 for Italian animals. This blue cluster is also strongly represented (0.30) in the Canadian Alpine population. Another ancestral cluster (green color in **Figure 5**) also seems to be widely shared between Swiss Alpine (0.69), Canadian Alpine (0.46), and Swiss Saanen (0.41).

# 4 DISCUSSION

## 4.1 Pedigree Connectedness

The connections between populations coming from the four different countries based on their pedigree information are an essential parameter for a successful international genetic evaluation, especially when using the single-step GBLUP method. On the other hand, to simplify the creation of a unified pedigree, it is important to have a unique identifier for each animal, which did not exist in goat populations in this study (and which is also rarely the case for cattle and sheep breeds). Here, some of the pedigree connections have been found, but there is still work to be carried out because some original pedigree of foreign animals is still untraceable. The importance and difficulty of exhaustive research of pedigrees have been discussed in previous studies, such as in beef cattle for Interbeef (Venot et al., 2007), dogs (Wang, 2018), and race horses (Viklund et al., 2015).

We also have disproportional datasets with larger amounts of data in France in comparison to the other countries. This situation has also been reported in the framework of Interbeef for the Limousin cattle breed (Bonifazi et al., 2020), in which the numbers of French animals (2,942,297 animals) were higher than in the other countries (between 30,843 and 172,229 animals). The authors evaluated the within-country rankings of the top 100 animals for age-adjusted weaning weight (AWW) for both international and national evaluations. They observed that the majority of the animals in the top 100 were French (between 84% and 100%) for the international evaluations, while they vary between 19% and 77% (100% being obtained in France) for the national evaluations. This is a situation that can potentially be reproduced in the international for dairy goat evaluations and could encourage the disproportional use of French breeding stock. Moreover, trade between countries has been mostly one-sided, with France exporting to all partner countries. Therefore, more research needs to be conducted to elucidate the best options for short- and long-term international and national genomic evaluations for the partner countries to maximize the benefits of the collaboration. In addition, genotyped animals represent only a small portion of shared animals between countries. Strategies for improving connections between countries need to be considered before implementing a multi-country genetic evaluation.

Inbreeding estimated in our population was also consistent with previous estimation found in the literature. For example, in Canada, Brito et al. (2017) reported an average pedigree inbreeding equal to 0.021 in Alpine (against 0.015 in our study) and 0.040 in Saanen (against 0.048 in our study). In France, Carillier et al. (2013) reported a pedigree inbreeding of around 0.02 for both Alpine and Saanen breeds which is consistent with our estimation. For Italy and Switzerland, it seems that no report was available that estimates inbreeding in these populations.

## 4.2 Linkage Disequilibrium and Consistency of the Gametic Phase

Population parameters such as LD and consistency of the gametic phase have implications for the design of across-population genomic evaluations. For a multi-population (here multi-country) genetic evaluation to be effective, there should be equivalent LD between SNPs and QTLs in each country and a relatively high consistency of gametic phases between populations from different countries (Mohammad Rahimi et al., 2020). For the French Saanen breed, the LD at 50 kb estimated in this study (0.19) is slightly higher than that observed in the study of Carillier et al. (2013) (0.17). In contrast, for Alpine, similar estimates were obtained (0.17 at 50 kb). The difference observed for Saanen can be explained by the difference in the numbers of animals used to calculate the LD values, which could impact the accuracy of the estimates. In the study of Carillier et al. (2013), the calculation of LD was determined for the Alpine breed on 470 Alpine genotypes compared to 2,968 in our study. For the Saanen breed, our study was based on 2,009 genotyped animals compared to 355 in Carillier et al. (2013). For Canada, the study by Brito et al. (2015) estimated the LD at 55 kb of around 0.14 for both breeds. We obtained higher values with 0.17 for Alpine and 0.19 for Saanen, which are identical to the estimates found for these breeds in France. Several factors can explain these differences in the estimates. The number of genotyped animals has increased substantially (403 vs. 793 Alpine and 318 vs. 903 Saanen), which contributes to obtaining more accurate estimates. On the other hand, although in both studies LD was estimated based on the $r^2$ metric, the bins used to group the SNPs are different. Between 10 and 100 kb, Brito et al. (2015) created intervals of 10 kb, while we used wider intervals of 20 kb. To the best of our knowledge, no study has investigated the LD in Italy and Switzerland goat populations. Our study shows that Saanen populations from these countries have similar levels of LD in comparison to the French Saanen population. For the Alpine breed, the LD in Italian and Swiss populations is lower than in French Alpine. In any case, the level of LD is very close at 50 kb between populations and sufficient to consider genomic evaluation, as was demonstrated by Carillier et al. (2014). However, this level of LD will likely require larger training populations in comparison to less genetically diverse populations to obtain similar GEBV accuracies.

The consistency of the gametic phase is a key parameter for determining the effectiveness of a multi-population genetic evaluation (Biegelmeyer et al., 2016; Deng et al., 2019). This is the first time, to our knowledge, that the consistency of gametic phases is estimated between North American and European dairy goat populations. We observed that the French and Italian populations (Alpine and Saanen) have very high consistency of the gametic phase up to large distances between SNP pairs, indicating that a joint genomic evaluation might be feasible for these two countries. The consistency of gametic phase values is lower than the Canadian population with the European populations. This is also the case for Swiss when compared to French and Italian populations. This may make it more difficult to implement an international genetic evaluation across all the four countries. Deng et al. (2019) suggested that using a higher-density SNP chip panel could be an alternative for increasing the consistency of the gametic phase between SNP pairs (especially at shorter distances between SNPs). However, there are no high-density SNP chip panels available for goats. The availability of a second version of the Goat SNP50 BeadChip did not add enough SNP to get a significantly higher density of SNPs across all the goat genomes.

The consistency of the gametic phase in Alpine and Saanen breeds is similar within countries until the SNP distance of (10, 20 kb) with a decrease from about 0.92 to 0.62. After this distance, the decrease of the consistency of the gametic phase shows different trends with a higher level for Italian, a lower level for

Canadian, and an intermediate level for Swiss and French populations. For French animals, these results are in accordance with those of Carillier et al. (2013), with a decrease from 0.88 to 0.56 for marker distance <50 kb vs. 0.89–0.63 in our study. For Canadian populations, Brito et al. (2015) reported a Pearson correlation of 0.69 at 20 kb between Alpine and Saanen breeds, which is also consistent with our study. Carillier et al. (2014) have shown that in the case of the French populations, multi-breed or single-breed genomic evaluations yielded similar GEBV accuracies. However, the number of genotyped animals was significantly smaller in their study. In the context of an international genomic evaluation, the interest of a multi-breed multi-country genomic evaluation will have to be evaluated in comparison to a single-breed multi-country evaluation, which could significantly increase the training population size per breed. However, the current genotypes provided by the partners are both unbalanced in number and in the years of birth of the animals. In particular, between Italy and Switzerland for both breeds, there is almost no overlap in the birth year of the genotyped animals. This study is, in fact, the first one that was carried out on such data for these two countries. Further analyses should be performed with larger genotyped populations to confirm our findings.

## 4.3 PCA and Admixture

The results of the PCA and admixture analyses contribute to determining the genetic relationship of the animals, including the breed and country of origin. The only populations with no clear distinction are French and Italian goat populations for both breeds. Italy is the country that imports most animals from France, which may explain the genetic proximity between these two populations. In contrast, the Canadian and European populations are more genetically distant. This might be explained by the little exchange of animals and the geographical distance that separates Canada and the European countries. Finally, these results are consistent with the results observed on the connections between countries based on pedigree information.

Several genetic diversity studies have been conducted in goats. In France, the study of Oget et al., 2019 was conducted on eight French goat populations, but it included a few genotypes of Alpine (45) and Saanen (38) animals. Our results, with more genotyped animals, confirmed what has been previously shown for these two breeds. The French Alpine and Saanen populations are genetically different. A second study performed by Brito et al. (2015) compared genotypes from Alpine (403 animals) and Saanen (318 animals) from Canada and found that these two populations are genetically different.

The comparison of Alpine and Saanen genotypes within one country is well-documented more than international comparisons of these breeds. Denoyelle et al. (2021) is one of the few examples of an international comparison, which was carried out as part of the VarGoats project (www.goatgenome.org/vargoats.html). This project sequenced goats of different breeds from all over the world including Alpine and Saanen from France, Italy, and Switzerland. They studied the phylogeny of these breeds using a neighbor-joining tree constructed with 100,000 SNPs. For the Italian, Swiss, and French populations, our results are in agreement with their study, where a close relationship between France and Italy (for both Alpine and Saanen) and a greater distance with the Swiss goat population were observed.

The first two components of the PCA represent less than 6% of the total variation, which is quite limited. Even if we observe two different clusters between European and Canadian populations, these populations seemed close enough in order to blend all genotypes and to analyze genotypes conjointly. Further investigations on differences along the genome between animals from different countries could be interesting to detect genomic regions specifically selected in each country.

## 4.4 Implications of the Results and Next Steps

This work aimed to combine and analyze pedigree information and genomic data from four countries. Our analyses showed that an international evaluation would be most beneficial to the European populations that are genetically closer. However, it is necessary to verify the impact of Canadian data into international genomic evaluations, especially if other European dairy goat populations are added such as Yorkshire dairy goats (Mucha et al., 2015). Yorkshire goats represent a composite population potentially more similar with Canadian dairy goats due to more similar crossbreeding events. Pedigree connectedness and genotype analysis remain the first step before implementing an international genomic evaluation. The following steps will be to combine and analyze the phenotypes commonly recorded among the different country partners of the project. The joint analysis of phenotypes, pedigree, and genotypes will enable the estimation of genetic and genomic parameters between breeds/countries that will be potentially used in future genomic evaluations.

# 5 CONCLUSION

The genetic diversity and pedigree analyses performed in this study showed that the French and Italian populations are both the most genetically connected and more genomically similar. On the other hand, for the Swiss and Canadian dairy goat populations, the genetic connections are limited to the importation of a few French animals. Besides, they are genomically more distant than the other populations. The genetic diversity and population connectedness between the studied populations indicate that an international genomic evaluation might be more feasible for French and Italian goats. Further studies will investigate the accuracy of genomic breeding values when combining the datasets from these four populations.

# DATA AVAILABILITY STATEMENT

The genotypes were produced by several companies in each of the partner countries. These companies share these genotypes for research collaboration. Within the framework of the SMARTER project, agreements have been signed to supervise the use and publication of these genotypes. These agreements specify that the genotype data must remain private. However, upon request and research collaboration, they can be shared for specific use in research projects. Requests to access these datasets should be directed to MT, marc.teissier@inrae.fr.

## ETHICS STATEMENT

Ethical review and approval was not required for the animal study. Data were collected during routine care in each country. Written informed consent was obtained from the owners for the participation of their animals in this study.

## AUTHOR CONTRIBUTIONS

HL, LB, and CR-G designed the study. MT analyzed the data, wrote all Python scripts, and drafted the first version of the manuscript. LB, FS, HL, GB, and BB provided the datasets and information on the current genetic evaluations in their respective countries. MT, HL, LB, FS, and CR-G interpreted the results and improved the manuscript. All authors read and approved the final manuscript. The authors declare they do not have any competing interests.

## FUNDING

The first author received financial support from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement No. 772787 (SMARTER) and the French Research National Research Institute for Agriculture, Food and Environment (INRAE—Animal Genetic division).

## ACKNOWLEDGMENTS

The authors are grateful to the CTIG (Centre de Traitement de l'Information Génétique) of INRAE Jouy-en-Josas, France, for providing computing resources. The authors would also like to acknowledge Laurent Maignel from the Canadian Centre for Swine Improvement (CCSI, Ottawa, ON, Canada) and Erin Massender (University of Guelph, Guelph, Canada) for their help in preparing the Canadian phenotypic and genomic datasets. The Canadian partners acknowledge that this research received partial funding from the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA; Guelph, ON, Canada) through the Ontario Agri-Food Innovation Alliance. The authors acknowledge Capgenes and, in particular, Pierre Martin, as well as Virginie Clément (Idele-GenEval) and Isabelle Palhière (INRAE-GenPhySE), for providing and preparing French data. The Swiss partners would like to acknowledge Steffen Werne (Research Institute of Organic Agriculture FiBL, CH-5070 Frick, Switzerland), Franz Seefried (Qualitas AG, CH-6300 Zug, Switzerland), Cord Drögemüller (University of Bern, CH-3012 Bern, Switzerland, Switzerland), Christine Flury (Bern University of Applied Sciences, CH-3052 Zollikofen, Switzerland), and Erika Bangerter (Schweizerischer Ziegenzuchtverband, CH-3052 Zollikofen, Switzerland) for their contributions. The Italian partners would like to acknowledge Silverio Grande (AssoNaPa, 00187 Roma, Italy), Riccardo Negrini (AIA, 00187 Roma, Italy), Andrea Galli (ARAL, 26013 Crema, Italy), Stefano Biffani (IBBA—CNR, 20133 Milano, Italy), Alessio Negro (AssoNaPa, 00187 Roma, Italy), and Giovanni Bailo (ARAL, 26013 Crema, Italy) for their contributions.

## REFERENCES

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

Baloche, G., Legarra, A., Sallé, G., Larroque, H., Astruc, J.-M., Robert-Granié, C., et al. (2014). Assessment of Accuracy of Genomic Prediction for French Lacaune Dairy Sheep. *J. Dairy Sci.* 97, 1107–1116. doi:10.3168/jds.2013-7135

Berry, D. P., Coffey, M. P., Pryce, J. E., de Haas, Y., Løvendahl, P., Krattenmacher, N., et al. (2014). International Genetic Evaluations for Feed Intake in Dairy Cattle through the Collation of Data from Multiple Sources. *J. Dairy Sci.* 97, 3894–3905. doi:10.3168/jds.2013-7548

Biegelmeyer, P., Gulias-Gomes, C. C., Caetano, A. R., Steibel, J. P., and Cardoso, F. F. (2016). Linkage Disequilibrium, Persistence of Phase and Effective Population Size Estimates in Hereford and Braford Cattle. *BMC Genet.* 17, 1–12. doi:10.1186/s12863-016-0339-8

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., et al. (2012). Genomic Selection in French Dairy Cattle. *Anim. Prod. Sci.* 52, 115–120. doi:10.1071/AN11119

Bonifazi, R., Vandenplas, J., Ten Napel, J., Michenet, A., Cromie, A., Veerkamp, R. F., et al. (2020). "International Single-step Genomic Evaluations in Beef Cattle," in 71st Annual Meeting of the European Federation of Animal Science, Porto, Portugal, August 31 to September 4th, 2020, 579.

Brito, L. F., Jafarikia, M., Grossi, D. A., Kijas, J. W., Porto-Neto, L. R., Ventura, R. V., et al. (2015). Characterization of Linkage Disequilibrium, Consistency of Gametic Phase and Admixture in Australian and Canadian Goats. *BMC Genet.* 16, 67. doi:10.1186/s12863-015-0220-1

Brito, L. F., Kijas, J. W., Ventura, R. V., Sargolzaei, M., Porto-Neto, L. R., Cánovas, A., et al. (2017). Genetic Diversity and Signatures of Selection in Various Goat Breeds Revealed by Genome-wide SNP Markers. *BMC Genomics* 18, 1–20. doi:10.1186/s12864-017-3610-0

Cardoso, F. F., Matika, O., Djikeng, A., Mapholi, N., Burrow, H. M., Yokoo, M. J. I., et al. (2021). Multiple Country and Breed Genomic Prediction of Tick Resistance in Beef Cattle. *Front. Immunol.* 12, 2189. doi:10.3389/fimmu.2021.620847

Carillier, C., Larroque, H., Palhière, I., Clément, V., Rupp, R., and Robert-Granié, C. (2013). A First Step toward Genomic Selection in the Multi-Breed French Dairy Goat Population. *J. Dairy Sci.* 96, 7294–7305. doi:10.3168/jds.2013-6789

Carillier, C., Larroque, H., and Robert-Granié, C. (2014). Comparison of Joint versus Purebred Genomic Evaluation in the French Multi-Breed Dairy Goat Population. *Genet. Sel. Evol.* 46, 67. doi:10.1186/s12711-014-0067-3

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. *Genetics* 179, 1503–1512. doi:10.1534/genetics.107.084301

Deng, T., Liang, A., Liu, J., Hua, G., Ye, T., Liu, S., et al. (2019). Genome-wide SNP Data Revealed the Extent of Linkage Disequilibrium, Persistence of Phase and Effective Population Size in Purebred and Crossbred Buffalo Populations. *Front. Genet.* 9, 688. doi:10.3389/fgene.2018.00688

Denoyelle, L., Talouarn, E., Bardou, P., Colli, L., Alberti, A., Danchin, C., et al. (2021). VarGoats Project: a Dataset of 1159 Whole-Genome Sequences to Dissect *Capra hircus* Global Diversity. *Genet. Sel. Evol.* 53, 1–14. doi:10.1186/s12711-021-00659-6

Fouilloux, M.-N., Minery, S., Mattalia, S., and Laloë, D. (2006). *Assessment of Connectedness in the International Genetic Evaluation of Simmental and*

*Montbéliard Breeds*. INTERBULL meeting in 2006 from 4th to 6th June in Kuopio, Finland 35, 129.

Hill, W. G., and Robertson, A. (1968). Linkage Disequilibrium in Finite Populations. *Theor. Appl. Genet.* 38, 226–231. doi:10.1007/BF01245622

Ibanez-Escriche, N., and Simianer, H. (2016). From the Editors: Animal Breeding in the Genomics Era. *Anim. Front.* 6, 4–5. doi:10.2527/af.2016-0001

Lund, M. S., de Roos, A. P., de Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., et al. (2011). A Common Reference Population from Four European Holstein Populations Increases Reliability of Genomic Predictions. *Genet. Sel. Evol.* 43, 43. doi:10.1186/1297-9686-43-43

Lund, M. S., Su, G., Janss, L., Guldbrandtsen, B., and Brøndum, R. F. (2014). Genomic Evaluation of Cattle in a Multi-Breed Context. *Livest. Sci.* 166, 101–110. doi:10.1016/j.livsci.2014.05.008

Massender, E., Brito, L. F., Maignel, L., Oliveira, H. R., Jafarikia, M., Baes, C. F., et al. (2022). Single-step Genomic Evaluation of Milk Production Traits in Canadian Alpine and Saanen Dairy Goats. *J. Dairy Sci.* 105, 2393–2407. doi:10.3168/jds.2021-20558

Mdladla, K., Dzomba, E. F., Huson, H. J., and Muchadeyi, F. C. (2016). Population Genomic Structure and Linkage Disequilibrium Analysis of South African Goat Breeds Using Genome-wide SNP Data. *Anim. Genet.* 47, 471–482. doi:10.1111/age.12442

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. H. (2002). BLUPF90 and Related Programs. Commun. No. 28–07. in Proc. 7th World Congr. Montpellier, France: Genet. Appl. Livest. Prod 98, skaa101. doi:10.1093/jas/skaa101

Misztal, I., Lourenco, D., and Legarra, A. (2020). Current Status of Genomic Evaluation. *J. Anim. Sci.* 98, skaa101. doi:10.1093/jas/skaa101

Mohammad Rahimi, S., Rashidi, A., and Esfandyari, H. (2020). Accounting for Differences in Linkage Disequilibrium in Multi-Breed Genomic Prediction. *Livest. Sci.* 240, 104165. doi:10.1016/j.livsci.2020.104165

Mucha, S., Mrode, R., MacLaren-Lee, I., Coffey, M., and Conington, J. (2015). Estimation of Genomic Breeding Values for Milk Yield in UK Dairy Goats. *J. Dairy Sci.* 98 (11), 8201–8208. doi:10.3168/jds.2015-9682

Oget, C., Servin, B., and Palhière, I. (2019). Genetic Diversity Analysis of French Goat Populations Reveals Selective Sweeps Involved in Their Differentiation. *Anim. Genet.* 50, 54–63. doi:10.1111/age.12752

Oliveira, H. R., McEwan, J. C., Jakobsen, J., Blichfeldt, T., Meuwissen, T., Pickering, N., et al. (2020). Genetic Connectedness between Norwegian White Sheep and New Zealand Composite Sheep Populations with Similar Development History. *Front. Genet.* 11, 371. doi:10.3389/fgene.2020.00371

Oliveira, H. R. d., McEwan, J. C., Jakobsen, J. H., Blichfeldt, T., Meuwissen, T. H. E., Pickering, N. K., et al. (2022). Across-country Genomic Predictions in Norwegian and New Zealand Composite Sheep Populations with Similar Development History. *J. Anim. Breed. Genet.* 139 (1), 1–12. doi:10.1111/jbg.12642

Pszczola, M., Strabel, T., Mulder, H. A., and Calus, M. P. L. (2012). Reliability of Direct Genomic Values for Animals with Different Relationships within and to the Reference Population. *J. Dairy Sci.* 95, 389–400. doi:10.3168/jds.2011-4338

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Rezende, F. M., Haile-Mariam, M., Pryce, J. E., and Peñagaricano, F. (2020). Across-country Genomic Prediction of Bull Fertility in Jersey Dairy Cattle. *J. Dairy Sci.* 103, 11618–11627. doi:10.3168/jds.2020-18910

Rupp, R., Mucha, S., Larroque, H., McEwan, J., and Conington, J. (2016). Genomic Application in Sheep and Goat Breeding. *Anim. Front.* 6, 39–44. doi:10.2527/af.2016-0006

van den Berg, I., MacLeod, I. M., Reich, C. M., Breen, E. J., and Pryce, J. E. (2020). Optimizing Genomic Prediction for Australian Red Dairy Cattle. *J. Dairy Sci.* 103, 6276–6298. doi:10.3168/jds.2019-17914

VanRaden, P. M. (2020). Symposium Review: How to Implement Genomic Selection. *J. Dairy Sci.* 103, 5291–5301. doi:10.3168/jds.2019-17684

Venot, E., Pabiou, T., Fouilloux, M.-N., Coffey, M., Laloë, D., Guerrier, J., et al. (2007). Interbeef in Practice: Example of a Joint Genetic Evaluation between France, Ireland and United Kingdom for Pure Bred Limousine Weaning Weights. INTERBULL meeting in 2006 from 4th to 6th June in Kuopio, Finland, 41.

Viklund, Å., Furre, S., Eriksson, S., Vangen, O., and Philipsson, J. (2015). Genetic Conditions of Joint Nordic Genetic Evaluations of Lifetime Competition Performance in Warmblood Sport Horses. *J. Anim. Breed. Genet.* 132, 308–317. doi:10.1111/jbg.12132

Wang, S. (2018). *International Breeding Programs to Improve Health in Pedigree Dogs*. Thesis. Paris, France: Institut agronomique, vétérinaire et forestier de France.

Weigel, K., Rekaya, R., Fikse, F., Zwald, N., and Gianola, D. (2000). Data Structure and Connectedness Issues in International Dairy Sire Evaluations. INTERBULL meeting in 2006 from 4th to 6th June in Kuopio, Finland, 26.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011). Genome Partitioning of Genetic Variation for Complex Traits Using Common SNPs. *Nat. Genet.* 43, 519–525. doi:10.1038/ng.823

**Conflict of Interest:** The author BB was employed by the company Qualitas AG.

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

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

ORIGINAL RESEARCH
published: 10 August 2022
doi: 10.3389/fgene.2022.943705

# Statistical Methods Revisited for Estimating Daily Milk Yields: How Well do They Work?

*Xiao-Lin Wu*[1,2]*, *George R. Wiggans*[1], *H. Duane Norman*[1], *Asha M. Miles*[3], *Curtis P. Van Tassell*[3], *Ransom L. Baldwin VI*[3], *Javier Burchard*[1] *and João Dürr*[1]

[1]*Council on Dairy Cattle Breeding, Bowie, MD, United States,* [2]*Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, United States,* [3]*USDA, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD, United States*

Cost-effective milking plans have been adapted to supplement the standard supervised twice-daily monthly testing scheme since the 1960s. Various methods have been proposed to estimate daily milk yields (DMY), focusing on yield correction factors. The present study evaluated the performance of existing statistical methods, including a recently proposed exponential regression model, for estimating DMY using 10-fold cross-validation in Holstein and Jersey cows. The initial approach doubled the morning (AM) or evening (PM) yield as estimated DMY in AM-PM plans, assuming equal 12-h AM and PM milking intervals. However, in reality, AM milking intervals tended to be longer than PM milking intervals. Additive correction factors (ACF) provided additive adjustments beyond twice AM or PM yields. Hence, an ACF model equivalently assumed a fixed regression coefficient or a multiplier of "2.0" for AM or PM yields. Similarly, a linear regression model was viewed as an ACF model, yet it estimated the regression coefficient for a single milk yield from the data. Multiplicative correction factors (MCF) represented daily to partial milk yield ratios. Hence, multiplying a yield from single milking by an appropriate MCF gave a DMY estimate. The exponential regression model was analogous to an exponential growth function with the yield from single milking as the initial state and the rate of change tuned by a linear function of milking interval. In the present study, all the methods had high precision in the estimates, but they differed considerably in biases. Overall, the MCF and linear regression models had smaller squared biases and greater accuracies for estimating DMY than the ACF models. The exponential regression model had the greatest accuracies and smallest squared biases. Model parameters were compared. Discretized milking interval categories led to a loss of accuracy of the estimates. Characterization of ACF and MCF revealed their similarities and dissimilarities and biases aroused by unequal milking intervals. The present study focused on estimating DMY in AM-PM milking plans. Yet, the methods and relevant principles are generally applicable to cows milked more than two times a day.

**Keywords: dairy cattle, days in milk, lactation, exponential growth function, milking interval**

OPEN ACCESS

***Edited by:***
*Li Ma,*
*University of Maryland, United States*

***Reviewed by:***
*Jose Denis-Robichaud,*
*Independent researcher*
*Navid Ghavi Hossein-Zadeh,*
*University of Guilan, Iran*

****Correspondence:***
*Xiao-Lin Wu*
*nick.wu@uscdcb.com*

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

***Received:*** *14 May 2022*
***Accepted:*** *23 June 2022*
***Published:*** *10 August 2022*

***Citation:***
*Wu X-L, Wiggans GR, Norman HD, Miles AM, Van Tassell CP, Baldwin RL, Burchard J and Dürr J (2022) Statistical Methods Revisited for Estimating Daily Milk Yields: How Well do They Work? Front. Genet. 13:943705. doi: 10.3389/fgene.2022.943705*

## INTRODUCTION

Accurate milking data are essential for herd management and genetic improvement in dairy cattle. In reality, lactation (305 days) yields are not directly measured, but they are calculated from the test-day yields, either with or without explicitly imputing DMY for non-test dates (VanRaden, 1997; Cole and VanRaden; Cole et al., 2009). For genetic evaluation programs, the standardization of lactation yields is practiced, ensuring that milking records are comparable between cows. The latter goal is to adjust variation due to, for example, the number of milking per day, lactation length, and age and month of calving (McDaniel, 1973; Schutz and Norman, 1994; Norman et al., 1995). Hence, the accuracies of test-day yields form the basis for the accuracies of lactation yields and the following standardization of lactation yields for genetic evaluation programs.

Nevertheless, test-day yields are not directly measured either. In the US, reduced-cost milking plans started to displace the standard supervised twice-daily, monthly testing scheme in the 1960s, motivated by reducing visits by a DHIA supervisor (Puttnam and Gilmore, 1968). Typically, cows are milked two or more times on a test day, but not all these milkings are measured. Porzio (1953) was the first to propose sampling the morning (**AM**) and evening (**PM**) milkings alternately on test days throughout lactation in the mountainous areas of Italy. This was known as the AM-PM milking plan, and the daily yield was taken to be approximately two times the yield of a single milking, assuming equal 12-h intervals for AM and PM milkings. In practice, however, AM and PM milking intervals can be different, and milk secretion rates may vary between day and night. Morning milking intervals tend to be longer than afternoon milking intervals. Hence, AM milk yields are usually higher than PM milk yields (Puttnam and Gilmore, 1970).

Various methods have been proposed to estimate daily milk, fat, and protein yields. The landmark developments date to the 1980s and 1990s, focusing on adjustment criteria in two broad categories, namely, additive (**ACFs**) and multiplicative correction factors (**MCFs**). ACFs provide additive adjustments beyond the two times AM or PM yields as the estimate of daily yields. Everett and Wadell (1970a) showed that the difference between AM and PM yields was a function of milking interval and days in milk (**DIM**). Significant factors affecting differences varied with cattle breeds, which also include lactation months, herd production, age classes, and so on (Everett and Wadell, 1970b). Hence, ACFs are evaluated by the average differences between AM and PM yields milk, say, in AM-PM milking plans, for various milking interval classes (**MICs**), and other categorical variables.

On the other hand, MCFs are ratios of daily yield to yield from a single milking, computed for each MIC. MCFs are also referred to as ratio factors. Multiplying a yield from a single milking by an appropriate ratio factor gives an estimate of daily yield. Various MCF forms have been proposed, yet the statistical interpretations differ (Wu et al., 2022). Shook et al. (1980) described the MCF as reciprocals of the AM or PM portions of daily yields, subject to quadratic smoothing. DeLorenzo and Wiggans (1986) proposed deriving MCF for AM-PM milking plans based on a linear regression model without intercept. They assumed heterogeneous means and variances and fitted separate regression models to each MIC. Wiggans (1986) proposed deriving MCFs for cows milked three times a day by regressing single-to-daily yield ratio on milking interval. Additional predictors such as DIM can be included in the model when applicable. MCF models are statistically challenged by "the ratio problem" because they have a ratio variable (i.e., proportional daily yield) as the dependent variable in the data density (Wiggans, 1986) or the smoothing functions (Shook et al., 1980; DeLorenzo and Wiggans, 1986). Consequences included possible biases in two aspects: omitted variable bias and measurement error bias (Lien et al., 2017). The former bias happens because main model effects are missing if the model is re-arranged by multiplying the denominator variable to both sides of the model equation. The latter bias occurs when there are measurement errors with the denominator variable of the response. Furthermore, the MCF models postulated a rational function between daily milky yield and milking, in which the numerator is 1, and the denominator is a linear function (DeLorenzo and Wiggans, 1986; Wiggans, 1986) or a quadratic function (Shook et al., 1980) of milking interval.

Previous studies almost exclusively assessed the accuracy of estimated daily yield in the same datasets from which the correction factors were derived (Putnam and Gilmore, 1968, 1970; Smith and Person, 1981; Liu et al., 2000). This type of in-sample evaluation essentially reflected more model-fitting accuracy than prediction accuracy. In the present study, our primary goal was to evaluate the performance of existing statistical models, including the recently proposed exponential regression model (Wu et al., 2022), in Holstein and Jersey cattle by cross-validation. Secondary goals included comparing model parameters and characterizing ACF and MCF obtained from various models, relative to the initial approach assuming a fixed multiplicative factor of 2.0 for AM or PM yields. Cross-validation, also referred to as out-of-sample testing, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent dataset (Stone, 1974; Geisser, 1993). Briefly, one round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (i.e., training set), and validating the analysis on the other subset (i.e., validation or testing set). To access variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are combined by averaging over the rounds to give an estimate of the model's predictive performance. Hence, cross-validation combines (averages) measures of fitness in prediction to derive a more accurate estimate of model prediction performance. Because cross-validation is a resampling method that uses different portions of the data to train and test a model across iterations, it also allows inferring the error origins by decomposing an MSE into the variance of the estimate and squared bias. The inverse of the variance provides a measure of precision for the estimates.

# MATERIALS AND METHODS

## AM-PM Milking Data

Milking records were extracted from the data repositories maintained by the Council for Dairy Cattle Breeding (**CDCB**). The data consisted of 9,218 milking records from 6,533 cows in 27 herds in 11 states, USA, collected from 2006 through 2009 (**Table 1**). Most milking records consisted of 82.7% Holsteins and 13.1% Jersey (13.1%) cows. The remaining (4.2%) milking records represented multiple breeds, including Ayrshire (0.7%), Brown Swiss (2.4%), Milking Shorthorn (0.01%), Red and White Holstein (0.04%), and unknown breeds (0.87%). Milking records from Holstein and Jersey cows were used in the present study. Data editing excluded records with missing or incomplete values for relevant columns (e.g., AM or PM milking yield, AM or PM milking interval, parity, lactation year or month, days in milk (**DIM**), and herd locations). The final dataset retained 7,544 Holstein milking records from 23 herds and 1,194 Jersey milking records from 9 herds. Approximately, one-third of records (30.6–39.9%) represented the first parity cows and two-thirds (59.4–69.4%) were the second parity cows in the two breeds. Milking records collected from parity 3 and greater were rate (0–10.7%).

## Statistical Methods

### Model 0 (M0): Doubling AM or PM Milking Yield

The initial AM-PM milking plan alternately sampled AM or PM milking on a test day throughout lactation, and the daily yield was obtained by doubling single milk weighed on each test day (Porzio, 1953). That is,

$$\hat{y}_{ij} = 2x_{ij}, \tag{1}$$

where $x_{ij}$ is the known AM ($j = 1$) or PM ($j = 2$) yield for cow $i$, and $\hat{y}_{ij}$ is an estimated DMY. Doubling AM or PM milk yield is equivalent to assuming a fixed multiplicative correction factor for all cows, assuming equal (12–12 h) AM and PM milking intervals.

### Model 1 (M1): ACF Model With Discrete Variables

Additive correction factors are evaluated by the expected values of the differences between AM and PM yields, computed locally for each MIC, coupled with other categorical variables such as lactation months (Everett and Wadell, 1970b). For example, let $z_{ijkl}$ be the difference between AM and PM milk yield for cow $i$, pertaining to MIC $k$ and lactation month (LM) $l$. Assume that the yield from milking $j$ is measured. The data model accounting for variations due to MIC and LM is the following:

$$z_{ijkl} = \mu_j + MIC_k + LM_l + (MIC * LM)_{kl} + \epsilon_{ijkl}, \tag{2}$$

where $\mu_j$ is the overall mean for milking $j$, $MIC_k$ and $LM_l$ are the main effects for MIC $k$ and LM $l$, respectively, $(MIC * LM)_{kl}$ is the interaction effect, and $\epsilon_{ijkl}$ is an error term. Then, ACF (denoted by $\Delta_{jkl}$) are computed by

$$\Delta_{jkl} = E\left(z_{ijkl}\right),$$

$$\approx \hat{\mu}_j + \widehat{MIC}_j + \widehat{LM}_k + (\widehat{MIC*LM})_{jk}. \tag{3}$$

Given the computed ACF and a single milk yield that has been measured for cow $i$ (denoted by $x_{ijkl}$), the estimated daily milk yield (**DMY**, denoted by $\hat{y}_{ijkl}$) is obtained as follows:

$$\hat{y}_{ijkl} = \Delta_{jkl} + 2x_{ijkl}. \tag{4}$$

In the aforementioned equation, we see that an ACF model is equivalent to a regression model assuming a fixed regression coefficient (2.0) for AM or PM yield. ACF models can be fit on AM or PM milk yields separately or jointly.

### Model 2 (M2A,B): ACF Model With Continuous Variables

An ACF model can also be fitted with continuous variables for milking interval (denoted by $t_{ij}$) and DIM (denoted by $d_{ij}$), assuming heterogeneous intercepts and common slopes for milking interval and DIM, respectively, as follows.

$$z_{ij} = \alpha_j + \beta t_{ij} + \gamma\left(d_{ij} - d_0\right) + \epsilon_{ij}, \tag{5}$$

where $z_{ij}$ is the difference between milking j and the other milking for cow $i$, $\alpha_j$ is the intercept for milking $j$, $\beta$ and $\gamma$ are the common regression coefficients for milking interval and DIM, respectively, $d_0$ is an arbitrary constant value for DIM, say, $d_0 = 158$, and $\epsilon_{ij}$ is an error. Here, DIM is used as a continuous variable, instead of the categorical LM.

Given the estimated model parameters, DMY is estimated by

$$\hat{y}_{ij} = \hat{\alpha}_j + \hat{\beta} t_{ij} + \hat{\gamma}\left(d_{ij} - d_0\right) + 2x_{ij}, \tag{6}$$

where $x_{ij}$ is the measured yield from milking $j$ for cow $i$. By this approach, the model is referred to as M2A. Alternatively, ACF are computed for discretized MIC, say MIC $k$ of milking $j$ (denoted by $\Delta_j^{(k)}$):

$$\Delta_j^{(k)} = \hat{\alpha}_j + \hat{\beta}\bar{t}_j^{(k)}, \tag{7}$$

where $\bar{t}_j^{(k)}$ is a midpoint of MIC $k$. Here, we used superscript "(k)" to pinpoint discretized MIC, which distinguishes from a subscript $k$ for a categorical variable for MIC in the model. This notation is used throughout this report. Then, DMY is estimated by

$$\hat{y}_{ij} = \Delta_j^{(k)} + \hat{\gamma}\left(d_{ij} - d_0\right) + 2x_{ij}. \tag{8}$$

With the latter approach (denoted by M2B), DMY is estimated through the ACF.

### Model 3 (M3A,B): Linear Regression With Linear Milking Interval and DIM

The linear model approach treats DMY as the response variable. Let $y_{ij}$ be a daily yield for cow $i$ on milking $j$, $x_{ij}$ be a yield from a single milking from milking $j$, $t_{ij}$ be the milking interval time, and $d_{ij}$ be the responding DIM for the test date. Then, the linear regression model accounting for the aforementioned variables is the following:

$$y_{ij} = \alpha_j + \beta t_{ij} + \gamma\left(d_{ij} - d_0\right) + bx_{ij} + \epsilon_{ij}. \tag{9}$$

In **(9)**, $\alpha_j$ is an overall mean specific to milking $j$, $\beta$, $\gamma$, and $b$ are common regression coefficients for milking interval, DIM, and single milk (AM or PM) yield, respectively, and $\epsilon_{ij}$ is an error.

Linear regression also offers two methods of estimating DMY. First, DMY for a cow can be estimated directly given the estimated model parameters in **(9)**, as follows:

$$\hat{y}_{ij} = \hat{\alpha}_j + \hat{\beta} t_{ij} + \hat{\gamma}\left(d_{ij} - d_0\right) + \hat{b} x_{ij}. \tag{10}$$

The aforementioned equation is referred to as the model M3A. Second, ACF can be computed on discretized MIC, following the same formula as (7), and then DMY are estimated by the following (denoted by M3B):

$$\hat{y}_{ij} = \Delta_j^{(k)} + \hat{\gamma}\left(d_{ij} - d_0\right) + \hat{b} x_{ij}. \tag{11}$$

## Model 4 (M4): Linear Regression With Linear and Quadratic Milking Interval and DIM

Linear regression models can be defined with varying complexity (Liu et al., 2000; Schaeffer et al., 2000). In the present study, we also evaluated a linear regression model with linear and quadratic variables for milking interval and DIM:

$$y_{ij} = \alpha_j + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \gamma_1\left(d_{ij} - d_0\right) + r_2\left(d_{ij} - d_0\right)^2 + b x_{ij} + \epsilon_{ij}. \tag{12}$$

Given the estimated model parameters, DMY is estimated directly as follows:

$$\hat{y}_{ij} = \hat{\alpha}_j + \hat{\beta}_1 t_{ij} + \hat{\beta}_2 t_{ij}^2 + \hat{\gamma}_1\left(d_{ij} - d_0\right) + \hat{\gamma}_2\left(d_{ij} - d_0\right)^2 + \hat{b} x_{ij}. \tag{13}$$

MCF could be derived similar to M2B, yet considering the quadratic terms, but they were not evaluated in the present study.

## Model 5 (M5): The 1980 Shook-Jensen-Dickimson MCF model

Shook et al. (1980) described MCF by the inverse of AM or PM proportion of daily milk yield. For example, MCF given PM yields are formulated as follows:

$$F_{jk} = \frac{AMP_{jk} + PMP_{jk}}{PMP_{jk}}, \tag{14}$$

where $j = 2$ (PM), and $AMP_k$ and $PMP_k$ stand for bulk AM and PM yields, respectively, for MIC $k$ in a population. Shook et al. (1980) employed a quadratic regression of the PM portion of DMY on MIC midpoints, and smoothed estimates of MCF were obtained as follows:

$$F_{jk} = \frac{1}{\hat{\alpha}_j + \hat{\beta}_{j1}\bar{t}_{jk} + \hat{\beta}_{j2}\bar{t}_{jk}^2}. \tag{15}$$

In the aforementioned equation, $\hat{\alpha}_j$, $\hat{\beta}_{j1}$, and $\hat{\beta}_{j2}$ are the estimated intercept and regression coefficients in the quadratic smoothing function, and $\bar{t}_{jk}$ is the midpoint of MIC $k$ for milking $j$. The quadratic smoothing also provided estimates for MIC with no or insufficient milking records.

Given the estimated PM MCF, the AM MCF can be computed indirectly (Shook et al., 1980), but this approach was not taken in the present study. Instead, we computed AM and PM MCF directly from the AM or PM milking data. Similar to (14), MCF given AM yields are formulated to be the inverses of the AM portion of daily yield, computed for each AM MIC ($j = 1$):

$$F_{jk} = \frac{AMP_{jk} + PMP_{jk}}{AMP_{jk}}. \tag{16}$$

Given the MCF ($F_{jk}$) and the yield from single milking $j$ for an animal, say $i$, measured on MIC $k$ ($x_{ijk}$), DMY for this animal is estimated by

$$\hat{y}_{ijk} = F_{jk} x_{ijk}. \tag{17}$$

## Model 6 (M6): The 1986 DeLorenzo and Wiggans MCF model

DeLorenzo and Wiggans (1986) derived MCF for cows milked twice a day based on a linear regression without intercept. They assumed heterogeneous means and variances and fitted separate linear regression models for different MIC.

$$y_{ijk} = b_{jk} x_{ijk} + \gamma_{jk}\left(d_{ijk} - d_0\right) + \epsilon_{ijk} \tag{18}$$

In **(18)**, $b_{jk}$ is the regression coefficient for single milk yield, and $\gamma_{jk}$ is the regression coefficient of DIM. Here, the regression coefficient, $b_{jk}$, coincides with the multiplicative correction factor, as defined by Shook et al. (1980) derived for MIC $k$ of milking $j$, assuming $E(d_{ijk} - d_0) = 0$. DeLorenzo and Wiggans (1986) employed a linear regression smoothing for the reciprocals of computed AM and PM factors, respectively:

$$F_j^{(k)} = \frac{1}{\hat{\alpha}_j + \hat{\beta}_j \bar{t}_{jk}}. \tag{19}$$

Given the computed MCF, DMY is estimated by

$$\hat{y}_{ij}^{(k)} = F_{ij}^{(k)} x_{ij}^{(k)} + \hat{\gamma}_{jk}\left(d_{ij}^{(k)} - d_0^{(k)}\right). \tag{20}$$

## Model 7 (M7A,B): The 1986 Wiggans MCF model

Wiggans (1986) proposed to derive MCF for cows milked three times a day by modeling the single-to-daily milk yield ratio as a linear function of milking interval and DIM when applicable:

$$\frac{x_{ij}}{y_{ij}} = \alpha_j + \beta t_{ij} + \gamma\left(d_{ij} - d_0\right) + \epsilon_{ij}. \tag{21}$$

The aforementioned model also applies to cows milked more than three times and, arguably, it applies to cows milked twice a day. In the latter case, however, the model is subject to the violation of linearity with a longer milking interval (Schmidt, 1960). In the present study, DMY is estimated directly based on the estimated model parameters from **(21)** or through computed MCF according to Wiggans (1986). In the former

**TABLE 1 |** Number (n) and percentage (%n) of milking records by parities, lactation years, and states in the Holstein and Jersey cattle, respectively.

| Variable | | Holstein | | Jersey | |
|---|---|---|---|---|---|
| | | n | %n | n | %n |
| Parity | 1 | 3,006 | 39.9 | 366 | 30.6 |
| | 2 | 4,482 | 59.4 | 831 | 69.4 |
| | 3+ | 56 | 0.70 | 0 | 0 |
| SUM | | 7,544 | 100 | 1,197 | 100 |
| Year | 2006 | 153 | 2.00 | 434 | 36.3 |
| | 2007 | 338 | 4.50 | 0 | 0 |
| | 2008 | 7,000 | 92.8 | 360 | 30.1 |
| | 2009 | 53 | 0.70 | 403 | 33.7 |
| SUM | | 7,544 | 100 | 1,197 | 100 |
| State | Vermont | 1,738 | 23.0 | 4 | 0.30 |
| | New York | 361 | 4.80 | 182 | 15.2 |
| | Pennsylvania | 1,224 | 16.2 | 333 | 27.8 |
| | Indiana | 375 | 5.00 | 206 | 17.2 |
| | Minnesota | 338 | 4.50 | 0 | 0 |
| | Iowa | 153 | 2.00 | 434 | 36.3 |
| | Delaware | 511 | 6.80 | 2 | 0.20 |
| | Maryland | 900 | 11.9 | 0 | 0 |
| | West Virginia | 252 | 3.30 | 0 | 0 |
| | Georgia | 945 | 12.5 | 36 | 3.00 |
| | Florida | 747 | 9.90 | 0 | 0 |
| SUM | | 7,544 | 100 | 1,197 | 100 |

case (denoted by M7A), DMYs are computed directly given the estimated model parameters, as follows:

$$\hat{y}_{ij} = \frac{x_{ij}}{\hat{\alpha}_j + \hat{\beta} t_{ij} + \hat{\gamma}\left(d_{ij} - d_0\right)}. \tag{22}$$

In the latter case (denoted by M7B), MCF are obtained by locally taking the expected value on both sides of **Equation 21**, assuming $E(\gamma(d_{ij}^{(k)} - d_0^{(k)})) = 0$ and $E(\epsilon_{ij}^{(k)}) = 0$. In other words,

$$F_j^{(k)} = \frac{1}{\alpha_j + \beta \bar{t}_j^{(k)}}. \tag{23}$$

Similarly, by taking the first-order Taylor series approximation of **(21)**, that is, $E(\frac{x_{ij}}{y_{ij}}) \approx \frac{E(x_{ij})}{E(y_{ij})}$, and assuming $E(\epsilon_{ij}) = 0$, DMY is estimated using the same formula as (20).

### Model 8 (M8A,B): Exponential Regression Model

Considering milking interval and days in milk, the exponential regression model for estimating DMY takes the following form (Wu et al., 2022):

$$y_{ij} = x_{ij}^b \, e^{\left(\alpha_j + \beta t_{ij} + \gamma\left(d_{ij} - d_0\right) + \epsilon_{ij}\right)}. \tag{24}$$

By noting $e \approx 2.718$, the exponential function is analogous to an exponential growth (or decay) function, given its initial value $y_0 = x_{ij}^b$:

$$y = y_0 (1 + r)^{t^*}, \tag{25}$$

where $r = 1.718$ is the rate of change, tuned by a time function, $E(t^*) = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0)$, as a linear function of milking interval and days in milk, and $y_0 = x^b$ is the initial state. Here, $y$ has an exponential growth when $t^* > 0$, or an exponential decay when $t^* < 0$.

The model parameters can be estimated by taking the following logarithm transformation:

$$log\left(y_{ij}\right) = \alpha_j + \beta t_{ij} + \gamma\left(d_{ij} - d_0\right) + blog\left(x_{ij}\right) + \epsilon_{ij}. \tag{26}$$

As a direct approach. DMY is estimated, given the model parameter estimates ($\hat{b}, \hat{\alpha}_j, \hat{\beta}$, and $\hat{\gamma}$) (denoted as the model M8A), assuming $E(\epsilon_{ij}) = 0$. In other words,

$$\hat{y}_{ij} = x_{ij}^{\hat{b}} \, e^{\left(\hat{\alpha}_j + \hat{\beta} t_{ij} + \hat{\gamma}\left(d_{ij} - d_0\right)\right)}. \tag{27}$$

Alternatively, MCF is computed locally for discretized MIC (Wu et al., 2022):

$$F_j^{(k)} = E\left(x_{ij}^{(k)}\right)^{b-1} \rho_j^{(k)} e^{\left(\hat{\alpha}_j + \hat{\beta} \bar{t}_j^{(k)}\right)}, \tag{28}$$

where $\rho_j^{(k)} = e^{\frac{1}{2}(V(y_{ij}^{(k)})E(y_{ij}^{(k)})^{-2} - bV(x_{ij}^{(k)})E(x_{ij}^{(k)})^{-2})}$, and $E(y_{ij}^{(k)}) = \bar{y}_j^{(k)}$ and $E(x_{ij}^{(k)}) = \bar{x}_j^{(k)}$ are the corresponding means for daily yield and AM (or PM) yield. Then, DMY is estimated by

$$\hat{y}_{ij}^{(k)} = F_j^{(k)} x_{ij}^{(k)} \times e^{\hat{\gamma}\left(d_{ij}^{(k)} - d_0^{(k)}\right)}. \tag{29}$$

The logarithm linear regression also suggests that ACF can be computed for estimating $log(y_{ij})$, and then, DMY in its original scale can be computed conveniently by taking an exponential transformation. The option for computing ACF based on the exponential regression model was not evaluated in the study.

## Cross-Validation of Accuracy

The performance of eight selected models and two strategies (**Table 2**) was evaluated for estimating DMY in the Holstein and Jersey milking datasets. The eight models included two ACF models, one with discrete MIC (M1) and the other with a continuous variable for milking interval (M2), a linear regression model M3 and M4, and three MCF models (M5, M6, and M7), according to Shook et al. (1980), DeLorenzo and Wiggans (1986), and Wiggans (1986), respectively, and the exponential regression model (M8), with doubling AM or PM (M0) as the benchmark model for comparison. For the two strategies, a model labeled "A" (M2A, M3A, M7A, and M8A) estimated DMY directly, given the estimates of model parameters, whereas a model labeled "B" (M2B, M3B, M7B, and M8B) estimated DMY indirectly *via* the computed ACF or MCF. Accuracy and decomposed mean squared errors (MSE) were evaluated for each model or model–strategy combination by cross-validation. Briefly, each dataset was divided into 10 approximately equal subsets. Then, nine subsets were pooled for training, and the remaining subset was used for testing the accuracy. The cross-validation process rotated 10 times, with each subset used for testing once and only once. To facilitate inference of the variance of the estimates, cross-validations were replicated 30 times, each with randomly selected subsets of data samples for training and testing.

**TABLE 2 |** Statistical methods and correction factors used in the present study[a,b,c].

| Model | Equation | Additive (Δ) or ratio (F) Factor |
|---|---|---|
| M0 | $y_{ij} = 2x_{ij}$ | $F \equiv 2$ |
| M1 | $y_{ijkl} = \mu_j + MIC_k\colon LM_l + 2x_{ijkl} + \epsilon_{ijkl}$ | $\Delta_{jk} = \hat{\mu}_j + \widehat{MIC_j\colon LM_k}$ |
| M2A | $y_{ij} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + 2x_{ij} + \epsilon_{ij}$ | --- |
| M2B | $y_{ij} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + 2x_{ij} + \epsilon_{ij}$ | $\Delta_j^{(k)} = \hat{\alpha}_j + \hat{\beta}\bar{t}_j^{(k)} + \hat{\gamma}E(d_{ij}^{(k)} - d_0^{(k)})$ |
| M3A | $y_{ij} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + bx_{ij} + \epsilon_{ij}$ | --- |
| M3B | $y_{ij} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + bx_{ij} + \epsilon_{ij}$ | $\Delta_j^{(k)} = \hat{\alpha}_j + \hat{\beta}\bar{t}_j^{(k)} + \hat{\gamma}E(d_{ij}^{(k)} - d_0^{(k)})$ |
| M4 | $y_{ij} = \alpha_j + \beta_1 t_{ij} + \gamma_1(d_{ij} - d_0) + \beta_2 t_{ij}^2 + \gamma_2(d_{ij} - d_0)^2 + bx_{ij} + \epsilon_{ij} -$ | --- |
| M5 | $\frac{\sum_i x_{ij}^{(k)}}{\sum_i y_{ij}^{(k)}} = \alpha_j + \beta_{j1}\bar{t}_j^{(k)} + \beta_{j2}(\bar{t}_j^{(k)})^2 + \epsilon_j$ | $F_j^{(k)} = \frac{1}{\hat{\alpha}_j + \hat{\beta}_{j1}\bar{t}_j^{(k)} + \hat{\beta}_{j2}(\bar{t}_j^{(k)})^2}$ |
| M6 | $y_{ijk} = b_{jk}x_{ijk} + \varepsilon_{ijk};\ (\hat{b}_j^{(k)})^{-1} = \alpha_j + \beta_j\bar{t}_j^{(k)} + \epsilon_j$ | $F_j^{(k)} = \frac{1}{\hat{\alpha} + \hat{\beta}\bar{t}_j^{(k)}}$ |
| M7A | $\frac{x_{ij}}{y_{ij}} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + \epsilon_{ij}$ | |
| M7B | $\frac{x_{ij}}{y_{ij}} = \alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + \epsilon_{ij}$ | $F_j^{(k)} = \frac{1}{\hat{\alpha} + \hat{\beta}\bar{t}_j^{(k)}}$ |
| M8A | $y_{ij} = x_{ij}^b e^{(\alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + \epsilon_{ij})}$ | --- |
| M8B | $y_{ij} = x_{ij}^b e^{(\alpha_j + \beta t_{ij} + \gamma(d_{ij} - d_0) + \epsilon_{ij})}$ | $F_j^{(k)} = \rho^* e^{\hat{\alpha}_j + \hat{\beta}\bar{t}_j^{(k)}}$ |

[a] *M0 = daily milk yield (DMY) estimated by doubling morning (AM) or evening (PM) milk yield; M1 = additive correction factor (ACF) model with categorical milking interval classes (MIC) and lactation months; M2A = ACF model with continuous variables for milking interval and days in milk (DIM); M2B = M2A with ACF computed on discretized MIC; M3A = linear regression of daily milk yield on milking interval and DIM; M3B = M3A with ACF computed on discretized MIC; M4 = M3A with quadratic terms for milking interval and DIM; M5 = multiplicative correction factor (MCF) model according to Shook et al. (1980); M6 = MCF model according to DeLorenzo and Wiggans (1986); M7A = linear regression of AM or PM proportion of DMY on milking interval and DIM (Wiggans, 1986); M7B = M7A with MCF computed for discretized MIC (Wiggans, 1986); M8A = exponential regression model (Wu et al., 2022); M8B = M8A with MCF computed on discretized MIC.*

[b] $\bar{t}_j^{(k)}$ *= midpoint of milking interval* k *of milking* j*, for j = 1 (AM milking) or 2 (PM milking):* $\rho^* = e^{\frac{1}{2}(V(y_{ij}^{(k)})E(y_{ij}^{(k)})^{-2} - bV(x_{ij}^{(k)})E(x_{ij}^{(k)})^{-2})} \times \frac{E(x_{ij}^{(k)})^b}{E(x_{ij}^{(k)})}$.

[c] *--- = computing yield correction factors is not required.*

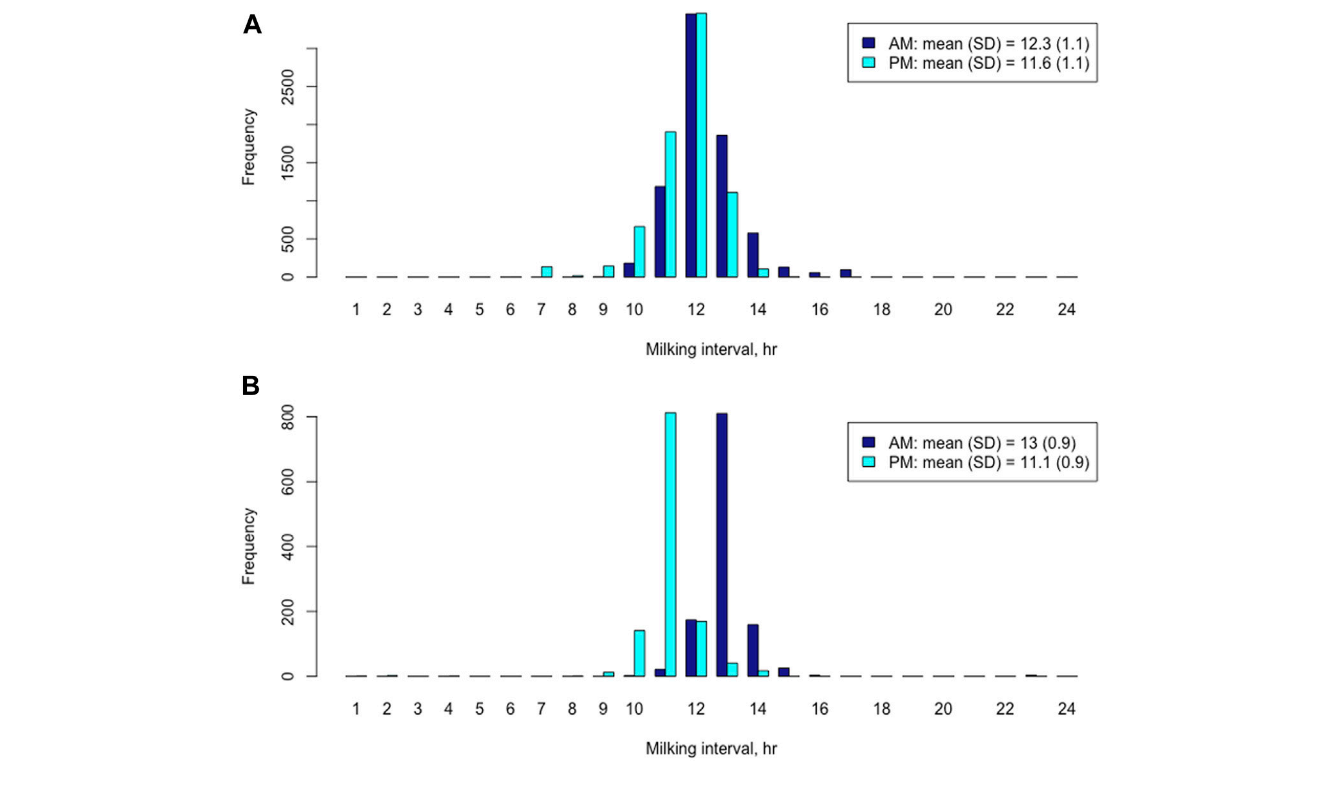


**FIGURE 1 |** Distributions of morning (AM) and evening (PM) milking interval time in Holstein cows **(A)** and Jersey cows **(B)**, respectively.

**TABLE 3 |** Decomposed mean squared error, $R^2$ accuracy, and correlation between estimated and actual daily milk yield obtained from 10-fold cross-validation [a,b,c].

| Method | Holstein | | | | | Jersey | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Varb | Bias$^2$ | MSE | Acc | Cor | Varb | Bias$^2$ | MSE | Acc | Cor |
| M0 | 0 | 22.8 | 22.8 | 0.821 (0) | 0.927 (0) | 0.000 | 14.54 | 14.54 | 0.798 (0) | 0.948 (0) |
| M1 | 0.003 | 11.3 | 11.3 | 0.902 (<0.001) | 0.951 (<0.001) | 0.012 | 6.718 | 6.730 | 0.895 (<0.001) | 0.952 (0.001) |
| M2A | <0.001 | 11.3 | 11.3 | 0.902 (<0.001) | 0.951 (<0.001) | 0.002 | 6.910 | 6.912 | 0.892 (<0.001) | 0.952 (<0.001) |
| M2B | <0.001 | 11.4 | 11.4 | 0.902 (<0.001) | 0.951 (<0.001) | 0.002 | 6.746 | 6.748 | 0.895 (<0.001) | 0.952 (<0.001) |
| M3A | <0.001 | 10.3 | 10.3 | 0.910 (<0.001) | 0.951 (<0.001) | 0.002 | 6.078 | 6.080 | 0.904 (<0.001) | 0.953 (<0.001) |
| M3B | <0.001 | 10.3 | 10.3 | 0.910 (<0.001) | 0.951 (<0.001) | 0.003 | 6.226 | 6.229 | 0.902 (<0.001) | 0.952 (<0.001) |
| M4 | <0.001 | 10.2 | 10.2 | 0.911 (<0.001) | 0.952 (<0.001) | 0.025 | 6.280 | 6.305 | 0.901 (<0.001) | 0.953 (<0.001) |
| M5 | 0.002 | 11.0 | 11.0 | 0.905 (<0.001) | 0.951 (<0.001) | 0.029 | 6.707 | 6.736 | 0.895 (<0.001) | 0.954 (<0.001) |
| M6 | 0.001 | 11.0 | 11.0 | 0.904 (<0.001) | 0.952 (<0.001) | 0.008 | 6.517 | 6.525 | 0.898 (<0.001) | 0.953 (<0.001) |
| M7A | <0.001 | 10.9 | 10.9 | 0.905 (<0.001) | 0.952 (<0.001) | 0.002 | 6.570 | 6.572 | 0.897 (<0.001) | 0.954 (<0.001) |
| M7B | <0.001 | 11.0 | 11.0 | 0.904 (<0.001) | 0.951 (<0.001) | 0.004 | 6.910 | 6.914 | 0.892 (<0.001) | 0.943 (<0.001) |
| M8A | 0.001 | 10.1 | 10.1 | 0.912 (<0.001) | 0.952 (<0.001) | 0.003 | 6.072 | 6.075 | 0.905 (<0.001) | 0.954 (<0.001) |
| M8B | 0.001 | 11.0 | 11.0 | 0.910 (<0.001) | 0.952 (<0.001) | 0.010 | 6.088 | 6.098 | 0.903 (<0.001) | 0.953 (<0.001) |

[a] *M0 = daily milk yield (DMY) estimated by doubling morning (AM) or evening (PM) milk yield; M1 = additive correction factor (ACF) model with categorical milking interval classes (MIC) and lactation months; M2A = ACF, model with continuous variables for milking interval and days in milk (DIM); M2B = M2A with ACF, computed on discretized MIC; M3A = linear regression of daily milk yield on milking interval and DIM; M3B = M3A with ACF, computed on discretized MIC; M4 = M3A with quadratic terms for milking interval and DIM; M5 = multiplicative correction factor (MCF) model according to Shook et al. (1980); M6 = MCF model according to DeLorenzo and Wiggans (1986); M7A = linear regression of AM or PM and proportion of DMY, on milking interval and DIM (Wiggans, 1986); M7B = M7A with MCF, computed for discretized MIC (Wiggans, 1986); M8A = exponential regression model (Wu et al., 2022); M8B = M8A with MCF, computed on discretized MIC.*

[b] *Var = variance; Bias$^2$ = squared bias; MSE, mean squared error; Acc = R$^2$ accuracy; Cor = correlation between the estimated and actual DMY.*

[c] *Numbers in the brackets were standard errors of the R$^2$ accuracy estimates.*

The correlation between the estimate and actual DMY and the following $R^2$ accuracy:

$$R^2 = \frac{\sigma^2}{\sigma^2 + MSE}. \tag{30}$$

Here, $\sigma^2$ was the true phenotype of DMY, assuming actual DMY was obtained without measurement error, and MSE was mean squared error. The $R^2$ accuracy was calculated per cross-validation population-wise or per individual animal. In the former case, MSE were obtained as the population parameter and the $R^2$ accuracy was calculated for each cross-validation replicate. Then, the mean and standard deviation (also referred to as the standard error) of the $R^2$ accuracy and correlation estimates were obtained across the 30 cross-validation replicates. In the latter case, the MSE was calculated as the average across the 30 replicates for each animal, and individual $R^2$ accuracy was calculated according to **Equation 30** per animal.

To infer the origin of errors, the mean squared error (MSE) of DMY estimates from the 10-fold cross-validation was decomposed into the variance ($Var(\hat{y}_i)$) and the squared bias ($Bias^2(\hat{y}_i)$), as follows:

$$\begin{aligned} MSE &= \frac{1}{n \times m} \sum\nolimits_{i=1}^{n} \sum\nolimits_{r=1}^{m} \left(\hat{y}_{ir} - y_i\right)^2 \\ &= \frac{1}{n \times m} \sum\nolimits_{i=1}^{n} \sum\nolimits_{r=1}^{m} \left(\hat{y}_{ir} - \overline{\hat{y}}_i\right)^2 + \frac{1}{n} \sum\nolimits_{i=1}^{n} \left(\overline{\hat{y}}_i - y_i\right)^2 \\ &= Var(\hat{y}_i) + Bias^2(\hat{y}_i). \end{aligned} \tag{31}$$

In the aforementioned equation, $Var(\hat{y}_i) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{r=1}^{m} (\hat{y}_{ir} - \overline{\hat{y}}_i)^2$ and $Bias^2(\hat{y}_i) = \frac{1}{N} \sum_{i=1}^{N} (\overline{\hat{y}}_i - y_i)^2$, where $n$ is the number of animals, $m$ is the number of replicates, $y_i$ was a true DMY for cow $i$, $\hat{y}_{ir}$ was an estimate of daily milking yield from the $rth$ replicate, and $\overline{\hat{y}}_i$ was the average of the estimated DMY across the 30 replicates.

## Cubic Smoothing Splines

Cubic smoothing splines of the individual $R^2$ accuracies and actual daily milk yields, respectively, were also fitted to provide approximations with weaker assumptions for relevant comparisons. Statistically, smoothing splines are function estimates (denoted by $\hat{f}(x)$) obtained from a set of noisy observations $y_i$ of the target $f(x_i)$, which balance a measure of goodness of fit of $\hat{f}(x)$ to $y_i$ with a derivative-based measurement of the smoothness of $\hat{f}(x)$ (Craven and Wahba., 1979). A $k$th order spline is a piecewise polynomial function of degree k, which is continuous and has continuous derivatives of orders 1,. . ., $k - 1$, at its knot points.

Let $\{x_i, y_i; i = 1, \ldots, n\}$ be a set of observations governed by the relation $y_i = f(x_i) + \epsilon_i$. The cubic smoothing spline estimate $\hat{f}$ of the function $f$ is defined to be the minimizer of the following, over the class of twice differentiable functions,

$$\sum\nolimits_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 + \lambda \int \hat{f}''(x)^2 dx. \tag{32}$$

In the aforementioned equation, $\lambda \geq 0$ is a smoothing parameter, controlling the trade-off between fidelity to the data and roughness of the function estimate. This is often estimated by generalized cross-validation or by restricted marginal likelihood (REML) which exploits the link between spline smoothing and Bayesian estimation (because the smoothing penalty can be viewed as being induced by a prior on the f). The integral is often evaluated over the whole real line, although it is also possible to restrict the range to that of $x_i$. As $\lambda \to 0$ (no smoothing), the smoothing spline converges to the interpolating spline. As $\lambda \to \infty$ (infinite smoothing), the roughness penalty becomes paramount and the estimate converges to a linear least squares estimate. The roughness penalty based on the second derivative is the most common in the modern statistics literature, although the method can easily be adapted to penalties based on other derivatives. The penalized sum of squares smoothing objective can be replaced by a *penalized likelihood* objective in which the sum of squares terms is replaced by another log-likelihood-based measure of fidelity to the data. The sum of squares term corresponds to penalized likelihood with a Gaussian assumption on the $\epsilon_i$.

# RESULTS AND DISCUSSION

## Summary of Milking Data for Holstein and Jersey Cows

In the Holstein cows, the mean and median of AM milking intervals were 12.3 h and 12.1 h, respectively, whereas the mean and median of PM milking intervals were 11.6 h and 11.9 h, respectively. The AM milking intervals had a wider range (5.6–23.67 h) than the PM milking intervals (5.0–18.4 h) (**Figure 1A**). A paired *t*-test showed that the mean AM milking interval was significantly longer than the mean PM milking interval in the Holsteins cows (t = 27.3, $p$ < 2.2e-16). The mean difference between AM and PM milking intervals was 0.688 h, with a 95% confidential interval between 0.639 h and 0.738 h. Similarly, the mean and median of AM milking intervals in the Jersey cows were 13.0 h and 12.9 h, respectively. The mean and median of PM milking intervals were 11.1 h and 11.0 h, respectively. The AM milking interval was significantly longer than the PM milking interval based on a paired *t*-test (t = 44.2; $p$ < 2.2e-16). The mean difference between AM and PM milking interval in the Jersey cows was 1.87 h, with a 95% confidential interval between 1.79 h and 1.95 h. The AM milking interval range (9.6–23.5 h) was also larger than the PM milking interval range (1.4–14.3 h) in Jersey cows (**Figure 1B**). The distribution of AM and PM milking intervals was approximately symmetric and bell shaped in the Holstein and Jersey cows, respectively (**Figure 1**).

Longer AM milking intervals led to greater average AM milk yields (**Figure 2**). In the Holstein cows, the mean AM milk yield (16.4 kg) was significantly larger than the average PM milk yield (15.3 kg) (t = 23.5; $p$ < 2.2e-16) (**Figure 2A**). The mean difference between AM and PM milk yield was 2.49 kg, with a 95% confidential interval between 2.29 and 2.70 kg, in the Holstein cows. Similarly, the mean AM milk yield (12.7 kg) was significantly larger than the average PM milk yield (11.0 kg) (t = 22.2; $p$ < 2.2e-16) in the Jersey cows (**Figure 2B**). The mean difference between AM and PM milk yield was 3.87 kg, with a 95% confidential interval between 3.53 and 4.21 kg, in the Jersey cows.

## Comparing Decomposed Mean Squared Errors and Accuracies

Accuracy and precision are two primary measures of observational or estimation errors. For estimating DMY, accuracy tells how close an estimated DMY is to the actual value, whereas precision shows how well the estimates agree with each other. Precision was measured by the inverse of the variance of DMY estimates. The smaller the variance, the greater the precision. Decomposed MSE were shown in **Table 3**. All the methods had close to zero variances for the DMY estimates, meaning they all had high precision of the estimated DMY. The variance of DMY estimates was not greater than 0.003 in Holstein cows and less than 0.03 in Jersey cows. The MSE were dominated by the portion of squared bias in Holstein and Jersey cows. Model M0 (doubling AM or PM milk yields) had the largest squared biases and the largest MSE, which were more than two times their counterparts for all the other models in Holstein and Jersey cows. Comparably speaking, the ACF models had larger squared biases and MSE than the MCF and linear regression models. The exponential regression model (M8A) had the smallest squared biases and the smallest MSE. Not including the model M0, the root MSE was between 3.18 and 3.38 kg in the Holstein cows and between 2.46 and 2.63 kg in the Jersey cows. The root MSE roughly agreed with two or three Schutz and Norman (2011), who reported a range of root MSE between 2.07 and 2.85 kg for cows milked twice a day. Higher root MSE for estimating DMY were reported in cows milked times a day (Schutz et al., 2008). It is worth mentioning that we used a 10-fold cross-validation, whereas Schutz and Norman (2011) employed an in-sample evaluation. Often, cross-validations tend to report higher errors than in-sample evaluations when applied to the same dataset. In-sample errors are the errors we get on the same data we used to train the prediction model, which tends to be optimistic, compared to the errors we would get from a new sample. The latter is referred to as out-of-sample errors. The reason is overfitting with in-sample evaluation (Harkins and Douglas, 2004). Overfitting occurs when the trained predictive model becomes sensitive to the noise in the sample. As a result, the function will perform well on the training set but not perform well on new data. The more overfitting occurs, the worse the predictive model will generalize to new data. When we get a new dataset, there will be different noises, so the accuracy will go down to some extent. Hence, in-sample errors are always less than out-of-sample errors, which leads to overestimated accuracy. Yet, the fact is, once we build a model on a sample of data that we have collected, we might want to test the realistic expectation of the predictive model as to how well it will perform on new data.

The standard deviation of the mean $R^2$ accuracy between the 30 CV replicates was 0 for M0 and less than 0.001 for all the remaining methods. Exactly, the standard deviation of $R^2$ accuracies between cross-validation replicates ranged between 0.00002 and 0.0001 for these methods in Holstein cows and between 0.0001 and 0.0005 in Jersey cows. By this definition, the $R^2$ accuracy is viewed as a population parameter. Based on paired *t*-test, we showed that the exponential regression model had highly significant mean $R^2$ accuracy than each of the existing methods (Holsteins: t = 584.8–37281; $p$ < 2.2–16; Jerseys: 1178.5–5861.4; $p$ < 2.2e-16). The model M0 had the lowest $R^2$ accuracy (0.821 in Holstein cows and 0.798 in Jersey cows). Compared to model M0, the ACF and MCF models, including the linear regression models, highly significantly improved the accuracies for estimating DMY (Holsteins: t = 8658.1–37281; $p <$ 2.2e-16; Jerseys: 11.67–5861.4; $p <$ 1.8e-12). The MCF and linear regression models had slightly higher accuracies of DMY estimates (0.904–0.912 in Holstein cows and 0.892–0.905 in Jersey cows) than the ACF models (0.902–0.910 in Holstein cows and 0.892–0.904 in Jersey cows). The exponential regression models, M8A and M8B, had the greatest $R^2$ accuracies of DMY estimates (0.910–0.912 in Holstein cows and 0.903–0.905 in Jersey cows). Based on a similar criterion, Liu et al. (2000) reported slightly higher $R^2$ accuracies (0.885) for doubling AM or PM approach in the German Holstein cows than ours in the US Holstein cows (0.821). The accuracies of estimated DMY (0.902–0.912) in the US Holstein cows that we obtained using the DeLorenzo and Wiggans (1986) model were within the accuracy range (0.900–0.914) in German Holstein cows obtained by Liu et al. (2000) using the same model. In addition to the genetic differences between German and the US Holstein cows, the accuracies of estimated DMY can vary with evaluation methods. Liu et al. (2000) employed in-sample evaluation, whereas we evaluated the accuracies by 10-fold cross-validation. As mentioned earlier, the accuracy obtained from cross-validation tends to be lower than that from in-sample evaluation because the former evaluations are prone to overfitting (Harkins and Douglas, 2004). Thus, comparing various methods is valid only when applied to the same dataset with the same evaluation strategy.

Correlation has been widely used to measure prediction accuracy, e.g., in genomic prediction and machine learning. However, correlation is not as informative as the $R^2$ accuracy for evaluating the performance of various models to estimate DMY. In the present study, all the models had similarly high correlations (0.951–0.952 in Holstein cows and 0.952–0.954 in Jersey cows) between the estimated and actual DMY, except that the model M0 had significantly lower corrections (0.927 in Holstein cows and 0.948 in Jersey cows). The standard deviation (i.e., standard error) of correlations between cross-validation replicates were all less than 0.0005. In statistics, correlation measures the degree of dependence between two random variables. Yet, correlation is not a precise measure of accuracy for two evident reasons. First, a correlation can be negative, but a valid accuracy measure is non-negative. Second and more importantly, a correlation does not account for estimation biases, meaning that two methods having identical corrections can vary drastically in the biases of the estimates. Hence, we recommend using the $R^2$ accuracy, instead of correction, as the measure of accuracy for estimating DMY.

A couple of reasons are worth noting for the lower accuracies with the ACF models than linear regression models. First, an ACF model is equivalent to assuming a fixed regression coefficient for partial milk yield, which can limit its predictability. For example, consider the models M2A and M2B. With some re-arrangements, these two models can be re-arranged into linear regression models of DMY on milk interval and DIM, plus a variable for AM or PM milk yield with a fixed regression coefficient ($b$ = 2.0). The re-arranged models have similar model settings for predictor variables as the linear regression models, M3A and M3B, except that the linear models treat regression coefficients as unknown and estimated from the data. Possibly, by relaxing the restriction $b$ = 2.0 and estimating it from the data, the linear regression models (M3A and M3B) predicted the data better than the ACF models (M2A and M2B). Second, specific to ACF models with discrete regression variables (e.g., M1), it was challenged by data missing or insufficient data for some MIC, which led to a loss of accuracy for estimating DMY. In reality, deriving ACF from a regression model with discrete variables is also challenged as the number of categorical variables increases. Hence, the computation can be highly intensive or even not practically operational. For example, 20 MIC, 4 herd location regions, 4 years, 4 seasons, and 2 parities were considered. Then, there would be $20 \times 4 \times 4 \times 4 \times 2 = 2,560$ specific classes for which ACF needed to be estimated if considering all these categorical variables at the same time.

Concerning an ACF or MCF model with continuous variables for milking intervals and DIM, discretizing a continuous variable to a categorical variable often leads to loss of information (and, therefore, accuracy) to some extent. Wu et al. (2022) showed analytically that computing ACF and MCF on discretized MIC led to a loss of accuracy of DMY estimates. This phenomenon was empirically observed in the Holstein and Jersey cows in the present study when comparing four pairs of models: M2A versus M2B, M3A versus M3B, M7A versus M7B, and M8A versus M8B. Each pair had the same model settings except that DMY were estimated with different strategies. The models labeled "A" (M2A, M3A, M7A, and M8A) estimated DMY directly based on estimated model parameters. Instead, the models labeled "B" (M2B, M3B, M7B, and M8B) computed ACF or MCF for discretized MICs after data fitting. Then, DMY were estimated through the calculated ACF or MCF. The models in group A consistently had smaller MSE and better accuracies than their counterparts in group B (**Table 3**). These results were an indication that discretizing milking interval time led to a loss of accuracy in estimated DMY. Hence, computing ACF or MCF without accounting for the loss due to discretizing MIC may be suboptimal when the linearity holds.

Relative to model M0 (doubling AM or PM milk yields), ACF and MCF models have considerably improved the DMY accuracy. To probe into the details, we computed the $R^2$ accuracies for individual cows based on three selected models, M0 (doubling AM or PM yields), one ACF model (M2B), and one MCF model (M7B). It came to our attention that mean individual $R^2$ accuracies were higher than the average $R^2$ accuracy population-wise across the 30 replicates. The distributions of individual $R^2$ accuracies in the Holstein cows obtained from these three models are shown in **Figure 3**. In particular, the distribution of individual R accuracies for the model M0 had a thicker tail than that for the model M2B or M7B. This was an indication that doubling AM or PM milk yields as the estimated DMY led to a higher percentage of the estimated DMY with lower accuracies, compared to the ACF and MCF models. The percentage of individual $R^2$ accuracies ≥ 0.90 were 59.6% (M0), 81.6% (M2B), and 83.4% (M7B). Average individual $R^2$ accuracy was 0.934 for M2B and 0.937 for M7B, respectively; both were substantially higher than the average individual $R^2$ accuracy (0.873) for M0. The medians of the $R^2$ accuracies were

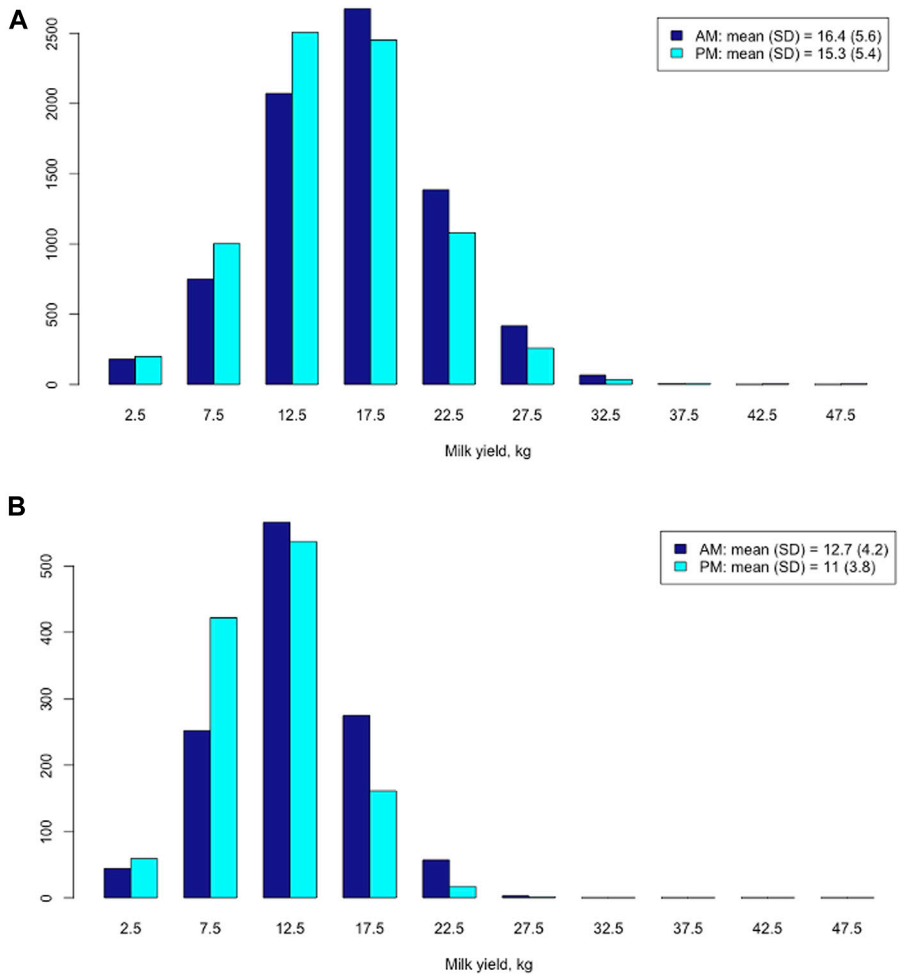


**FIGURE 2 |** Distribution of morning (AM) and evening (PM) milk yields in Holstein cows **(A)** and Jersey cows **(B)**, respectively.

0.927 for M0, 0.976 for M2B, and 0.980 for M7B, respectively, in the Holstein cows. The medians were consistently larger than the means. The MCF model (M7B) had a slightly higher mean $R^2$ accuracy than the ACF model (M2B). Unlike the standard deviations of the average $R^2$ accuracies between cross-validation replicates, which were all close to zero, the standard deviation of the individual $R^2$ accuracy was 0.135 for M0, 0.116 for M2B, and 0.115 for M7B. By Student's *t*-test, the mean $R^2$ accuracies between M2B and M7B was not significantly different (t = 1.69, $p$ = 0.091), yet they both were highly significantly greater than the mean $R^2$ accuracy of M0 (t = 29.4–31.1, $p$ < 2.2e-16). Similar trends were observed in Jersey cows.

Furthermore, the cubic smoothing spline (**CSS**) means of individual $R^2$ accuracies obtained from the three models were plotted against milking interval time in hours. (**Figure 4**). All three models had comparable means of individual $R^2$ accuracies when AM and PM milking intervals were approximately 12 h. Still, the average individual $R^2$ accuracy with the model M0 dropped drastically as the milking interval deviated from 12 h. The further it deviated from 12 h, the lower the average $R^2$ accuracy it had. In contrast, average individual $R^2$ accuracies for models M2B and M7B remained consistently high for milking intervals between 10 h and 16 h. They dropped slightly outside that range due to insufficient milking data. Hence, doubling AM or PM yield is equivalent to assuming a fixed multiplicative factor of 2.0 for AM and PM milk yields. It is valid (or approximately so) only for equal (12–12 h) AM and PM milking intervals but subject to large errors with unequal AM and PM milking intervals. Instead, ACF and MCF effectively provided adjustments to unequal milking intervals, leading to substantially improved DMY accuracies.

## Comparing Model Parameters

Model parameters were estimated and compared for four selected models (M2A, M3A, M7A, and M8A) using all milking data in Holstein and Jersey cows; each was implemented for AM or PM milkings separately and jointly (**Table 4**). The first two models, M2A and M3A, are the baseline models for the ACF models M2B and M3B. Both models (M2A and M3A) were implemented similarly yet with slightly different modeling assumptions. The model M2A equivalently assumed a fixed regression coefficient

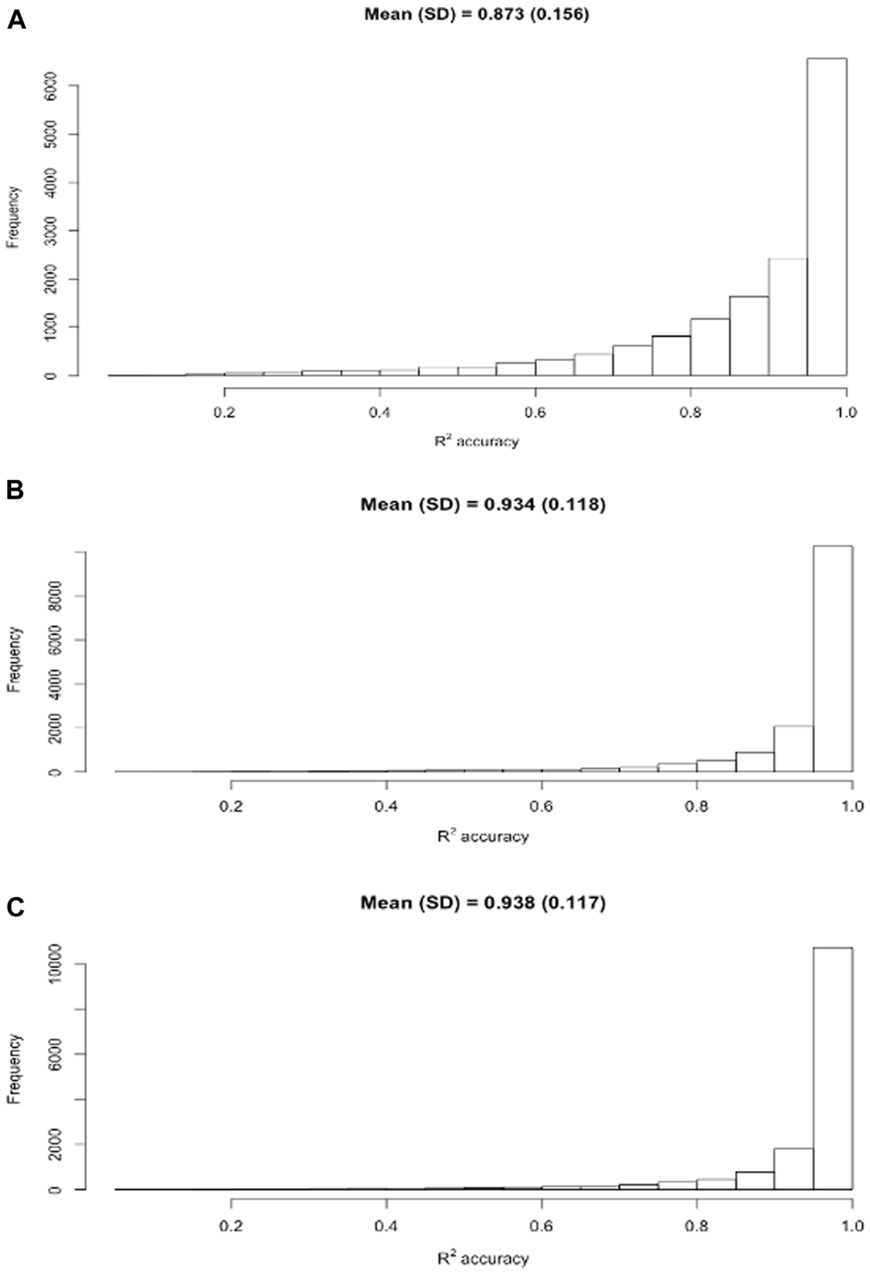


**FIGURE 3 |** Distribution of individual $R^2$ accuracies of the estimated daily milk yield obtained using three models, M0 **(A)**, M2B **(B)**, and M7B **(C)**, respectively. M0 = two times AM or PM yield as the estimate of test-day milk yield; M2B = additive correction factor model implemented by regressing the difference between AM and PM yields on milking interval and days in milk; M7B = multiplicative correction factor model according to Wiggans (1986).

"2.0" for AM or PM milk yields, whereas the model M3A estimated the regression coefficient for AM or PM yield from the data. For example, the estimated regression coefficient with model M3A was 1.749 in Holstein cows and 1.750 in Jersey cows when AM and PM milk yields were analyzed jointly. Hence, the model M2A provided additive adjustments to two times AM or PM milk yields as the DMY estimates, whereas the model M3A provided additive adjustments to approximately 1.75 times AM or PM milk yields as the estimated DMY. Owing to this difference, other model parameters varied between both models. Overall, the model M3A had a slightly larger intercept than the model M2A in both datasets. The regression coefficients for milking intervals were all negative for both models. The absolute value of the regression coefficient for milking interval in the model M2A was larger than that in the model M3A. The model M3A would coincide precisely with the ACF model M2A if we could fix the regression coefficient for AM or PM milk yield to be 2.0 in the model M3A.

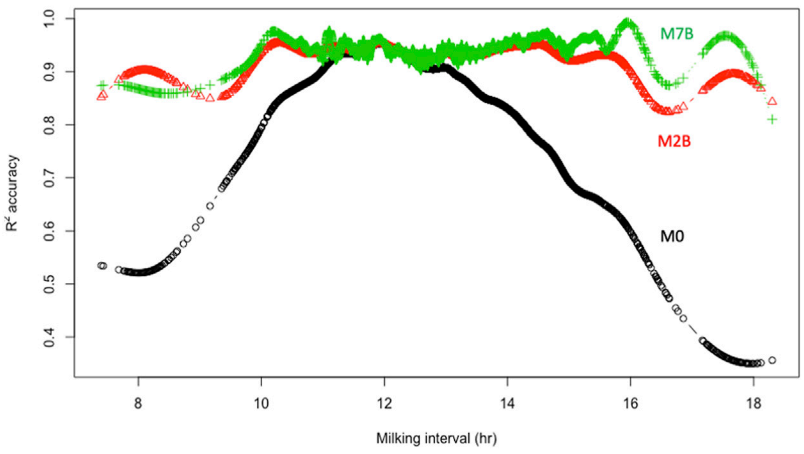


**FIGURE 4 |** Relationships between smooth splines means of individual $R^2$ accuracies of the estimated daily milk yield and milking interval for three models, M0, M2B, and M7B. M0 = two times AM or PM yield as the estimate of test-day milk yield; M2B = additive correction factor model implemented by regressing the difference between AM and PM yields on milking interval and days in milk; M7B = multiplicative correction factor model according to Wiggans (1986).

**TABLE 4 |** Estimated parameters obtained from four models (M2A, M3A, M7A, and M8A), each implemented separately or jointly for known morning (AM) or evening (PM) milk yields [a,b].

| Statistical model | Model parameter | Holstein | | | Jersey | | |
|---|---|---|---|---|---|---|---|
| | | AM | PM | Joint | AM | PM | Joint |
| M2A | $\alpha_1$ | 25.80 (0.431) | --- | 26.04 (0.302) | 9.593 (1.170) | --- | 9.789 (0.807) |
| | $\alpha_2$ | --- | 27.01 (0.870) | 26.79 (0.285) | --- | 11.84 (0.951) | 11.64 (0.692) |
| | $\beta$ | −2.190 (0.035) | −2.222 (0.034) | −2.206 (0.024) | −0.898 (0.090) | −0.905 (0.085) | −0.889 (0.062) |
| | $\gamma$ | 0.001 (3E-4) | −0.001 (3E-4) | -4.7E-5 (2E-4) | 0.001 (0.001) | -0.001 (0.001) | -1.4E-4 (4E-04) |
| M3A | $\alpha_1$ | 27.76 (0.404) | --- | 26.64 (0.283) | 13.52 (1.402) | --- | 11.22 (0.701) |
| | $\alpha_2$ | --- | 28.02 (0.382) | 27.35 (0.267) | --- | 12.49 (0.947) | 12.90 (0.652) |
| | $\beta$ | −1.898 (0.033) | −1.934 (0.034) | −1.909 (0.024) | −0.797 (0.078) | −0.782 (0.086) | −0.746 (0.059) |
| | $\gamma$ | −0.005 (3E-4) | −0.005 (3E-4) | −0.005 (2E-4) | −0.003 (0.001) | −0.003 (0.001) | −0.003 (0.001) |
| | $b$ | 1.720 (0.008) | 1.780 (0.008) | 1.749 (0.005) | 1.664 (0.017) | 1.860 (0.022) | 1.750 (0.014) |
| M7A | $\alpha_1$ | 0.071 (0.008) | --- | 0.068 (0.005) | 0.269 (0.029) | --- | 0.268 (0.020) |
| | $\alpha_2$ | --- | 0.053 (0.007) | 0.056 (0.005) | --- | 0.231 (0.024) | 0.231 (0.017) |
| | $\beta$ | 0.036 (0.001) | 0.037 (0.001) | 0.037 (4E-04) | 0.021 (0.002) | 0.021 (0.002) | 0.021 (0.002) |
| | $\gamma$ | 7E-06 (5E-06) | -5E-06 (5E-06) | 8E-07 (4E-06) | 2E-05 (1E-05) | -2E-05 (1E-05) | 3.3E-06 (1E-05) |
| M8A | $\alpha_1$ | 1.779 (0018) | --- | 1.856 (0.013) | 1.580 (0.067) | --- | 1.575 (0.048) |
| | $\alpha_2$ | --- | 1.946 (0.017) | 1.877 (0.012) | --- | 1.621 (0.060) | 1.638 (0.042) |
| | $\beta$ | −0.059 (0.001) | −0.070 (0.001) | −0.065 (0.001) | −0.037 (0.005) | −0.025 (0.005) | −0.032 (0.004) |
| | $\gamma$ | -2E-04 (1E-05) | -2E-04 (1E-05) | -2E-04 (9E-06) | -3E-04 (3E-05) | -3E-04 (4E-05) | -3E-04 (3E-05) |
| | $b$ | 0.861 (0.004) | 0.852 (0.004) | 0.856 (0.003) | 0.812 (0.010) | 0.757 (0.011) | 0.784 (0.008) |

[a] *M2A = additive correction factor model with continuous variables for milking interval and days in milk (DIM); M3A = linear regression of daily milk yield (DMY) on milking interval and DIM; M7A = linear regression of AM or PM and proportion of DMY, on milking interval and DIM (Wiggans, 1986); M8A = exponential regression model (Wu et al., 2022).*

[b] *$\alpha_1$ = intercepts for AM milk yield; $\alpha_2$ = intercept for PM milk yield; $\beta$ = common regression coefficient for milking interval; $\gamma$ = common regression coefficient for DIM; b = common regression coefficient for AM (or PM) milk yield (M3A) or the logarithm of AM or PM milk yield (M8A).*

The models M7A and M8A are the baseline models for the MCF models, M7B and M8B. The MCF models represented substantially different modeling strategies (**Table 2**). For example, the former model (M7A) fitted AM or PM proportion of DMY as a linear function of milking interval and days in milk (Wiggans, 1986). In contrast, the latter (M8A) was an exponential regression model (Wu et al., 2022). We show that the model M8A was equivalent to a linear regression of the logarithm DMY on milking interval, days in milk, and the logarithm AM (or PM) milk yields through reparameterization. When AM and PM milk yields were analyzed jointly, the regression coefficient for milking interval was positive (0.037 in Holstein cows and 0.021 in Jersey cows) in the model M7A, whereas it was negative (0.065 in Holstein cows and 0.032 in Jersey cows) in the model M8A. The regression coefficient for the logarithm AM (or PM) milk yield was less than 1.0 (0.856 in Holstein cows and 0.784 in Jersey cows) in the model M8A.

Analyzing AM and PM milk yields separately led to slightly different model parameters in Holstein and Jersey cows (**Table 4**).

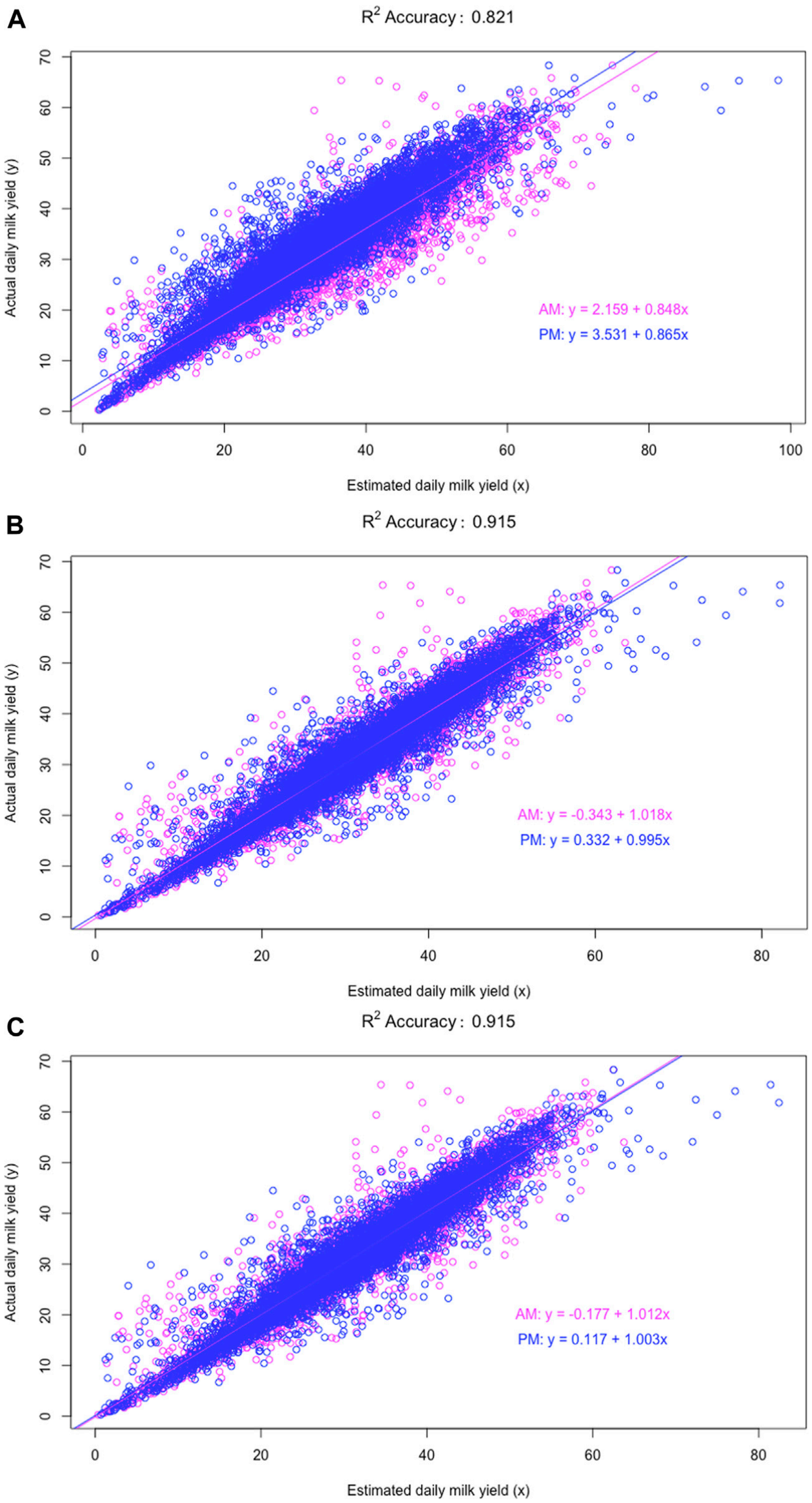


**FIGURE 5 |** Scatterplot and linear regression fits of the actual daily milk yield against estimated daily milk yields under three scenarios: **(A)** estimating daily milk yield (DMY) by doubling morning (AM) or evening (PM) milk yields (model M0); **(B)** estimating DMY for known morning (AM) and evening (PM) milkings separately using the exponential regression model (model M8A; separate analysis); **(C)** estimating DMY for known AM and PM milkings jointly using the exponential regression model (model M8A; joint analysis).

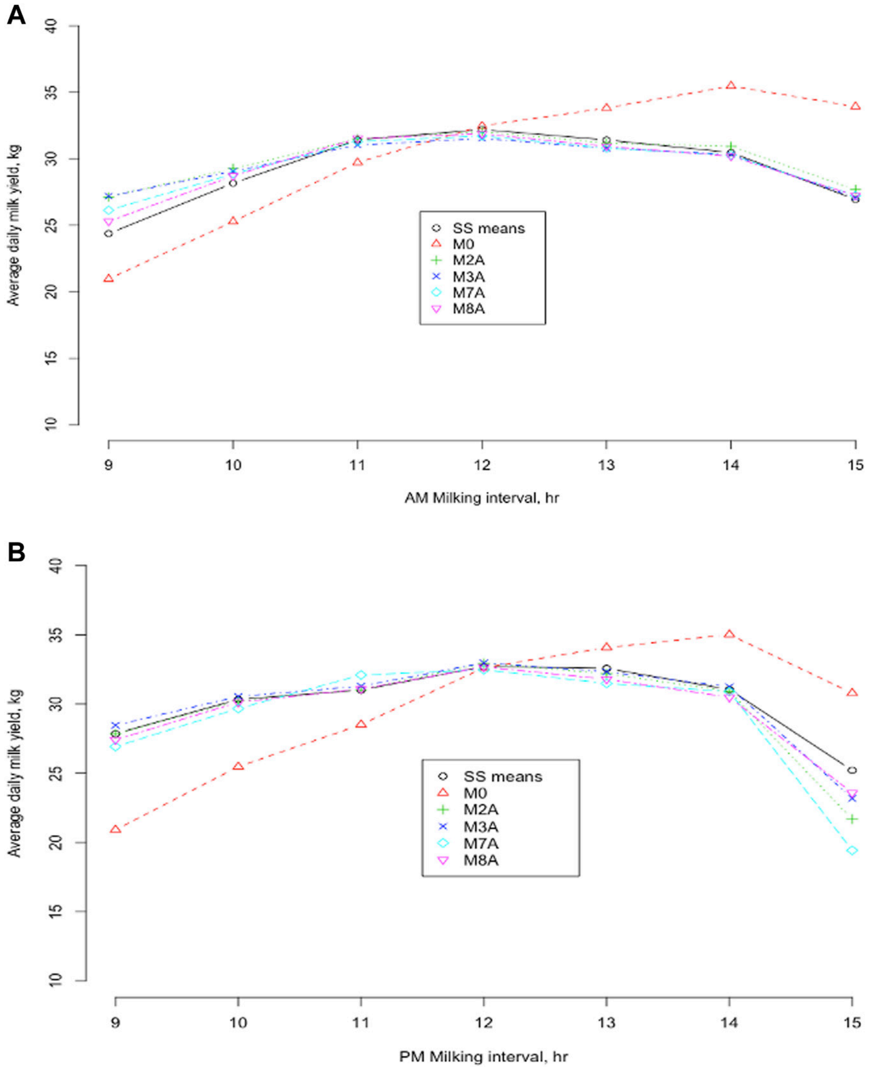


**FIGURE 6 |** Average daily milk yields were obtained from five models and smooth spline (SS) means of the daily milk yield against morning **(A)** and evening **(B)** milking intervals from 9 to 15 h, respectively. M0 = daily milk yield (DMY) estimated as two times AM or PM yield; M2A = linear regression of the difference between morning (AM) and evening (PM) milk yields on milking interval and days in milk (DIM); M3A = linear regression of DMY on the milking interval and DIM; M7A = linear regression of AM or PM proportion of DMY on milking interval and DIM (Wiggans, 1986); M8A = exponential regression model (Wu et al., 2022).

Overall, the joint model had a smaller standard deviation of model parameters because the size of the data used to estimate these parameters doubled. Therefore, the joint analysis improved the precision of estimated model parameters by pooling AM and PM milk yields. Nevertheless, the accuracies of estimated DMY from separate analyses for AM or PM milkings increased only slightly compared to the joint analyses. Plots of actual and estimated DMY for the exponential regression model (M8A), implemented separately or jointly for AM and PM milkings, are shown in **Figure 5B,C** compared to the mode M0 (**Figure 5A**). The plots showed slight stratification between AM and PM milkings, which explained why separate analyses had better, although slightly, linear regression fits between the actual and estimated DMY than joint analyses. For the model M8A, separate analyses had smaller intercepts and the regression coefficient was closer to 1, indicating improved accuracies with the separate analyses. However, the extent of improved accuracies was very slight. The $R^2$ accuracy was 0.9151 for the separate accounting and 0.9147 for the joint analysis; both rounded to 0.915. Here, we show that the accuracy obtained from the in-sample evaluation was higher than that (0.912) from the 10-fold cross-validation. Similarly, the $R^2$ accuracies from separate analyses were almost identical to or slightly better than joint analyses for the other models. For example, the $R^2$ accuracies were 0.9040 with the joint analysis and 0.9042 with the separate analysis for the model M2A, 0.9128 (joint) and 0.9131 (separate) for the model M3A, and 0.9062 (joint) and 0.9063 (separate) for the model M7A. The differences in the $R^2$ accuracies were seen only in the third or fourth decimal points. Compared to model M8A, model M0 had considerably larger intercepts and the regression coefficients

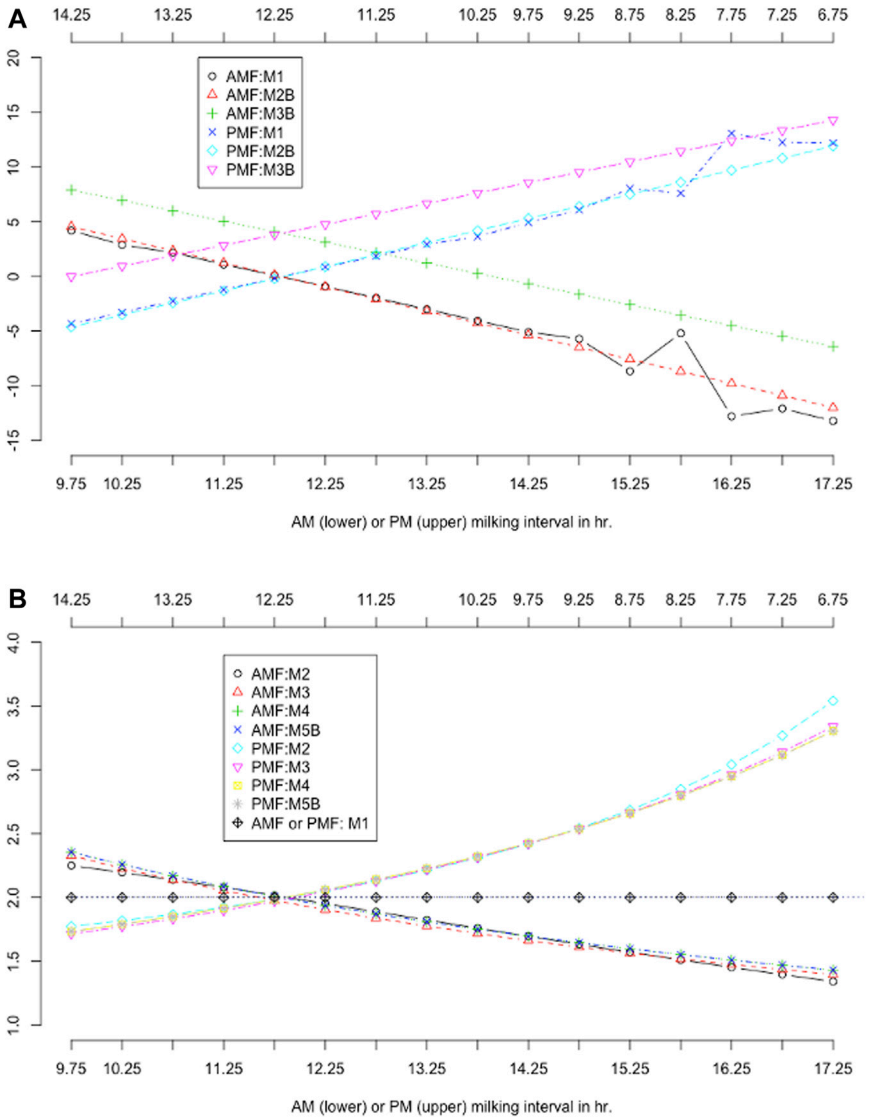


**FIGURE 7 |** Comparison of additive correction factors **(A)** and multiplicative correction factors **(B)** obtained using different models. AMF = morning milk yield correction factors; PMF = evening milk yield correction factors. M0 = daily milk yield (DMY) estimated as two times AM or PM yield; M1 = additive correction factors (ACF) model with categorical milking interval (MIC) and lactation months; M2B = ACF model with continuous milking interval and days in milk (DIM); M3B = linear regression of DMY on milking interval and DIM, with ACFs computed for discretized MIC; M5 = multiplicative correction factor (MCF) model according to Shook et al. (1980); M6 = MCF model according to DeLorenzo and Wiggans (1986); M7B = MCF model according to Wiggans (1986); M8B = MCF model based on the exponential regression model (Wu et al., 2022).

deviated substantially from 1.0. In other words, the model M8A has improved the DMY accuracies substantially compared to model M0. Similar conclusions hold all the ACF and MCF models and linear regression models, compared to doubling AM or PM yield as the daily yields.

Average DMY by milking intervals between 9 and 15 h were computed based on the estimated model parameters by joint analyses for the four selected models (M2A, M3A, M7A, and M8A), compared to the model M0 and the CSS means of actual DMY over milking interval (**Figure 6**). All the methods gave an average DMY comparable to the CSS means when AM and PM milking intervals were equal (12–12 h for AM and PM milking intervals). Still, they showed larger deviations with unequal AM and PM milkings. The model M0 had the largest deviations from the CSS means of DMY. Overall, the model M0 underestimated DMY with milking interval <12 h and overestimated DMY with milking interval >12 h. The more the AM (PM) milking interval departed from 12 h, the larger its deviation from the actual DMY. For the model M0, the average absolute deviation from the SCC means was 3.23 kg in Holstein and Jersey cows. Nevertheless, the deviations were much smaller for the ACF models (M2A and M3A) and the MCF models (M7A and M8A). The exponential regression model M8A had the smallest average absolute deviations from the CSS means of DMY (0.543 kg in the

Holstein cows and 0.598 in the Jersey cows). For the other models M2A, M3A, and M7A, the average absolute deviation from the CSS mean varied from 0.568 (M3A) to 0.773 (M2A) in the Holstein cows and from 0.649 (M3A) to 0.914 (M2A) in the Jersey cows. These results also showed that the relation between the smoothed average DMY and milking interval time from 9 h to 15 h was not precisely linear (**Figure 6**). Early studies showed that DMY (including fat and solid-not-fat) were not linear with intervals beyond 12 h (Ragsdale et al., 1924; Bailey et al., 1955; Elliott and Brumby, 1955; Schmidt, 1960). In particular, Atashi and Hostens (2021) showed that milk and component productions, in relation to the interval between the current milking and the previous milking, showed an exponential increase at the beginning and later leveled off to an asymptote. This exponential behavior for milk production was assumed to be the result of cell degradation and milk present in the udder (Neal and Thornley, 1983).

## Comparing Additive and Multiplicative Correction Factors

Additive and multiplicative factors were computed based on the parameter values of the data density functions or smoothing functions. Plots of ACF and MCF by MIC are shown in **Figure 7**. The ACF models were implemented with slightly different model assumptions, yet they resulted in drastically different ACF values in two groups (**Figure 7A**). The two classic ACF models, M1 and M2B, equivalently assumed a fixed regression coefficient of 2.0 for AM or PM milk yield. Hence, both models gave roughly comparable ACF per MIC, except that ACF from M2B were smoothed, but those from M1 were not (**Figure 7A**). The M1 model had considerably large fluctuations of ACF when the milking interval was less than 9 h or greater than 15 h due to insufficient milking records. Instead, the model M2B fitted the data on a continuous variable for milking interval, and ACF were computed on discretized MIC regardless of the data size for a specific MIC. Hence, the model M2B was robust to insufficient milking records per MIC, provided that the data are sufficient in general. Within MIC, the sum of AM and PM ACF for each model was close to zero (which ranged −0.031 with M1 to −0.108 with M2B). The ACF computed from the linear regression model (M3B) were considerably larger than those based on the two ACF models (M1 and M2B). This was because the estimated regression coefficients (approximately 1.75; **Table 4**) from the linear regression models were less than the fixed regression coefficients (2.0) assumed in the ACF model. Hence, the classic ACF models provided additive adjustments to two times AM or PM milk yields as the estimated DMY. Still, the linear regression models provided additive adjustments to approximately 1.75 times AM or PM milk yields. Because of this difference, the ACF from the linear regression model should be larger than those from the ACF models. The sum of AM and PM ACF within MIC was greater than zero (i.e., 8.24 kg) for the model M3B, with the average ACF being 4.11 kg, in the Holstein cows. The average ACF from the linear regression model can be verified as follows. In the Holstein cows, the average AM and PM milk yields were 16.4 and 15.3 kg, respectively. The regression coefficients for AM and PM milk yields by the separate analyses were 1.72 and 1.78, respectively. Hence, the average difference in ACF between the linear regression model and the ACF model was approximately estimated to be

$$(2.0 - 1.72) * 16.4 + (2.0 - 1.78) * 15.3 \approx 4.0.$$

With equal (12–12 h) AM and PM milking intervals, the ACF obtained from the M1 and M2B models were all close to zero (0.09–0.123 kg in Holstein cows and -0.67–0.41 kg in Jersey cows). Because these two models each assumed a fixed regression coefficient of 2.0 for the AM or PM milk yield, we concluded that doubling AM or PM milk yields provided an approximate estimate of DMY with equal AM and PM milking intervals. Put in another way. With equal AM and PM milking intervals, the additive correction amount was zero beyond two-time AM or PM milk yield as the estimated DMY. The results agreed with some early studies. For example, Everet and Wadell (1970b) showed that the mean AM excluding PM milk production was -0.51–0.19 kg in Holstein cows and −0.35–0.27 kg in Jersey cows with approximately equal AM and PM milking intervals (720–749 min). In the present study, the average AM minus PM milk yield was 1.13 kg in Holstein cows and 1.75 kg in Jersey cows. Similarly, Everet and Wadell (1970b) reported that the average AM minus PM milk yield in Holstein cows was 1.28 kg in Holstein cows and 0.89 kg in Jersey cows. Both studies agreed with each other concerning the average AM minus PM milk yield, despite a 50-year gap. Nevertheless, the ranges of AM minus PM milk yield (and ACF) in our study were significantly larger than the ranges in Everet and Wadell (1970b) because daily milk production has increased considerably over the past decades.

Unlike the ACF model, the MCF models implemented substantially different modeling strategies (**Table 2**). Nevertheless, the computed MCF from various models all corresponded to ratios of daily-to-single milk yields, despite their statistical interpretations varied (Wu et al., 2022). Hence, MCF obtained using various methods were approximately comparable in the Holstein cows (**Figure 7B**). MCF agreed well between the four MCF models given AM milking between 11 and 15 h, or PM milking between 9 and 13 h. Yet, large differences were observed out of this range. MCF were approximately 2.0 when AM and PM milking intervals were both 12 h. The AM MCF was greater than 2.0 when the AM milking interval was less than 12, and it was less than 2.0 when the AM milking interval was greater than 12 h. A precisely opposite trend was observed with the PM MCF. These results again suggested that two times AM or PM milk yield was an approximate estimate of DMY with equal AM and PM milking intervals. Still, such approximation did not hold with uneven AM and PM milking intervals. Similar results were observed in Jersey cows as well.

# CONCLUSION

Estimated milk yields by doubling AM or PM milk yields were taken approximately assuming equal AM and PM milking intervals, but they were subject to large errors when AM and PM milking intervals were unequal. The more deviations of AM and PM milking intervals from 12–12 h, the larger errors it generated. ACF and MCF provided effective adjustments to the estimated DMY with unequal AM and PM milkings. ACF provided additive adjustments, evaluated by the expected difference between AM and PM milk yield for each MIC and other categorical variables when applicable. An ACF model equivalently assumed a fixed multiplier (2.0) for AM or PM milk yields. In reality, ACF models with many discrete variables are challenged by insufficient or missing data points for specific MIC categories. Similarly, a linear regression model was implemented as an ACF model which nevertheless estimated the multiplier (regression coefficient) for AM or PM milk yield from the data. Relaxing the limitation on the fixed multiplier for AM and PM milkings allowed linear regression models to fit and predict the data better than ACF models. Multiplicative correction factors were computed by ratios of daily yield to yield from a single milking. Thus, multiplying a known AM or PM yield by an MCF gave an estimated DMY. Overall, the MCF models outperformed the ACF models, providing more accurate DMY estimates in the Holstein and Jersey cows. Nevertheless, computed ACF or MCF on discretized milking interval time suffered from losing information, leading to larger errors and lower accuracies. The exponential regression model (Wu et al., 2022) had the smallest MSE and the greatest accuracies of DMY estimates. This new model is analogous to an exponential growth (or decay) function for DMY with the observed yield from single milking as the initial state and the change rate tuned by a linear function of milking interval and other variables when applicable. This exponential regression model provides a promising alternative tool for estimating DMY.

The present study represented a preliminary effort to revisit the existing statistical methods for estimating DMY, compared to the newly proposed exponential regression model, using milking data collected between 2006 and 2009. In a continuing effort, large-scaled high-resolution milking data are being collected for follow-up studies, jointly supported by the US Council on Dairy Cattle Breeding, the USDA Agricultural Genomics and Improvement Laboratories, and the National Dairy Herd Information Association. This is a 3-year data collection project. We expect that MCF in use will be updated by then. Finally, we illustrated the methods for estimating DMY in AM and PM milking plans. Yet, these methods and principles are generally applicable, either directly or with necessary modifications, to cows milked more than two times a day.

## DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: The Holstein and Jersey milking record data maintained by CDCB are not publicly available, but can be requested to JD, subject to signing an official agreement for non-commercial use only. Requests to access these datasets should be directed to Joao Durr, Joao, durr@uscdcb.com.

## AUTHOR CONTRIBUTIONS

X-LW conceived and designed this study, with a series of discussion meetings with GW, HDN, AM, CVT, RB, JB, and JD. X-LW carried out the data extraction and data analyses and drafted this manuscript. All authors have reviewed and approved the final manuscript.

## FUNDING

This project was financially supported by CDCB, a non-profit organization for dairy genetic/genomic evaluations and data storage.

## REFERENCES

Bailey, G. L., Clough, P. A., and Dodd, F. H. (1955). The Rate of Secretion of Milk and Fat. *J. Dairy Res.* 22, 22–36. doi:10.1017/s0022029900007512

Cole, J. B., Null, D. J., and VanRaden, P. M. (2009). Best Prediction of Yields for Long Lactations. *J. Dairy Sci.* 92, 1796–1810. doi:10.3168/jds.2007-0976

Craven, P., and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions. *Numer. Math.* 31, 377–403.

DeLorenzo, M. A., and Wiggans, G. R. (1986). Factors for Estimating Daily Yield of Milk, Fat, and Protein from a Single Milking for Herds Milked Twice a Day. *J. Dairy Sci.* 69, 2386–2394. doi:10.3168/jds.s0022-0302(86)80678-6

Elliott, G. M., and Brumby, P. J. (1955). Rate of Milk Secretion with Increasing Interval between Milking. *Nature* 176, 350–351. doi:10.1038/176350a0

Everett, R. W., and Wadell, L. H. (1970a). Relationship between Milking Intervals and Individual Milk Weights. *J. Dairy Sci.* 53, 548–553. doi:10.3168/jds.s0022-0302(70)86251-8

Everett, R. W., and Wadell, L. H. (1970b). Sources of Variation Affecting the Difference between Morning and Evening Daily Milk Production. *J. Dairy Sci.* 53, 1424–1429. doi:10.3168/jds.s0022-0302(70)86410-4

Geisser, S. (1993). *Predictive Inference*. New York, NY: Chapman & Hall.

Hawkins, D. M., and Douglas, M. (2004). The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12. doi:10.1021/ci0342472

Lien, D., Hu, Y., and Liu, L. (2017). A Note on Using Ratio Variables in Regression Analysis. *Econ. Lett.* 150, 114–117. doi:10.1016/j.econlet.2016.11.019

Liu, Z., Reents, R., Reents, R., Reinhardt, F., and Kuwan, K. (2000). Approaches to Estimating Daily Yield from Single Milk Testing Schemes and Use of a.m.-p.M. Records in Test-Day Model Genetic Evaluation in Dairy Cattle. *J. Dairy Sci.* 83, 2672–2682. doi:10.3168/jds.s0022-0302(00)75161-7

McDaniel, B. T. (1973). Merits and Problems of Adjusting to Other Than Mature Age. *J. Dairy Sci.* 56, 959–967. doi:10.3168/jds.s0022-0302(73)85286-5

Neal, H. D. S. C., and Thornley, J. H. M. (1983). The Lactation Curve in Cattle: A Mathematical Model of the Mammary Gland. *J. Agric. Sci.* 101, 389–400. doi:10.1017/s0021859600037710

Norman, H. D., Meinert, T. R., Schultz, M. M., and Wright, J. R. (1995). Age and Seasonal Effects on Holstein Yield for Four Regions of the United States over Time. *J. Dairy Sci.* 78, 1855–1861. doi:10.3168/jds.s0022-0302(95)76810-2

Porzio, G. (1953). A New Method of Milk Recording. *An. Br. Abstr.* 21, 344.

Putnam, D. N., and Gilmore, H. C. (1970). Factors to Adjust Milk Production to a 24-hourbasiswhen Milking Intervals Are Unequal. *J. Dairy Sci.* 53, 685.

Putnam, D. N., and Gilmore, H. C. (1968). The Evaluation of an Alternate AM-PM Monthly Testing Plan and its Application for Use in the DHIA Program. *J. Dairy Sci.* 51, 985.

Ragsdale, A. C., Turner, C. W., and Brody, S. (1924). The Rate of Milk Secretion as Affected by an Accumulation of Milk in the Mammary Gland. *J. Dairy Sci.* 7, 249–254. doi:10.3168/jds.s0022-0302(24)94019-7

Schaeffer, L. R., Jamrozik, J., Van Dorp, R., Kelton, D. F., and Lazenby, D. W. (2000). Estimating Daily Yields of Cows from Different Milking Schemes. *Livest. Prod. Sci.* 65, 219–227. doi:10.1016/s0301-6226(00)00153-6

Schmidt, G. H. (1960). Effect of Milking Intervals on the Rate of Milk and Fat Secretion. *J. Dairy Sci.* 43, 213–219. doi:10.3168/jds.s0022-0302(60)90143-0

Schutz, M. M., Bewley, J. M., and Norman, H. D. (2010). Derivation of Factors to Estimate Daily Milk Yield from One Milking of Cows Milked Three Times Daily. *J. Dairy Sci.* 93 (E-Suppl. 1), 595.

Schutz, M. M., Bewley, J. M., and Norman, H. D. (2008). Derivation of Factors to Estimate Daily Yield from Single Milkings for Holsteins Milked Two or Three Times Daily. *J. Dairy Sci.* 91 (E-Suppl. 1), 106–107.

Schutz, M. M., and Norman, H. D. (1994). Adjustment of Jersey Milk, Fat, and Protein Records across Time for Calving Age and Season. *J. Dairy Sci.* 72 (E-Suppl. 1), 267.

Schutz, M. M., and Norman, H. D. (2011). Verification of Factors to Estimate Daily Milk Yield from One Milking of Cows Milked Twice Daily. *J. Dairy Sci.* 94 (E-Suppl. 1), 28–29.

Shook, G., Jensen, E. L., and Dickinson, F. N. (1980). Factors for Estimating Sample-Day Yield in Am-Pm Sampling Plans. *DHI Lett.* 56, 25–30.

Smith, J. W., and Pearson, R. E. (1981). Development and Evaluation of Alternate Testing Procedures for Official Records. *J. Dairy Sci.* 64, 466–474. doi:10.3168/jds.s0022-0302(81)82595-7

Stone, M. (1974). Cross-validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B Methodol.* 36, 111–133. doi:10.1111/j.2517-6161.1974.tb00994.x

VanRaden, P. M. (1997). Lactation Yields and Accuracies Computed from Test Day Yields and (Co)variances by Best Prediction. *J. Dairy Sci.* 80, 3015–3022. doi:10.3168/jds.s0022-0302(97)76268-4

Wiggans, G. R. (1986). Estimating Daily Yields of Cows Milked Three Times a Day. *J. Dairy Sci.* 69, 2935–2940. doi:10.3168/jds.s0022-0302(86)80749-4

Wu, X-L., Wiggans, G., Norman, H. D., Miles, A., Van Tassell, C. P., Baldwin, R. L. V. I., et al. (2022). *Daily Milk Yield Correction Factors: What Are They?* doi:10.48550/arXiv.2203.09606

**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.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 26 August 2022
DOI 10.3389/fgene.2022.963654

OPEN ACCESS

EDITED BY
Ran Di,
Institute of Animal Sciences (CAAS),
China

REVIEWED BY
Xiujin Li,
Zhongkai University of Agriculture and Engineering, China
Peipei Ma,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Xiangdong Ding,
xding@cau.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 07 June 2022
ACCEPTED 01 August 2022
PUBLISHED 26 August 2022

CITATION
Jiang Y, Song H, Gao H, Zhang Q and Ding X (2022), Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals.
*Front. Genet.* 13:963654.
doi: 10.3389/fgene.2022.963654

# Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals

Yifan Jiang[1†], Hailiang Song[2†], Hongding Gao[3], Qin Zhang[4] and Xiangdong Ding[1*]

[1]National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, College of Animal Science and Technology, China Agricultural University, Beijing, China, [2]Beijing Key Laboratory of Fisheries Biotechnology, Fisheries Science Institute, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [3]Natural Resources Institute Finland (Luke), Helsinki, Finland, [4]Shandong Provincial Key Laboratory of Animal Biotechnology and Disease Control and Prevention, Shandong Agricultural University, Taian, China

Genotype imputation from BeadChip to whole-genome sequencing (WGS) data is a cost-effective method of obtaining genotypes of WGS variants. Beagle, one of the most popular imputation software programs, has been widely used for genotype inference in humans and non-human species. A few studies have systematically and comprehensively compared the performance of beagle versions and parameter settings of farm animals. Here, we investigated the imputation performance of three representative versions of Beagle (Beagle 4.1, Beagle 5.0, and Beagle 5.4), and the effective population size (Ne) parameter setting for three species (cattle, pig, and chicken). Six scenarios were investigated to explore the impact of certain key factors on imputation performance. The results showed that the default Ne (1,000,000) is not suitable for livestock and poultry in small reference or low-density arrays of target panels, with 2.47%–10.45% drops in accuracy. Beagle 5 significantly reduced the computation time (4.66-fold–13.24-fold) without an accuracy loss. In addition, using a large combined-reference panel or high-density chip provides greater imputation accuracy, especially for low minor allele frequency (MAF) variants. Finally, a highly significant correlation in the measures of imputation accuracy can be obtained with an MAF equal to or greater than 0.05.

KEYWORDS

**imputation, accuracy, whole genome sequencing, livestock, poultry**

## 1 Introduction

Genotype imputation (Yun et al., 2009), which uses linkage disequilibrium knowledge from haplotypes of a known reference panel to predict genotypes of missing or ungenotyped markers, is a commonly used procedure for obtaining more genotypes. This is achieved by imputing low-to high-density single nucleotide polymorphism (SNP) markers, and even whole-genome sequencing (WGS) SNP markers. It has played a crucial role in whole-genome studies such as genomic selection (GS) (Vanraden et al., 2017; Raymond et al., 2018; Zhang et al., 2018) and genome-wide association studies (GWAS) (Jonathan and Bryan, 2010; Kelemen et al., 2015; Yan et al., 2017; van den Berg et al., 2019). The availability of next-generation sequencing techniques has made it possible to obtain WGS and SNP markers at reasonable cost. However, sequencing all individuals is not realistic in livestock and poultry breeding programs. Thus, one of the most used strategies is to sequence a subset of a population that is used as a reference panel to perform genotype imputation with high accuracy. For example, using the comprehensive reference panels provided by the 1000 Genomes Project and 1000 Bull Genomes Project consortium to impute to whole-genome-level SNPs has recently become more common in humans and other genomic studies (Kelemen et al., 2015; Liu et al., 2015; Pausch et al., 2016).

Since its first release in 2009 (Browning and Browning, 2009), Beagle has been widely used for genotype imputation and phasing. Beagle uses Bayesian methods with the Markov Chain Monte Carlo (MCMC) algorithm. As one of the most popular imputation software programs, it has been widely used in humans and non-human species, such as cattle (Frischknecht et al., 2017), dogs (Jenkins et al., 2021), pigs (Yang et al., 2021; Li et al., 2022), and chickens (Ye et al., 2019a; Li et al., 2020b), etc. In the past 13 years, it has been continuously updated from Beagle 3 to Beagle 5.4 (as of 25 May 2022). Beagle 4.1 was developed for genotype imputation of millions of reference samples (Browning and Browning, 2016). Beagle 5.0 was developed to further reduce the computational cost of imputation from large reference panels (Browning et al., 2018). Since version 5.2, Beagle has employed a two-stage phasing algorithm to make it faster and more memory efficient (Browning et al., 2021). However, the differences of these version, and their effects of the parameter settings on livestock and poultry, have not been fully compared. Research have shown that the parameter effective population size (Ne) has the greatest impact on the error rate of imputation in chicken and maize populations (Pook et al., 2020). Thus, the effect of Ne on the imputation accuracy is considered in our study.

Factors affecting imputation accuracy, such as reference panel size and chip density, have already been studied based on both simulated and empirical data (Pausch et al., 2013; Ventura et al., 2016; Pausch et al., 2017). However, most of them were carried out with default parameters and were not intended to compare different imputation programs or parameter settings (Zheng et al., 2012; Pausch et al., 2013; Ventura et al., 2016), and the calculation of imputation accuracy is not similar between studies. For example, in some studies, only random masked sites were used for the calculation of imputation, and some used all imputed sites but only a part of the individuals (Frischknecht et al., 2017; Ye et al., 2018; Yuan et al., 2018). In addition, the commonly used measures of genotype imputation accuracy include genotype concordance, the correlation between imputed and true genotypes, and Allele R-Squared (AR2) and Dosage R-Squared (DR2) in different versions of Beagle (Pausch et al., 2017; Rowan et al., 2019; Song et al., 2019; van den Berg et al., 2019). Some studies only used one method to measure, which made the reliability of comparison between the studies low. Therefore, it is crucial to devise an optimal strategy for improving the accuracy of genotype imputation in GS and GWAS studies, or in livestock and poultry breeding programs, regardless of the chip density in the target panel. We performed a comprehensive and systematic investigation of these factors on imputation accuracy across three species: cattle, pigs, and chickens.

In the current study, we investigated the performance of three representative versions of Beagle (Beagle 4.1, Beagle 5.0, and Beagle 5.4) and the effects of parameter settings on three farm animals (cattle, pigs, and chickens) to devise an optimal strategy from the SNP array to whole genome sequencing data of livestock and poultry. In addition, we explored the effects of chip density, reference population size, and the relationship between the target panel and the reference panel on imputation accuracy. Finally, the correlation between the measures of imputation accuracy and minor allele frequency (MAF) was also explored.

# 2 Materials and methods

## 2.1 Whole genome sequencing data and BeadChip data

WGS and BeadChip data based on three livestock and poultry, including cattle, pigs, and chickens, were used in this study. The framework of the genotype imputation is shown in Figure 1. The detailed information is as follows.

### 2.1.1 Cattle

WGS data, of Beagle-phased SNP calls, were obtained from RUN 5 of the 1000 Bull Genomes Project, released in 2017 (Daetwyler et al., 2014). A total of 1,682 whole-genome sequenced animals were provided by the 1000 Bull Genomes Project (Run 5), which included 1,602 *Bos taurus*, 53 *Bos indicus*, and 27 Chinese yellow cattle (Daetwyler et al., 2014). Detailed information regarding the breeds of the animals used is provided in Supplementary Table S1. A total of 67.33 million variants were discovered in these animals, of which 64.80 million were SNPs and 2.53 million were indels. Further details about variant calling, genotyping, and filtering of variants, in the 1000 Bull Genomes Project, were presented by Daetwyler et al. (2014).

Genotype imputation included two panels: the reference panel and the target panel. In the genotype imputation analysis scenarios we investigated, two main target panels and two main reference panels were considered. One of the target groups consisted of 100 Holstein cattle, randomly selected from

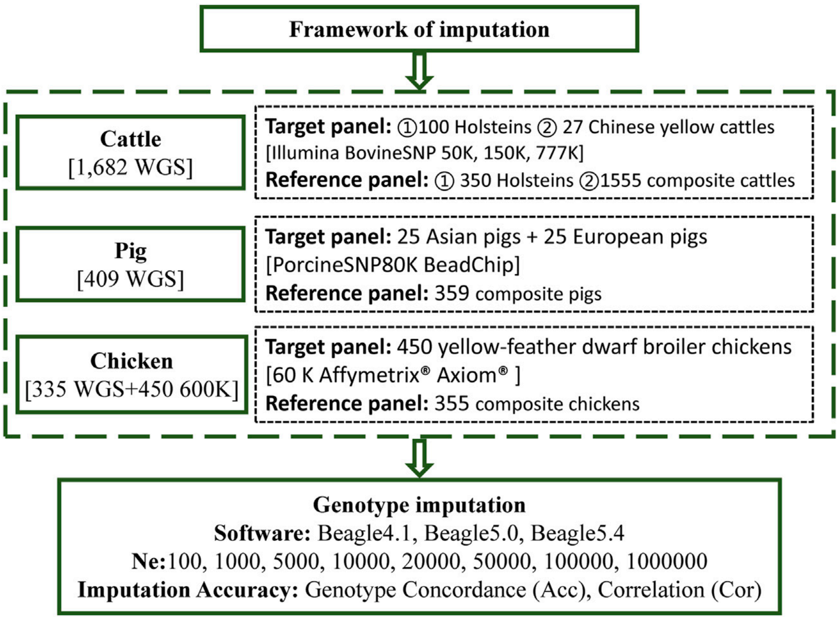

Framework of imputation
Cattle
[1,682 WGS]
Target panel: ①100 Holsteins ② 27 Chinese yellow cattles
[Illumina BovineSNP 50K, 150K, 777K]
Reference panel: ① 350 Holsteins ②1555 composite cattles
Pig
[409 WGS]
Target panel: 25 Asian pigs + 25 European pigs
[PorcineSNP80K BeadChip]
Reference panel: 359 composite pigs
Chicken
[335 WGS+450 600K]
Target panel: 450 yellow-feather dwarf broiler chickens
[60 K Affymetrix® Axiom® ]
Reference panel: 355 composite chickens
Genotype imputation
Software: Beagle4.1, Beagle5.0, Beagle5.4
Ne:100, 1000, 5000, 10000, 20000, 50000, 100000, 1000000
Imputation Accuracy: Genotype Concordance (Acc), Correlation (Cor)

FIGURE 1
The framework of the imputation.

TABLE 1 Number of SNPs used across chromosomes under different panels in cattle.

| Chr (Cattle) | Chr length (bp) | Reference panel | | Target panel | | | IMP sites (ref350) | | | IMP sites (ref1555) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ref350 | ref1555 | 50 K | 150 K | 777 K | 50 K | 150 K | 777 K | 50 K | 150 K | 777 K |
| chr1 | 158,337,067 | 1,265,065 | 3,068,377 | 3,067 | 6,781 | 39,186 | 1,262,134 | 1,258,329 | 1,229,993 | 3,065,312 | 3,061,599 | 3,029,193 |
| chr7 | 112,638,659 | 847,075 | 2,063,921 | 2,064 | 5,386 | 28,133 | 845,108 | 841,843 | 822,367 | 2,061,858 | 2,058,537 | 2,035,791 |
| chr21 | 71,599,096 | 562,318 | 1,387,487 | 1,296 | 3,071 | 17,712 | 561,083 | 559,274 | 546,263 | 1,386,192 | 1,384,416 | 1,369,777 |
| chr29 | 51,505,224 | 470,173 | 1,101,854 | 962 | 2,190 | 12,038 | 469,260 | 467,998 | 458,834 | 1,100,892 | 1,099,665 | 1,089,816 |
| Total | 394,080,046 | 3,144,631 | 7,621,639 | 7,389 | 17,428 | 97,069 | 3,137,585 | 3,127,444 | 3,057,457 | 7,614,254 | 7,604,217 | 7,524,577 |

Chr, chromosome; IMP sites, imputed sites and locus used to calculate imputation accuracy.

the 450 Holstein cattle in the sequencing data, and the other consisted of 27 Chinese yellow cattle. Correspondingly, the remaining 350 Holsteins served as the reference panel of purebreds (ref350). All the remaining 1,555 cattle in the 1000 Bull Genomes Project, RUN5, served as a composite reference panel (ref1555).

To investigate the influence of different imputation scenarios, the genotypes of the target panel were masked, using bovine chips of different densities, to mimic the scenario from which the animals were genotyped. The low-, medium-, and high-density chips corresponded to Illumina BovineSNP 50, 150, and 777 K BeadChip chips, of cattle with 54,609, 138,892, and 777,962 SNPs, respectively. In addition, the positions of all SNPs, in BeadChip, were based on the *B. taurus* UMD3. We used one reference genome (Zimin et al., 2009) obtained from the UCSC liftover (http://genome.ucsc.edu/cgi-bin/hgLiftOver), which was consistent with a genome of the 1000 Bull Genomes Project RUN5. After removing variants with minor allele counts (less than one), and variants with more than two alleles across all reference individuals, Table 1 presents detailed information on variants in the imputation. Then, imputation from chip variants to whole genome sequence variants was performed and the imputation accuracy of imputation (IMP) sites of the target population was compared to the WGS data of these target individuals.

TABLE 2 Number of SNPs used across chromosomes in pigs and chickens.

| Species | Chr | Chr length (bp) | Reference panel | Target panel | IMP sites |
|---|---|---|---|---|---|
| Pig | chr1 | 315,321,322 | 2,756,826 | 5,014 | 2,751,812 |
| | chr6 | 157,765,593 | 1,782,136 | 3,693 | 1,778,443 |
| | chr12 | 63,588,571 | 893,925 | 2,138 | 891,787 |
| | chr18 | 61,220,071 | 879,515 | 1,439 | 878,076 |
| | Total | 597,895,557 | 6,312,402 | 12,284 | 6,300,118 |
| Chicken | chr1 | 196,202,544 | 7,158,664 | 9,841 | 80,339 |
| | chr3 | 111,302,122 | 4,079,325 | 5,506 | 44,859 |
| | chr6 | 35,467,016 | 1,479,613 | 2,117 | 17,537 |
| | chr28 | 4,974,273 | 190,787 | 534 | 4,187 |
| | Total | 347,945,955 | 12,908,389 | 17,998 | 146,922 |

Chr, chromosome; IMP sites, imputed sites and locus used to calculate imputation accuracy.

TABLE 3 Scenarios used to evaluate imputation performance.

| Scenario | Description | Species | Target panel | Reference panel | Software | Ne |
|---|---|---|---|---|---|---|
| S1 | Effects of beagle version and Ne parameter size on imputation accuracy in three species | Cattle | 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle4.1, Beagle5.0, Beagle5.4 | 100, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000, 1,000,000 |
| | | Pig | 25 Asian pigs + 25 European pigs (80 K) | 359 pigs | Beagle4.1, Beagle5.0, Beagle5.4 | 100, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000, 1,000,000 |
| | | Chicken | 450 yellow-feather dwarf broiler chickens (60 K) | 355 chickens | Beagle4.1, Beagle5.0, Beagle5.4 | 100, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000, 1,000,000 |
| S2 | Chip density and reference panel size on the imputation accuracy | Cattle | 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle4.1, Beagle5.0, Beagle5.4 | 100,000 |
| S3 | Imputation accuracy against minor allele frequency | Cattle | 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle5.4 | 100,000 |
| S4 | The relationship of the measure of imputation accuracy (Acc, Cor, AR2, DR2) | Cattle | 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle4.1 (for AR2), Beagle5.4 | 100,000 |
| S5 | The relationship between target panel and reference panel on the imputation accuracy | Cattle | 27 Chinese yellow cattle (50, 150, 777 K) and 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle5.4 | 100,000 |
| S6 | Time consuming | Cattle | 100 Holstein (50, 150, 777 K) | ref350, ref1555 | Beagle4.1, Beagle5.0, Beagle5.4 | 100,000 |

Ne, effective population size; AR2, allelic R-squared; DR2, dosage R-squared; Acc, genotype concordance; Cor: correlation.

### 2.1.2 Pig

WGS data for pigs were downloaded from the Genome Variation Map (GVM; http://bigd.big.ac.cn/gvm/) database, which collected and integrated genome variations for 47 species (as of 25 May 2022) (Li et al., 2020a). A total of 409 pigs, with 90.90 million SNPs (based on the *Sus scrofa* 10. 2 reference genome), were provided by GVM, which included 213 Asian pigs, 181 European pigs, and 15 *Sus* pig species (Supplementary Table S2). Variants with a missing rate of more than 0.2, and a minimum allele frequency of less than 0. 01, were removed for subsequent analysis. Phasing was executed using Beagle (version5.4) (Browning et al., 2021), with its default parameters. We randomly selected 25 European and 25 Asian pigs, as the target population and the remaining 359 pigs were a part of the reference population. The genotypes of the target panel were masked to a PorcineSNP80K BeadChip (Illumina, San

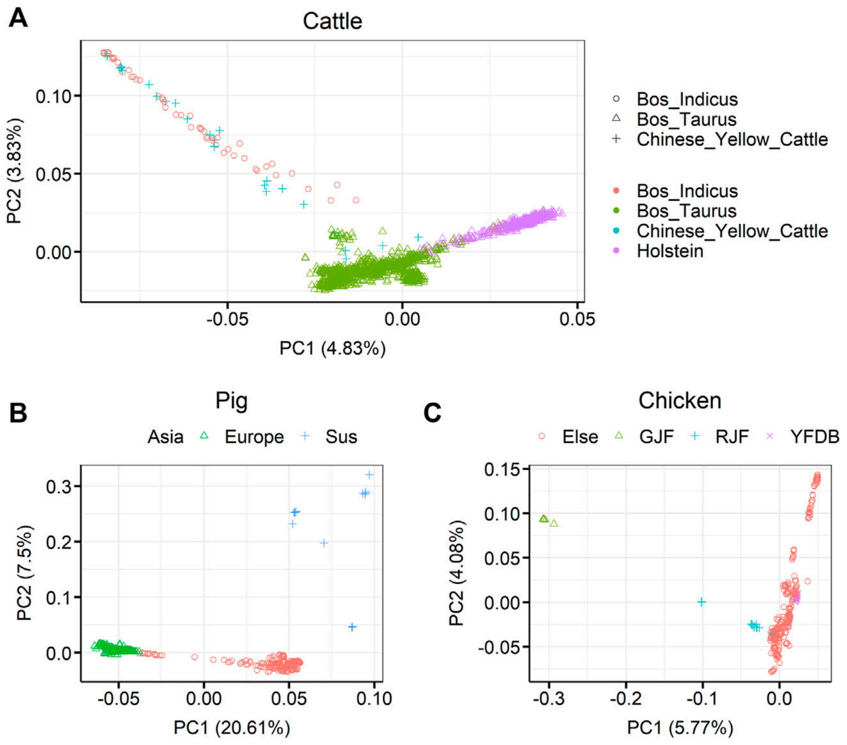


FIGURE 2
Principal component analysis (PCA) showing the population structure of the three farm animals (cattle, pigs, and chickens). **(A)** PCA showing the population structure of 1,682 sequenced cattle in the RUN5 of the 1000 bull genome project. **(B)** PCA showing the population structure of 409 sequenced pigs in genome variation map database **(C)** PCA showing the population structure of 335 sequenced chickens. GJF, green jungle fowl; RJF, red jungle fowl; YFDB, yellow feather dwarf broiler. Different colors and symbols represent different classes.

Diego, CA, United States). After imputation, like that of cattle, the imputation accuracy was calculated by comparing the IMP sites in the target population with the WGS data of target individuals. The statistics of the number of SNPs are listed in Table 2.

### 2.1.3 Chicken

This dataset was adopted from Ye et al.'s studies (Ye et al., 2018; Yuan et al., 2018). A total of 335 chickens were sequenced using WGS technology (based on the galGal5 reference genome), and 450 yellow-feather dwarf broiler chickens were genotyped using the 600 K Affymetrix® Axiom® high-density genotyping array (Supplementary Table S3). The WGS panel contains diverse breeds including red junglefowl, green junglefowl, Tibetan chickens, fighting chickens, white leghorn chickens and so on. It is worth mentioning that 24 key individuals of the yellow-feather dwarf broiler population were included in the 355 WGS populations. Following Ye et al. (2018), Ye et al. (2019b), the supposed 60 K chip data were generated by sampling the first SNP in each bin of adjacent 10 SNPs, of the 600 K SNP chip as the target panel for imputation. The 450 chickens with a 60 K BeadChip chip were used as the target panel, and the 335 WGS chickens were used as the reference panel for imputation. After the imputation was performed, the IMP sites coincident with 600 K were used to calculate the imputation accuracy, as shown in Table 2.

## 2.2 Genotype imputation strategy

To improve computational efficiency, four autosomes across large, medium, and small chromosomes, were separately selected for cattle (chr1, chr7, chr21, chr29), pig (chr1, chr6, chr12, chr18), and chicken (chr1, chr3, chr6, chr28). The variant information of the genotype imputation in this study is listed in Tables 1 and 2.

We compared the effect of Beagle versions, setting effective population size (Ne), chip density, reference panel sizing, and the relationship between the target and reference panels on imputation accuracy, as shown in Table 3. To explore the effect of the Beagle version and the parameter of effective population size (Ne) on the imputation accuracy of the three livestock and poultry, the imputation were performed by Beagle 4.1 (Beagle.27Jan18.7e1.jar) (Browning and Browning, 2016),

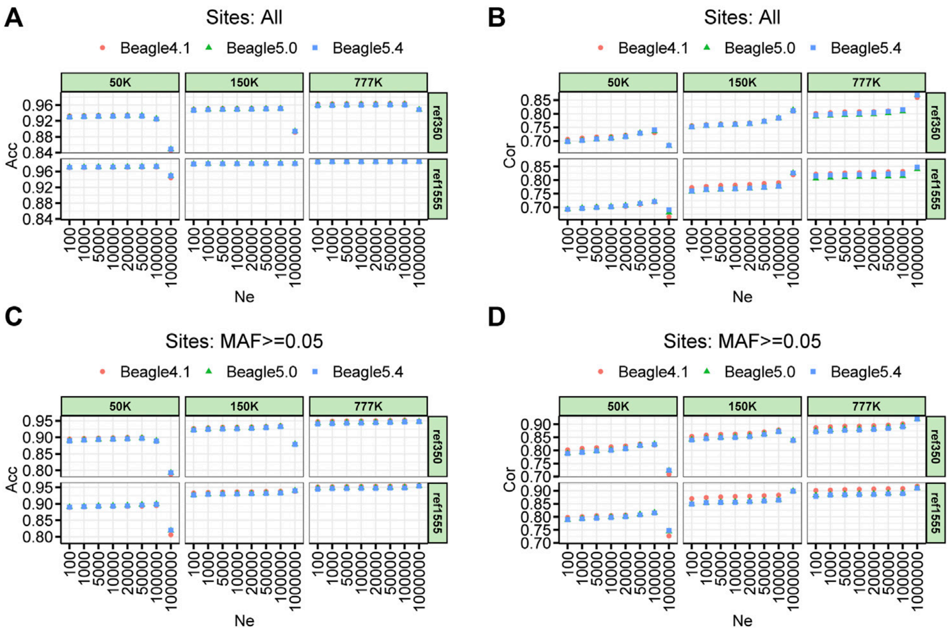


FIGURE 3
Accuracy of imputation for three density BeadChip chips, two reference population sizes and three imputation software with a range of effective population size (Ne) sets in cattle. **(A)** Imputation accuracy measured by the genotype concordance (Acc). **(B)** Imputation accuracy measured by the correlation (Cor) **(C,D)** corresponds to **(A)** and **(B)** with minor allele frequency sites less than 0.05 removed.

Beagle 5.0 (beagle.12Jul19.0df.jar) (Browning et al., 2018), and Beagle 5.4 (beagle.19Apr22.7c0.jar) (Browning et al., 2021) with the parameters of effective population size (Ne) set to 100, 1,000, 5,000, 10,000, 20,000, 50,000, 100,000, and 1,000,000 for the three livestock. In both Beagle 4.1 and Beagle 5.0, the default parameter of Ne was 1,000,000, but in Beagle 5.4, the default parameter of Ne was 100,000. Furthermore, in cattle populations, the effects of reference population size and chip density on imputation accuracy were also explored. Furthermore, using cattle as an example, we explored the relationship between the target and reference panels on the accuracy of genotype imputation with 27 Chinese yellow cattle as the target panel. Meanwhile, the imputation accuracy against minor allele frequency, the correlation of the measure of imputation accuracy, and the time used was explored using cattle datasets.

## 2.3 Evaluation of imputation accuracy

Two criteria were used to measure the imputation performance: 1) correlation between true and imputed genotypes (Cor), which were coded as 0, 1, and 2 for genotypes AA, AB, and BB, respectively; 2) genotype concordance (Acc), which was defined as the proportion of genotypes of the imputed variants that were the same as the true genotypes. In addition, Allele R-Squared (AR2, estimated squared correlation between the most probable REF dose and true REF dose) and Dosage-R2 (DR2, estimated squared correlation between estimated REF dose and true REF dose) output by Beagle (Beagle 4.1 generates both the AR2 and DR2, Beagle 5 only generates DR2) were also used to make a comparison of these imputation accuracy measurements.

## 2.4 Population structural analysis

The population structure was demonstrated by principal component analysis (PCA), using GCTA (version 1.92.0 beta2) software (Yang et al., 2011), and the first 20 eigenvectors were output and then plotted using the R program (Valero-Mora, 2010). Variants with an MAF of less than 0.05 were removed for this analysis.

# 3 Results

## 3.1 Population structure

Principal component analysis (PCA) for the three livestock, cattle, pig, and chicken, is shown in Figure 2. For cattle, it can be seen that *Bos taurus* and *Bos indicus* were first separated by PC1 in 1,682 individuals, and then the individuals were separated

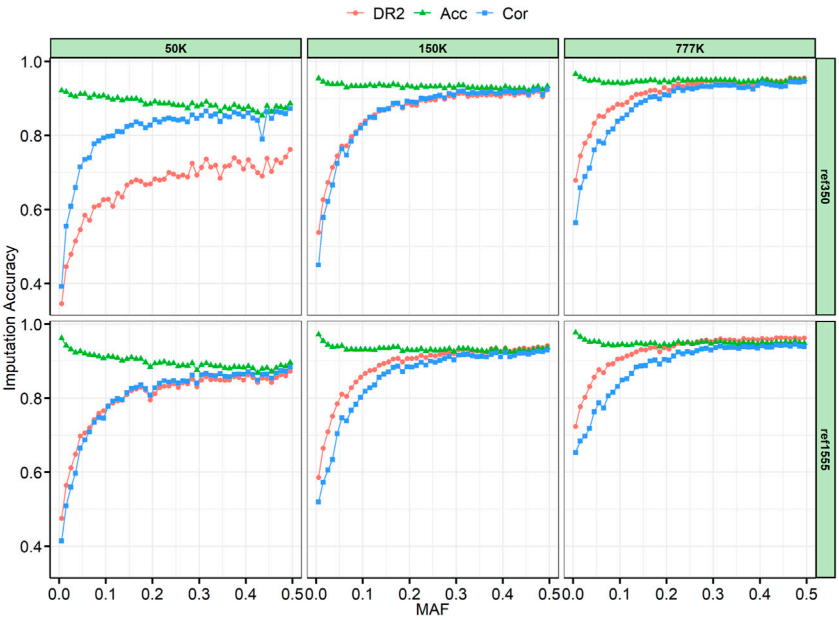


FIGURE 4
Imputation accuracy by minor allele frequency (MAF) class. The SNPs were divided into bins of 0.01 per increment according to their MAF. AR2, allelic R-squared; DR2, dosage R-squared; Acc, genotype concordance; Cor, correlation.

into *B. taurus*, *B. indicus*, or Chinese yellow cattle by PC2. Among the *B. taurus*, Holstein cattle had the largest number of individuals were Holstein cattle (450 samples). For pigs, it was clearly shown that European pigs, Asian pigs, and *Sus* species pigs were separate from 409 pigs. For chickens, it was clearly shown that red jungle fowl and green jungle fowl separate from the other samples in 335 chickens. The detailed breed compositions are presented in Supplementary Tables S1–S3.

## 3.2 Beagle versions and the parameter of the effective population size settings on imputation accuracy

The imputation accuracy of the parameter setting on the effective population size (Ne) of three different Beagle versions for cattle, pigs, and chickens are shown in Figure 3 and Supplementary Figures S1, S2, respectively. For the comparison of imputation accuracy of the different versions of Beagle, we found that the three versions of Beagle software achieved almost the same accuracy in different scenarios with only slight differences. Beagle 5.0 and Beagle 5.4 performed nearly the same imputation accuracy across all scenarios. Compared with Beagle 4.1, Beagle 5 (including Beagle 5.0 and Beagle 5.4) showed 0.1% and 0.6% improvement in Acc, and 0.4% and 0.9% improvement in Cor for pigs and chickens, respectively, when Ne was equal to 100,000 (Supplementary Figures S1, S2). Similarly, the imputation accuracy varies by a few tenths of thousands of beagles. However, the size of Ne has a significant impact on imputation accuracy. In the case where the default Ne size of Beagle 4.1 and Beagle 5.0 (Ne = 1,000,000), the imputation accuracy for the cattle's 50 K was significantly reduced, whether imputed with ref350 (Acc and Cor dropped by 7.72% and 5.28%, respectively, for all imputed sites; Acc and Cor dropped by 9.77% and 10.45%, respectively, for the imputed sites with MAF ≥ 0.05) or ref1555 (Acc and Cor dropped by 2.47% and 4.13%, respectively, for all imputed sites; Acc and Cor dropped by 8.24% and 7.55%, respectively, for the imputed sites with MAF ≥ 0.05). The imputation accuracy also decreased when the imputation was performed from the 150 K chip to the WGS, with ref350 (Acc dropped by 5.88% and no drop in Cor for all imputed sites; Acc and Cor dropped by 5.40% and 3.62%, respectively, for the imputed sites with MAF ≥ 0.05). The other panels in cattle imputation, such as imputation from 777 K to WGS and from 150 K to WGS with ref350, Ne had less impact on the accuracy of imputation (Figure 3). In addition, we noticed that in pig and chicken imputation, the default Ne size in Beagle 4.1 and Beagle 5.0 (Ne = 1,000,000) also reduced the imputation accuracy (Supplementary Figures S1, S2). All these results suggest that the impact of different Beagle versions on the imputation accuracy is small, but the default value of Ne has a great impact on the imputation accuracy, especially for the imputation of low-density chips or small reference panels.

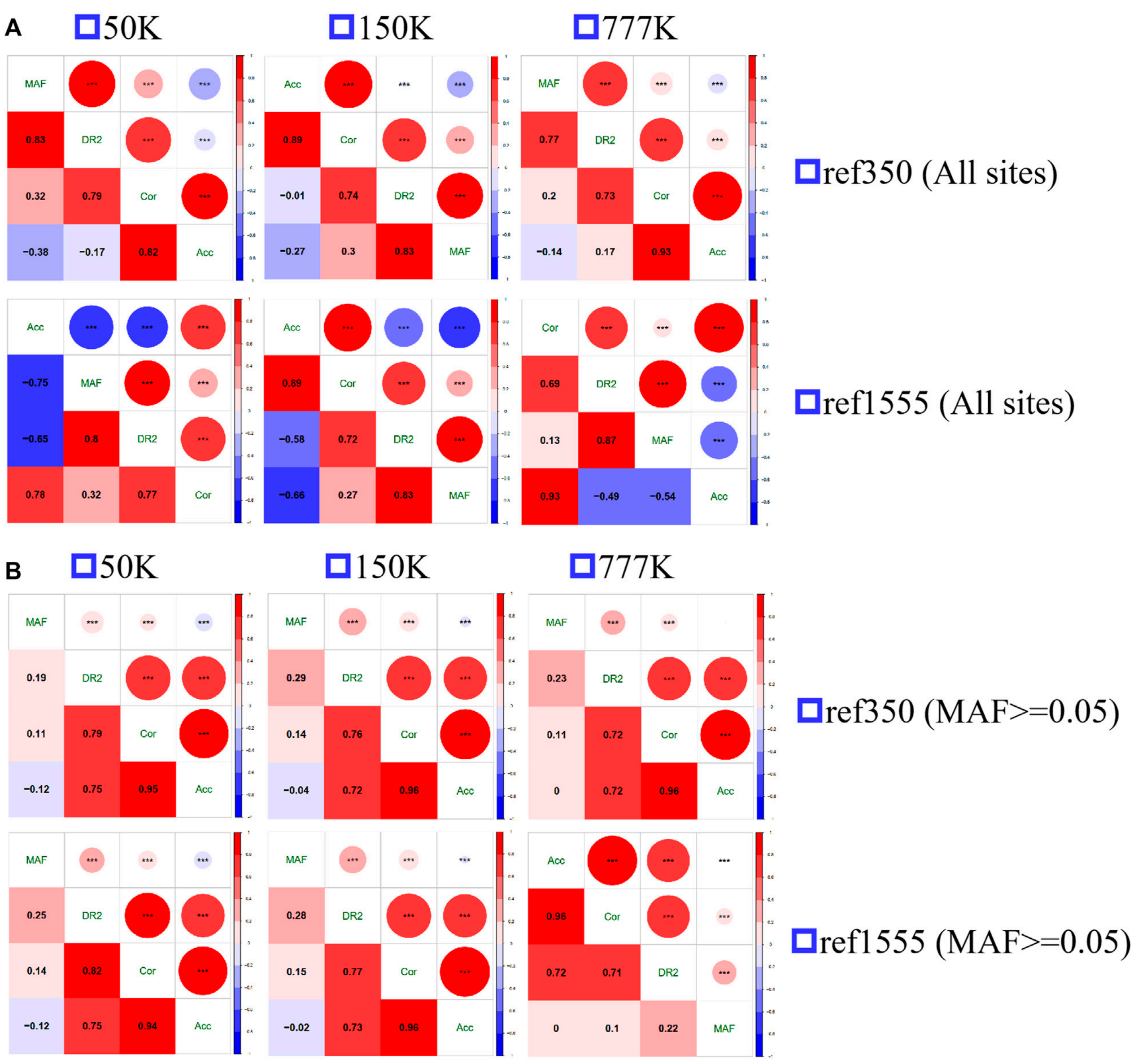


FIGURE 5
The spearman correlation of the three measures of imputation accuracy and minor allele frequency (MAF) among each other. **(A)** All sites **(B)** the sites with minor allele frequency no less than 0.05.

## 3.3 Impact of reference population size and chip density on imputation accuracy

The imputation accuracy when using different chip densities with different reference population sizes is shown in Figure 3. Overall, the large reference populations resulted in higher imputation accuracy. The imputation accuracies measured by genotype concordances (Acc) at 50, 150, and 777 K, using ref350, were 0.925, 0.952, and 0.962, respectively, and their corresponding Cor values were 0.735, 0.784, and 0.812, respectively. The imputation accuracies measured by Acc at 50, 150, and 777 K, using ref1555, were 0.972, 0.981, and 0.985, respectively, and their corresponding Cor values were 0.720, 0.781, and 0.823, respectively. In general, the higher the chip density of the target panel, and the larger the number of reference panels, the higher the imputation accuracy. The Acc at chip densities of 50, 150, and 777 K were improved by 4.73%, 2.88%, and 2.30%, respectively, when the reference population was increased from ref350 to ref1555.

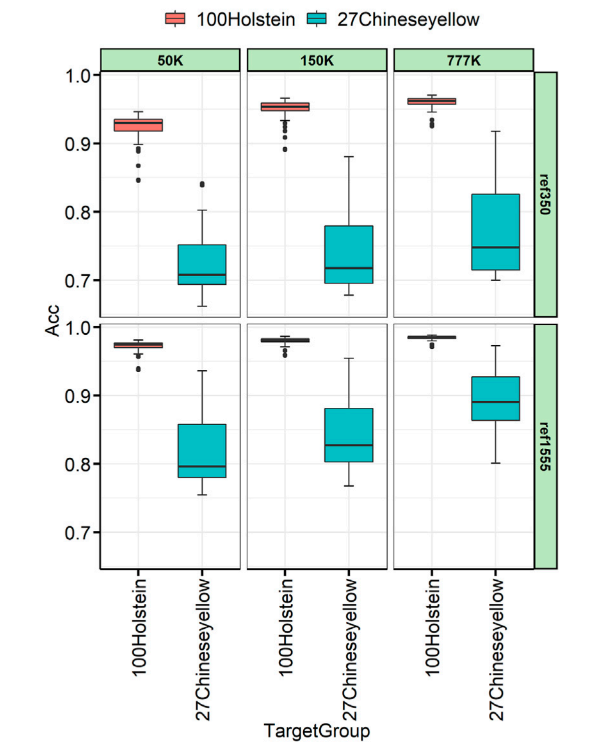


FIGURE 6
Genotype concordance calculated in the individual lever for 100 Holstein and 27 Chinese yellow cattle.

## 3.4 Effect of minor allele frequency on the imputation accuracy

The SNPs were divided into 50 successive bins according to their MAF, with 0.01 step increments. Generally, we found that Acc was slightly decreased when Cor and DR2 were high (Figure 4). Because AR2 is only generated by Beagle 4.1, we also provide an example of the imputation results from 150 K to the two reference panels, which are like the results in Supplementary Figure S3, and AR2 is slightly lower than DR2. As expected, the imputation accuracy increased with an increase in MAF, and the accuracy changed rapidly when the MAF was less than 0.05. In addition, we can see that with a large reference panel, the imputation accuracy of low-MAF sites can be significantly improved. When the reference panel was increased from ref350 to ref1555, the lowest classified MAF site bin (MAF ≤ 0.01) imputation accuracy of Acc for 50, 150, and 777 K chips increased by 4.05%, 1.78%, and 1.16%, respectively; Cor increased by 2.28%, 6.97%, and 8.93%, and DR2 increased by 13.05%, 4.81%, and 4.52%, respectively. In the case of the same reference panel, with the increase in chip density, the imputation accuracy of the low-MAF sites will also be greatly improved. When the chip density was increased from 50 to 777 K, the imputation accuracy of Acc, Cor, and DR2 for ref350 increased by 4.44%, 17.21%, and 33.32%, respectively, and by 1.54%, 23.86%, and 24.80%, respectively, for ref1555.

TABLE 4 The imputation accuracy for 27 Chinese yellow cattle.

| Breed | Individual number | Imputation accuracy |
|---|---|---|
| Menggu | 2 | 0.953 |
| Yanbian | 2 | 0.941 |
| Hasake | 2 | 0.919 |
| Xizang | 1 | 0.888 |
| Qinchuan | 2 | 0.871 |
| Luxi | 2 | 0.869 |
| Guanling | 2 | 0.837 |
| Dengchuan | 2 | 0.832 |
| Wenling | 2 | 0.808 |
| Dehong | 2 | 0.805 |
| Dabieshan | 2 | 0.802 |
| Fujian | 2 | 0.802 |
| Liping | 2 | 0.789 |
| Nanyang | 2 | 0.768 |

Imputation accuracy was measured using genotype concordance (Acc). This imputation was performed from 150 K to WGS with ref1555 using Beagle 5.2 with Ne = 1,00,000.

## 3.5 The correlation between different measures of imputation accuracy

The Spearman correlation between MAF and the three measures of imputation accuracy was calculated and plotted, as shown in Figure 5. All the correlations were significant, with strong positive correlations between Acc and Cor (range from 0.78 to 0.93 with an average of 0.87), Cor and DR2 (range from 0.69 to 0.79, with an average of 0.74) at all loci, and Acc was moderately negatively correlated with MAF (range from −0.38 to −0.14, with an average of −0.26 for ref350, ranging from −0.75 to −0.53, with an average of −0.65 for ref1555), while DR2 had a strong positive correlation with MAF (range from 0.77 to 0.87, with an average of 0.83). Since the inconsistency between Acc and other accuracy measures was mainly in the case of MAF < 0.05 (Figure 4; Supplementary Figure S3), we also calculated the correlation after removing the sites with MAF less than 0.05. Here, we found a strong positive correlation between Acc, Cor, and DR2, with Acc and Cor being 0.96, Acc and DR2 being 0.73, and Cor and DR2 being 0.76. There was a weak negative correlation between Acc and MAF (−0.05), and a weak positive correlation between Cor and MAF (0.12), DR2, and MAF (0.24). Similarly, we also evaluated the correlation of AR2 with other metrics using imputation from 150 K to the two reference panels. As expected, there is a high correlation between AR2 and DR2 (0.98 for all sites and 1 for the sites with MAF greater than or equal to 0.05)

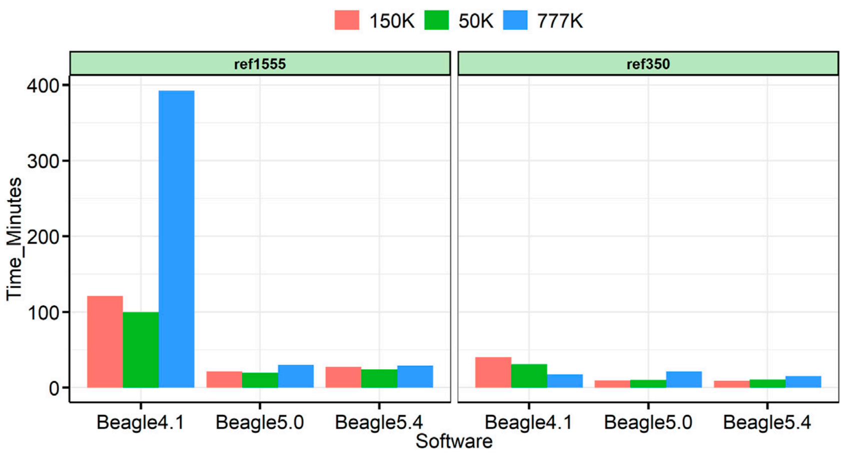


FIGURE 7
Time utilized for each imputation.

(Supplementary Figure S4), which may be the reason why only DR2, and no AR2, output was observed after subsequent Beagle 5.0 version analysis.

## 3.6 The relationship between target and reference individuals on the imputation accuracy

To better understand the relationship between the target and the reference individuals, 27 Chinese yellow cattle were used as the target panel for imputation, which had a complex history between *B. taurus* and *indicus*. The imputation accuracy varies across individuals, as shown in Figure 6. The variance among Chinese yellow cattle was much larger than that among the Holstein target individuals. Taking the imputation from 150 K to WGS, with ref1555, as an example, the imputation accuracies ranged from 0.768 to 0.953, and Mongolian cattle achieved the highest imputation accuracy, followed by Yanbian, Hasake, and Xizang cattle, which were 0.953, 0.941, 0.919, and 0.888, respectively. Nanyang cattle achieved the lowest imputation accuracy, followed by Liping, which was 0.768 and 0.789, respectively, as shown in Table 4. Other breeds such as Qinchuan, Luxi, Guanling, Dengchuan, Wenling, Dabieshan, and Fujian ranged from 0.802 to 0.871.

## 3.7 Running time

All analyses were run on a 22-core 2.10 GHz Linux computer, with Intel(R) Xeon(R) Gold 6,238 processors, and 1,007 GB of memory. Beagle was run on 24 threads. Figure 7 shows the computation time for each panel of cattle. In all cases, Beagle 5 is significantly faster than Beagle 4.1, and Beagle 5.0 is comparable to Beagle 5.4. In many reference panels, the obvious advantages of Beagle 5 can be obtained at 4.6-fold, 5.0-fold, and 13.2-fold, faster than Beagle 4.1 for the imputation of 50, 150, and 777 K, respectively.

# 4 Discussion

Imputation has been widely adopted, in the genomic era, as an important approach to boost the power of genetic studies of animal and human traits. By using the genotypes obtained from the 1000 Bull Genomes Project as the benchmark, with the incorporation of pig datasets from the GVM database, as well as the chicken datasets (Yuan et al., 2018), we systematically assessed the imputation performance of three representative versions of Beagle software with sets of effective population size across the three livestock and poultry. We also identified the influence of several key factors on imputation accuracy, such as chip density, the size of the reference panel, the relationship between the target panel and the reference panel, and the correlation between the measures of accuracy and the MAF. Overall, these key factors must be considered before performing an imputation.

With the continuous update of Beagle versions, various versions of the Beagle software were used in the published research, and the vast majority of studies used the default parameters (Li et al., 2020b; Li et al., 2022). However, in our study, we discovered that it is not suitable to use the default parameter Ne (default Ne = 1,000,000 for Beagle4.1 and Beagle5.0) when the number of reference panels is small or the chip density of the target panel is low, which will drop sharply, with drops ranging from 2.47% to 10.45% under our imputation cases (Figure 3). A similar result was reported by Pook et al. (2020) for the maize population, the imputation error rate increased when using default parameters in Beagle. It is worth noting that the default size of Ne in Beagle's latest version, 5.4, is 100,000; in this case, all three versions of Beagle can obtain high imputation accuracy. This reminds us that it is more appropriate to set Ne to 100,000 to obtain higher accuracy because there is no reference panel as large as cattle, in other livestock and poultry (Supplementary Figures S1, S2). All these results indicate that the default Ne parameters are better changed when using earlier Beagle versions. Furthermore, there is little difference in imputation accuracy among the three versions (only thousandths of the change), but Beagle 5 can significantly sped up the computation, especially with large reference panels (13.24-fold faster in our cases) (Figure 7).

Our results showed that using a large mixed-breed reference population attained a much higher imputation accuracy than using a small single-breed reference population of the same breed as the target population, which is in agreement with the studies of Brøndum et al. (2014), Pausch et al. (2017). The reason for this high imputation accuracy for large mixed reference panels may be the variety of haplotypes in the reference panel, and their ability to facilitate the identification of long-shared haplotypes. With the development of next-generation sequencing technology, sequencing has decreased by five orders of magnitude (Smith, 1993), and the size of data sets used as reference panels for genotype imputation has increased rapidly. Especially for the genome projects implemented, such as the Human Project (Michael, 2005; Adam et al., 2015), the 1000 Bull Genomes Project (Daetwyler et al., 2014), and the dog genome projects (https://www.broadinstitute.org/scientific-community/science/projects/mammals-models/dog/dog-genome-links), which greatly facilitated imputation.

Previous studies have suggested that a low allele frequency may play an important role in complex traits (Manolio et al., 2009). However, it is challenging to correct the imputation of variants at low MAF and rare variants. Similar to previous studies (Teng et al., 2022), we also found that the accuracy dropped sharply for variants with MAF less than 0.05. In agreement with published research (Brøndum et al., 2014; Pausch et al., 2017), a multibreed combined reference panel increased imputation accuracy at low MAF variants. In addition, we found that the increase in chip density and imputation accuracy could also be improved at low MAF variants (Figure 4).

Across studies, there are different measures to evaluate the accuracy of imputation (Pausch et al., 2017; Yan et al., 2017; Song et al., 2019), including the genotype concordance, which counts the proportion of the correctly imputed sites to all imputed sites (Acc) and it is equal to 1 minus imputation error rates (the number of incorrectly imputed sites), the Pearson correlation between true and imputed genotypes (Cor), allelic R-squared (AR2, estimated squared correlation between the most probable REF dose and true REF dose), and Dosage R-Squared (DR2, estimated squared correlation between estimated REF dose and true REF dose) proposed in Beagle. The calculation of Acc and Cor requires the true genotype value, which is generally used to compare imputation methods. AR2 and DR2 are output by Beagle and are proposed as useful measures of imputation accuracy, usually used without knowledge of the true genotype information of the individuals belonging to the target panel. Our results indicated a significantly high correlation between AR2 and DR2 (Supplementary Figure S4), which may explain why only DR2 was the output after the Beagle 5 version. After removing the variants with MAF less than 0.05, a significantly high correlation was observed among the measures of accuracy, as well as a low correlation between MAF. This suggests that one of the metrics may be sufficient to measure the imputation accuracy.

For the imputation from low-density Beadchip to whole genome sequence variants, there are two approaches, one is the one-step imputation, referred direct imputed from low-density chip to WGS, the other is two-step imputation approach, referred imputed from low-density Beadchip to high-density Beadchip at first, and then impute to WGS. Part of the previous studies showed that the two-step imputation suggested to be advantageous in comparison to the one-step imputation approach with regard to imputation accuracy (Binsbergen et al., 2014). However, it had also been shown that the one-step imputation method yields higher imputation accuracy compared to the two-step imputation when fewer animals are available in the intermediate imputation steps (Korkuć et al., 2019). And the two-step imputation is difficult to implement in animals other than cattle since the need for high-density chip populations in large number individuals, and can be affected by the population structure of the high density mediated population. Thus, only one-step imputation was concerned in this study.

## 5 Conclusion

In summary, this study investigated the performance of three representative versions of Beagle (Beagle 4.1, Beagle 5.0, and Beagle 5.4) and the effects of parameter settings on three livestock and poultry (cattle, pig, and chicken) breeds. We found that the default parameter Ne, for the earlier version of Beagle, is not suitable for livestock and poultry in small reference panels or low-density BeadChip chips of target panels. Beagle 5 significantly reduced the computation time without a loss of accuracy, especially for large reference panels. Overall, a large, combined reference panel, or high-density chip, provided greater imputation accuracy, particularly for low minor allele frequency variants. Furthermore, AR2 or DR2 can be used to measure imputation accuracy in the absence of a true genotype. Our findings provide insights into the imputation from BeadChip data to whole-genome sequence variants of livestock and poultry, as well as other non-human species.

## Data availability statement

Publicly available datasets were analyzed in this study. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

## Author contributions

YJ and HS performed all the analyses and wrote the manuscript. XD designed the study and prepared the manuscript. HG made further revisions to the manuscript and provided valuable comments. QZ helped with the design and analysis. All authors have read and approved the final manuscript.

## Funding

This work was supported by the National Key Research and Development Project (2019YFE0106800), China Agriculture Research System of MOF and MARA (CARS-35), the National Natural Science Foundation of China (32070568).

## Acknowledgments

We thank the whole-genome sequence data from RUN5 of the 1000 Bull Genomes Project (http://www.1000bullgenomes.com) for this study. We thank Zhe Zhang and Shaopan Ye for providing chicken data. We thank Zhiyan Zhang for his valuable comments.

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.963654/full#supplementary-material

## References

Adam, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Hyun Min, K., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Binsbergen, R. V., Bink, M. C., Calus, M. P., Eeuwijk, F. A. V., Hayes, B. J., Hulsegge, I., et al. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 46 (1), 41. doi:10.1186/1297-9686-46-41

Brøndum, R. F., Guldbrandtsen, B., Sahana, G., Lund, M. S., and Su, G. (2014). Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *Bmc Genomics* 15 (1), 728. doi:10.1186/1471-2164-15-728

Browning, B., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi:10.1016/j.ajhg.2009.01.005

Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98 (1), 116–126. doi:10.1016/j.ajhg.2015.11.020

Browning, B. L., Tian, X., Zhou, Y., and Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* 108 (10), 1880–1890. doi:10.1016/j.ajhg.2021.08.005

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103 (3), 338–348. doi:10.1016/j.ajhg.2018.07.015

Daetwyler, H. D., Aurélien, C., Hubert, P., Paul, S., Rianne, V. B., Klopp, C., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46 (8), 858–865. doi:10.1038/ng.3034

Frischknecht, M., Pausch, H., Bapst, B., Signer-Hasler, H., Flury, C., Garrick, D., et al. (2017). Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. *Bmc Genomics* 18 (1), 999. doi:10.1186/s12864-017-4390-2

Jenkins, C. A., Consortium, D., Aguirre, G., André, C., and Ricketts, S. L. (2021). Improving the resolution of canine genome-wide association studies using genotype imputation: A study of two breeds. *Anim. Genet.* 52 (2), 703. doi:10.1111/age.13117

Jonathan, M., and Bryan, H. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11 , 499–511. doi:10.1038/nrg2796

Kelemen, L. E., Kate, L., Jonathan, T., Qiyuan, L., Lee, J. M., Ji-Heui, S., et al. (2015). Genome-wide significant risk associations for mucinous ovarian carcinoma. *Nat. Genet.* 47 (8), 888–897. doi:10.1038/ng.3336

Korkuć, P., Arends, D., and Brockmann, G. A. (2019). Finding the optimal imputation strategy for small cattle populations. *Front. Genet.* 10, 52. doi:10.3389/fgene.2019.00052

Li, C., Duan, D., Xue, Y., Han, X., Wang, K., Qiao, R., et al. (2022). An association study on imputed whole-genome resequencing from high- throughput sequencing data for body traits in crossbred pigs. *Anim. Genet.* 53 (2), 212–219. doi:10.1111/age.13170

Li, C., Tian, D., Tang, B., Liu, X., Song, S., Zhao, W., et al. (2020a). Genome variation map: A worldwide collection of genome variations across multiple species. *Nucleic Acids Res.* 49 (1), D1186–D1191. doi:10.1093/nar/gkaa1005

Li, W., Liu, R., Zheng, M., Feng, F., Wen, J., Guo, Y., et al. (2020b). New insights into the associations among feed efficiency, metabolizable efficiency traits and related QTL regions in broiler chickens. *J. Anim. Sci. Biotechnol.* 11 (1), 65. doi:10.1186/s40104-020-00469-8

Liu, Q., Cirulli, E. T., Han, Y., Yao, S., Liu, S., and Zhu, Q. (2015). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Brief. Bioinform.* 16 (4), 549–562. doi:10.1093/bib/bbu035

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461 (7265), 747–753. doi:10.1038/nature08494

Michael, O. (2005). A haplotype map of the human genome. *Physiol. Genomics* 13 (1), 3–9. doi:10.1152/physiolgenomics.00178.2002

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K. U., and Fries, R. (2013). Imputation of high-density genotypes in the Fleckvieh cattle population. *Genet. Sel. Evol.* 45 (1), 3. doi:10.1186/1297-9686-45-3

Pausch, H., Emmerling, R., Schwarzenbacher, H., and Fries, R. (2016). A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genet. Sel. Evol.* 48 (1), 14. doi:10.1186/s12711-016-0190-4

Pausch, H., Macleod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D., et al. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genet. Sel. Evol.* 49 (1), 24. doi:10.1186/s12711-017-0301-x

Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C. C., et al. (2020). Improving imputation quality in BEAGLE for crop and livestock data. *G3 (Bethesda)* 10 (1), 177–188. doi:10.1534/g3.119.400798

Raymond, B., Bouwman, A. C., Schrooten, C., Houwing-Duistermaat, J. J., and Veerkamp, R. F. (2018). Utility of whole-genome sequence data for across-breed genomic prediction. *Genet. Sel. Evol.* 50 (1), 27. doi:10.1186/s12711-018-0396-8

Rowan, T. N., Hoff, J. L., Crum, T. E., Taylor, J. F., and Decker, J. E. (2019). A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genet. Sel. Evol.* 51 (1), 77. doi:10.1186/s12711-019-0519-x

Smith, L. M. (1993). The future of DNA sequencing. *Science* 262 (5133), 530–532. doi:10.1126/science.8211178

Song, H., Ye, S., Jiang, Y., Zhang, Z., and Ding, X. (2019). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. *Genet. Sel. Evol.* 51 (1), 58. doi:10.1186/s12711-019-0500-8

Teng, J., Zhao, C., Wang, D., Chen, Z., Tang, H., Li, J., et al. (2022). Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. *J. Dairy Sci.* 105, 3355. doi:10.3168/jds.2021-21360

van den Berg, S., Vandenplas, J., van Eeuwijk, F. A., Bouwman, A. C., Lopes, M. S., and Veerkamp, R. F. (2019). Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. *Genet. Sel. Evol.* 51 (1), 2. doi:10.1186/s12711-019-0445-y

Vanraden, P. M., Tooker, M. E., O'Connell, J. R., Cole, J. B., and Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* 49 (1), 32. doi:10.1186/s12711-017-0307-4

Ventura, R. V., Miller, S. P., Dodds, K. G., Auvray, B., Lee, M., Bixley, M., et al. (2016). Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genet. Sel. Evol.* 48 (1), 71. doi:10.1186/s12711-016-0244-7

Valero-Mora, P. M. (2010). ggplot2: Elegant graphics for data analysis. *Journal of Statistical Software, Book Reviews* 35 (1), 1–3. doi:10.18637/jss.v035.b01

Yan, G., Qiao, R., Zhang, F., Xin, W., Xiao, S., Huang, T., et al. (2017). Imputation-based whole-genome sequence association study rediscovered the missing QTL for lumbar number in sutai pigs. *Sci. Rep.* 7 (1), 615. doi:10.1038/s41598-017-00729-0

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 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Yang, R., Guo, X., Zhu, D., Tan, C., Hu, X., Ren, J., et al. (2021). Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy. *GigaScience* 10 (7), giab048. doi:10.1093/gigascience/giab048

Ye, S., Gao, N., Zheng, R., Chen, Z., Zhang, Z., Yuan, X., et al. (2019a). Strategies for obtaining and pruning imputed whole-genome sequence data for genomic prediction. *Front. Genet.* 10, 673. doi:10.3389/fgene.2019.00673

Ye, S., Yuan, X., Huang, S., Zhang, H., Chen, Z., Li, J., et al. (2019b). Comparison of genotype imputation strategies using a combined reference panel for chicken population. *Animal* 13 (6), 1119–1126. doi:10.1017/S1751731118002860

Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., et al. (2018). Imputation from SNP chip to sequence: A case study in a Chinese indigenous chicken population. *J. Anim. Sci. Biotechnol.* 9 (2), 30. doi:10.1186/s40104-018-0241-5

Yuan, S., Huang, Zhang, Chen, Z., ZHang, H., Chen, Z., Li, J., et al. (2018). Comparison of genotype imputation strategies using a combined reference panel for chicken population. *Animal* 13, 1119–1126. doi:10.1017/S1751731118002860

Yun, L., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10 (1), 387–406. doi:10.1146/annurev.genom.9.081307.164242

Zhang, C., Kemp, R. A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., et al. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet. Sel. Evol.* 50 (1), 14. doi:10.1186/s12711-018-0387-9

Zheng, H. F., Ladouceur, M., Greenwood, C., and Richard, J. B. (2012). Effect of genome-wide genotyping and reference panels on rare variants imputation. *J. Genet. Genomics* 39, 545–550. doi:10.1016/j.jgg.2012.07.002

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 (4), R42. doi:10.1186/gb-2009-10-4-r42

TYPE Original Research
PUBLISHED 05 September 2022
DOI 10.3389/fgene.2022.940650

Check for updates

OPEN ACCESS

EDITED BY
Yunxia Zhao,
Huazhong Agricultural University, China

REVIEWED BY
Wuttigrai Boonkum,
Khon Kaen University, Thailand
Xiujin Li,
Zhongkai University of Agriculture and Engineering, China
Navid Ghavi Hossein-Zadeh,
University of Guilan, Iran

*CORRESPONDENCE
Peipei Ma,
peipei.ma@sjtu.edu.cn
Yuchun Pan,
panyuchun1963@aliyun.com

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

RECEIVED 10 May 2022
ACCEPTED 04 August 2022
PUBLISHED 05 September 2022

CITATION
Liu D, Xu Z, Zhao W, Wang S, Li T, Zhu K, Liu G, Zhao X, Wang Q, Pan Y and Ma P (2022), Genetic parameters and genome-wide association for milk production traits and somatic cell score in different lactation stages of Shanghai Holstein population.
*Front. Genet.* 13:940650.
doi: 10.3389/fgene.2022.940650

# Genetic parameters and genome-wide association for milk production traits and somatic cell score in different lactation stages of Shanghai Holstein population

Dengying Liu[1], Zhong Xu[2], Wei Zhao[1], Shiyi Wang[1], Tuowu Li[1], Kai Zhu[3], Guanglei Liu[3], Xiaoduo Zhao[3], Qishan Wang[4], Yuchun Pan[4]* and Peipei Ma[1]*

[1]Shanghai Key Laboratory of Veterinary Biotechnology, Department of Animal Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China, [2]Hubei Key Laboratory of Animal Embryo and Molecular Breeding, Institute of Animal Husbandry and Veterinary, Hubei Provincial Academy of Agricultural Sciences, Wuhan, China, [3]Shanghai Dairy Cattle Breeding Centre Co, Ltd, Shanghai, China, [4]Department of Animal Breeding and Reproduction, College of Animal Science, Zhejiang University, Hangzhou, China

The aim of this study was to investigate the genetic parameters and genetic architectures of six milk production traits in the Shanghai Holstein population. The data used to estimate the genetic parameters consisted of 1,968,589 test-day records for 305,031 primiparous cows. Among the cows with phenotypes, 3,016 cows were genotyped with Illumina Bovine SNP50K BeadChip, GeneSeek Bovine 50K BeadChip, GeneSeek Bovine LD BeadChip v4, GeneSeek Bovine 150K BeadChip, or low-depth whole-genome sequencing. A genome-wide association study was performed to identify quantitative trait loci and genes associated with milk production traits in the Shanghai Holstein population using genotypes imputed to whole-genome sequences and both fixed and random model circulating probability unification and a mixed linear model with rMVP software. Estimated heritabilities (h2) varied from 0.04 to 0.14 for somatic cell score (SCS), 0.07 to 0.22 for fat percentage (FP), 0.09 to 0.27 for milk yield (MY), 0.06 to 0.23 for fat yield (FY), 0.09 to 0.26 for protein yield (PY), and 0.07 to 0.35 for protein percentage (PP), respectively. Within lactation, genetic correlations for SCS, FP, MY, FY, PY, and PP at different stages of lactation estimated in random regression model were ranged from -0.02 to 0.99, 0.18 to 0.99, 0.04 to 0.99, 0.04 to 0.99, 0.01 to 0.99, and 0.33 to 0.99, respectively. The genetic correlations were highest between adjacent DIM but decreased as DIM got further apart. Candidate genes included those related to production traits (*DGAT1*, *MGST1*, *PTK2*, and *SCRIB*), disease-related (*LY6K*, *COL22A1*, *TECPR2*, and *PLCB1*), heat stress–related (*ITGA9*, *NDST4*, *TECPR2*, and *HSF1*), and reproduction-related (*7SK* and *DOCK2*) genes. This study has shown that there are differences in the genetic mechanisms of milk production traits at different stages of lactation. Therefore, it is necessary to conduct research on milk production traits at different stages of lactation as different traits. Our results can also provide a theoretical basis for subsequent molecular breeding, especially for the novel genetic loci.

KEYWORDS

**Shanghai Holstein population, milk production traits, genetic parameter, genome-wide association study, different stages of lactation**

## Introduction

Chinese Holstein cattle are derived from grading crossbreeding and selection between the local yellow cattle and Holstein, a breed that was mostly imported from Canada, the United States, France, and northern Europe and renamed by the Chinese Ministry of Agriculture in 1992 (Huang et al., 2010; Ferreri et al., 2011). Since then, China has continuously imported live proven cattle, frozen semen, and embryos from most temperate countries for use in crossbreeding aimed at improving the productivity of Chinese native cattle by combining the environmental adaptation features of Chinese cattle with the high milk yield (MY) potential of foreign cattle (Ferreri et al., 2011; Zhang and Sun, 2021). Therefore, the genetic architecture of the Chinese Holstein population is different from other populations. China occupies a larger area and a larger span of north–south latitudes. Accordingly, topography, climate, herd management system, and other environments vary greatly in different regions, and the different climatic zones have differential contributions to population genetic characteristics, with Holstein in different countries or provinces having its own genetic characteristics (Pérez-Cabal et al., 2012; Liu et al., 2019). The Shanghai Holstein cattle population is raised under a subtropical environment and an intensive pasture system that is maintained below the level of severe thermal stress throughout the day in the summer season. At the same time, Shanghai is the main center for providing Holstein semen to various farms throughout China. Currently, Shanghai Holstein cattle are susceptible to mastitis. The average number of lactations for Shanghai Holstein cattle was 2.23, which makes it difficult to maintain production efficiency and meet the demands of the dairy industry, and the MY is much less than that in the United States (Mao, 2015; Liu et al., 2021).

Since 1994, the Dairy Herd Improvement has been carried out in Shanghai, where millions of test day records are collected (Sun et al., 2008). Milk production and quality, including MY, fat yield (FY), fat percentage (FP), protein yield (PY), protein percentage (PP), and somatic cell score (SCS), are the most important traits in the dairy industry. There are complex traits influenced by management practices and environmental conditions and the physiological stages (e.g., age and stage of lactation) and genetic merits of the animals. Genetic parameters such as heritability are the core of breeding work to accelerate genetic progress and also the most important properties of a population (Meyer, 1989; Akanno and Ibe, 2005). Evaluating genetic parameters is the basis for research such as genome-wide association study (GWAS) and genome-wide selection. However, the heritability of a phenotype in GWAS is too low, resulting in the reduced possibility of detecting the actual association between single nucleotide polymorphisms (SNPs) and traits or non-detection (Shao et al., 2021). Recently, there has been considerable interest in using the random regression model (RRM) to model individual test-day records for the genetic evaluation of milk traits (Khanzadeh et al., 2013; Silva et al., 2020; Soumri et al., 2020).

GWAS is an effective method for identifying the genetic variations involved in complex traits. With the rapid development of high-throughput sequencing technology, many researchers have reported that the power of GWAS based on imputed whole-genome sequencing (WGS) variants on different traits in livestock, such as cattle (Sanchez et al., 2017), pig (Wu et al., 2019; Yan et al., 2021), and chicken (Ye et al., 2020), was improved. Compared to microarray, WGS data cover all SNPs, including causative mutations. However, sequencing thousands of individuals of interest is expensive. Imputation from SNP panels to WGS data is an attractive and less expensive approach to obtain WGS data. Selection of the imputation reference panel is very important for genomic prediction with imputed WGS data. Nowadays, numerous GWASs are conducted on cattle by using the 1000 Bull Genomes Project to impute WGS data on genotyped animals (Iheshiulor et al., 2016; Meuwissen et al., 2021).

Thus far, many researchers have studied the Holstein population in different countries and provinces, including the north of China (Ferreri et al., 2011; Jiang et al., 2012; Liu et al., 2020; Silva et al., 2020). A previous study of the Shanghai Holstein population used the genotyping by genome reducing and sequencing (GGRS) of 1,092 cattle and revealed some SNPs associated with MY, FP, PP, and SCS (Chen Z. et al., 2018), but the study had a small sample size and only conducted association analysis of part of milk production traits using GGRS data. The use of imputed WGS data has been shown that can increase GWAS power and ability to detect causal mutations of complex traits. Therefore, the aim of the present study was to estimate the genetic parameters for milk production and quality traits by using RRM and find new genetic loci by using imputed WGS and a much larger population. In this study, we emphasized the different physiological stages of the mammary gland across the lactation stage. To the best of our knowledge, this is the first time that a GWAS for milk production traits was conducted using imputed WGS data in the south of China, where the Holstein population is suffering heavy heat stress.

TABLE 1 Descriptive statistics of milk production and quality traits in Shanghai Holstein population.

| Traits | No. of records | No. of animals | Mean | Standard deviation | Minimum | Maximum | CV |
|---|---|---|---|---|---|---|---|
| Milk yield (MY, kg/d) | 1,859,464 | 240,681 | 27.80 | 8.35 | 0.1 | 300 | 0.30 |
| Fat yield (FY, kg/d) | 1,855,585 | 240,678 | 0.998 | 0.36 | 0.01 | 7.99 | 0.36 |
| Protein yield (PY, kg/d) | 1,843,598 | 240,680 | 0.866 | 0.25 | 0.003 | 6.944 | 0.29 |
| Fat (FP, %) | 1842807 | 240,679 | 3.64 | 0.88 | 0.02 | 15.90 | 0.24 |
| Protein (PP, %) | 1843717 | 240,681 | 3.15 | 0.38 | 0.1 | 15.90 | 0.12 |
| SCS | 1668583 | 240240 | 2.84 | 1.95 | 0.00 | 9.00 | 0.59 |

# Material and methods

## Data

To evaluate the genetic parameter of milk production traits, we collected the test-day records from the farms of Shanghai Bright Dairy and Food Co., Ltd. from primiparous cows born between 1995 and 2020 with the regular and standard performance of DHI. In total, there are 1,968,589 records for the first lactation of 305,031 cows from 260 farms with the following criteria (Aerts et al., 2021; Mbuthia et al., 2021): 1) age at first calving between 19 and 37 months; 2) test day from 5 to 305 DIM, of which only 12% records out of the range; 3) milk yield of 1.0–65 kg, fat percentage of 0.5–8.5%, protein percentage of 0.5–7.5%, SCC less than 2 million cells per milliliter (Yang et al., 2013); 4) a minimum of three test-day records were required for a cow observation to be included in the analysis (Soumri et al., 2020), of which one was before DIM 45 (Bignardi et al., 2009); 5) the calving date was required to be before December 2019 so that all cows had the opportunity to finish the complete first lactation. A summary of data set used in this analysis is given in Table 1. The somatic cell count (SCC) was log-transformed in SCS as follows: SCS = log2 (SCC/100) + 3; FY was calculated as (FP*MY)/100; PY was calculated as (PP*MY)/100. The distribution of phenotypes is illustrated in Supplementary Figure S1. DMU Trace program was used for tracing ancestors and creating the full pedigree of the animals (Madsen, 2012). The pedigree was built by tracing the ancestors back as far as possible by using the sire-dam structure. Consequently, the pedigrees included 529,011 animals in total, which was recorded during the 1985–2019 period, including 4,945 sires and 19,867 dams, respectively. The inbreeding coefficients for the individuals with test-day records were calculated by going back only three generations in the pedigree. This data set included 226,602 animals. Estimates of the inbreeding coefficient were obtained using the R package “nadiv” (Wolak, 2012).

## Random regression test-day model

The derivative-free approach to multivariate analysis (DMU) package was used to estimate breeding values using the random regression test-day model (RRM) (Jakobsen et al., 2002a; Schaeffer, 2004). Due to problems with convergence, single trait RRM was used to estimate the genetic parameters for different traits. We considered herd-test date, calving month–age, and calving year–season as fixed effects, and individual additive genetic effects and permanent environment effects as random regression effects (Liu et al., 2020). Both random regressions were modeled using fifth-order Legendre polynomial. The model equation is as follows:

$$Y_{ijklmn} = HTD_i + Age_j + CDSD_k + \sum_{m=0}^{5} a_{lm} X_m(\omega) + \sum_{m=0}^{5} p_{lm} X_m(\omega) + e_{ijklmn}$$

Here, $Y_{ijklmn}$ is the test-day records; $HTD_i$ is the fixed effect of the $i$th herd-test day; $Age_j$ is the fixed effect of $j$th calving month–age; $CDSD_k$ is the fixed effect of the $k$th calving year–season; $a_{lm}$ is random regression coefficient for additive genetic effects specific to cow $l$; $p_{lm}$ is random regression coefficient for permanent environment effects specific to cow $l$; $X_m(\omega)$ is the $m$th covariate of Lengendre polynomial; $\omega$ is the days of lactation after standardization; and $e_{ijklmn}$ is the random residual effects.

The variance-covariance matrix is as follows:

$$Var\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{p} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G} \otimes \boldsymbol{A} & 0 & 0 \\ 0 & \boldsymbol{I} \otimes \boldsymbol{P} & 0 \\ 0 & 0 & \boldsymbol{R} \end{bmatrix}$$

Here, $\boldsymbol{a}$ is additive genetic random regression coefficient vector; $\boldsymbol{p}$ is permanent environment random regression coefficient vector; $\boldsymbol{G}$ is the variance–covariance matrix of additive genetic random regression coefficient; $\boldsymbol{A}$ is the numerator relationship matrix; $\boldsymbol{P}$ is the variance–covariance matrix of permanent environment random regression coefficient; $\boldsymbol{I}$ is the identity matrix; and $\boldsymbol{R}$ is the diagonal matrix of residual variance ($\boldsymbol{I}\sigma_e^2$), which hypothesized the residuals are homogeneous. The homogeneous option dramatically reduces computing time without sacrifice as there is a minimal difference between the homogeneous model and the heterogenous model (López-Romero and Carabaño, 2003; Li J. et al., 2020).

## Genotyping, quality control, and imputation

Data from 3,489 genotyped animals were used in this study. In addition, 222 bulls from Run 2 of the 1000 Bull Genome Project were included (Daetwyler et al., 2014). The 3,489 animals were genotyped using different panels: GGP Bovine 50K chip (47,843 SNPs, GeneSeek Genomic Profiler, Neogen Corp., Lincoln, NE, United States, n = 294), GGP Bovine 150 K chip (140,668 SNPs, n = 1,744), GGP Bovine LD v4 (30,108 SNPs, n = 145), Illumina Bovine SNP50K v2 (54,609 SNPs, Illumina, San Diego, CA, United States, n = 1,100) and the extremely low-coverage whole genome sequencing with coverage at 0.5–1× (n = 206).

The extremely low-coverage whole genome sequencing used the Illumina Hiseq4000 platform to sequence the genomic DNA extracted from cow hair-follicle according to the manufacturer's protocol. All of the raw sequence data were filtered using Fastp v0.20.0 (Chen S. et al., 2018) with default parameters and were then aligned to the pig genome build UMD3.1 using BWA mem algorithm implemented in samtools v1.10 (Li and Durbin, 2010). After removing PCR duplicates by Picard Tools v2.0.1 (http://broadinstitute.github.io/picard/), local realignment around indels and base quality scores recalculation were conducted using GATK v3.6 (McKenna et al., 2010) based on known indels and SNPs from in dbSNP database build 152. Sequenced individuals (n = 206) were used to carry out SNP calling *via* both bcftools v1.9 (Li, 2011) (set 1) and GATK UnifiedGenotyper (set 2), simultaneously. The overlapping SNPs between set 1 and set 2 were further filtered *via* GATK VQSR using known variants from the dbSNP database. Finally, a total of 12,396,463 autosomal SNPs with PASS flag and minor allele frequency (MAF) larger than 0.05 were retained. STITCH v1.5.3 (Davies et al., 2016) was used to impute the missing genotypes of the extremely low-coverage whole genome sequencing.

For all the genotype data, only the autosomal chromosomes and SNPs with known positions in the UMD 3.1 bovine assembly map were considered. Genotype quality control for all the panels excluded SNPs with a call rate lower than 0.90, SNPs with deviations from the Hardy–Weinberg equilibrium ($p < 10^{-6}$) as calculated by means of the Fisher's Exact Test, and SNPs with MAF lower than 0.05. For the quality control of the samples, animals with a call rate lower than 0.95 were excluded from the analysis.

The imputation of WGS genotypes from LD and 50K was performed in two steps. First, the LD and 50K genotypes were imputed to 150K, respectively. Then, in the second step, all imputed and real 150K genotypes were imputed to sequence data using 222 bulls from Run 2 of the 1000 Bull Genome Project (Daetwyler et al., 2014) and the UMD3.1 reference sequence. All the abovementioned steps used BEAGLE v4.1 (Browning and Browning, 2009) software. For the imputed extremely low-coverage whole genome sequencing, we used BEAGLE v4.1 to impute to WGS genotypes using 222 bulls as reference sequence described earlier.

All the genotypes imputed to WGS were merged using "bcftools merge--force-samples" (v1.3). We used Perl script to match phenotype samples ID with genotype samples ID to obtain the genotype file which has phenotype. Finally, genotype data were filtered by PLINK v1.9 with the parameters "--geno 0.1 --hwe 0.000001 --maf 0.05 --mind 0.05". Only autosomal SNPs were considered in this study, and IDs without phenotypes were excluded.

## Principal component analysis

To determine the level of population stratification, we plotted the population structure by PCA. Principal component analysis (PCA) was conducted using GCTA v64 (Yang et al., 2011) on 3,016 cows genotyped with 8,686,483 markers covering the whole genome to study the population structure. The first two eigenvectors are selected to make a scatter plot, and according to the results of the scatter plot, it can be known whether the population is divided into several subgroups.

## GWAS analysis

We performed powerful GWAS analyses of six milk production traits (MY, FP, FY, PP, PY, and SCS) in different lactation stages (early lactation [TD7], peak lactation [TD35 and TD50], mid lactation [TD140], and late lactation [TD280]) in the Shanghai Holstein population using FarmCPU (Fixed and random model Circuitous Probability Unification) and MLM (mixed linear model) based on imputed WGS data with the rMVP software (Yin et al., 2021). FarmCPU method is a multi-locus linear mixed model which implements marker tests with associated markers as covariates in a fixed effect model and optimization on the associated covariate markers in a random effect model separately (Liu et al., 2016). As is known, population stratification is an important factor that can cause false positives in association studies. Therefore, the present study fitted the first three principal components (PCs) as covariate variables in the GWAS models to adjust for the population stratification. The model can be written as follows:

$$\boldsymbol{y} = \boldsymbol{Tw} + \boldsymbol{Pq} + \boldsymbol{m_k h_k} + \boldsymbol{e}$$

Here, $\boldsymbol{y}$ is the vector of EBVs of individual; $\boldsymbol{w}$ is a matrix of fixed effect for the top three PCs; $\boldsymbol{q}$ is the pseudo quantitative trait nucleotides (QTNs) effects, which was used as the fixed effects, initiated as an empty set; $\boldsymbol{T}$ and $\boldsymbol{P}$ are the corresponding design matrices for $\boldsymbol{w}$ and $\boldsymbol{q}$, respectively; $\boldsymbol{m_k}$ is the genotype of the $\boldsymbol{k}$ marker; $\boldsymbol{h_k}$ is the corresponding ; and $\boldsymbol{e}$ is the vector of residuals with assuming $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$. The random effect model was used to select the most appropriate pseudo QTNs. The model can be written as follows:

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{e}$$

Here, $\boldsymbol{y}$ is the vector of EBVs of individual; $\boldsymbol{u}$ is the genetic effect of the individual, and $\boldsymbol{u} \sim N(0, 2\boldsymbol{K}\sigma_u^2)$, in which $\boldsymbol{K}$ is the kinship matrix derived from the pseudo QTNs, and $\sigma_u^2$ is an unknown genetic variance; and $\boldsymbol{e}$ is the residual effect vector.

## The MLM can be written as follows:

$$y = \boldsymbol{Wb} + \boldsymbol{Zc} + \boldsymbol{Sa} + \boldsymbol{e}$$

Here $\boldsymbol{y}$ is the vector of EBVs of individual; $\boldsymbol{c}$ is the vector of the same fixed effects as in the FarmCPU model; $\boldsymbol{b}$ is the vector of the SNP substitution effects, and $\boldsymbol{a}$ is the vector of random additive genetic effects with $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma_a^2)$, where $\boldsymbol{G}$ is the genomic relationship matrix, and $\sigma_a^2$ is the additive variance. $\boldsymbol{W}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are the incidence matrices for $\boldsymbol{b}$, $\boldsymbol{a}$, and $\boldsymbol{c}$, respectively.

As suggested by Ji et al. (2019), we used $5 \times 10^{-8}$ and $5 \times 10^{-6}$ as genome-wide and suggestive significance threshold to correct false positive findings due to multiple testing (Ji et al., 2019).

## Enrichment analysis of candidate genes

We extended the positions of significant SNPs 150 Kb upstream and downstream and then updated to the Ensembl (UMD3.1 genome version). Identification of the closest genes to significant SNPs was obtained using Ensembl annotation of the UMD3.1 genome version. GO enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of the candidate genes were performed using the DAVID 6.8 Functional Annotation Tool (https://david.ncifcrf.gov/). In all analyses, the $p$-value < 0.05 was considered significantly different.

Another cost-effective approach to compare, confirm, and locate the most candidate genes related to important traits was to align our results with the QTLdb of UMD3.1, which contains 95,332 QTLs/associations. We identified all the QTLs (<1 Mb) that contained or overlapped with the candidate genes. After matching, the number and function of variants were identified, and these variants were used for subsequent analyses.

# Results

## Phenotypic and estimated genetic parameters

Supplementary Figure S1 shows that all the phenotypes follow normal distributions, which can be used for subsequent genome-wide association analysis. The range of inbreeding coefficient for 226,602 animals is 0–0.42. The number of inbred animals is 1,997. Heritabilities for milk yield, fat yield, protein yield, fat percentage, protein percentage, and SCS estimated with the random regression model for DIM are shown in Figure 1. The heritabilities for all phenotypes, except for PP, were reduced from early lactation, were lowest in the peak lactation stage, and increased gradually, remaining quite constant at the mid and late of the lactation stage. Generally, heritabilities for MY, FY, PY, FP, PP, and SCS ranged from 0.16–0.27, 0.11–0.23, 0.13–0.26, 0.12–0.22, 0.17–0.35, and 0.04–0.14 during the lactation, respectively. As expected, heritabilities for SCS are the lowest in all phenotypes.

Genetic correlations between test-day MY, test-day FY, test-day PY, test-day FP, test-day PP, and test-day SCS at different stages of lactation estimated in RRM are ranged from 0.04 to 0.99, 0.04 to 0.99, 0.01 to 0.99, 0.18 to 0.99, 0.33 to 0.99, and -0.02 to 0.99, respectively (Supplementary Table S1). For the six traits, the highest genetic correlation estimates were observed between adjacent test days and the lowest correlations between more distant test days. The genetic correlations for TD5 and TD7, TD50 and TD65, TD140 and TD95, TD125, TD155, TD185, TD215, TD280 and TD245, TD275, TD305 in all traits were larger than 0.95. For SCS, we obtained negative genetic correlations between TD5 and TD215 and TD245. In this study, we emphasized the different physiological stages of the mammary gland across lactation.

## Imputation and quality control

The imputation accuracy was 0.95, which was evaluated by the internal information score generated by STITCH itself for the extremely low-coverage whole genome sequencing. The imputation accuracy for GGP Bovine LD v4 and GGP Bovine 50K imputed to GGP Bovine 150K was 0.98 and 0.99, respectively. Then the imputation accuracy for imputed GGP Bovine LD v4 and GGP Bovine 50K to WGS (222 bulls from Run 2 of the 1000 Bull Genome Project as reference panel) was 0.97 and 0.97, respectively. The accuracy for GGP Bovine 150K to WGS was 0.97. All the genotypes imputed to WGS were merged, and the final genotype data file contained 19,105,311 SNPs. After the filtration, 8,686,483 loci and 3,016 individuals were retained to be used in the GWAS. Figure 2 displayed the distribution of SNPs across all autosomes.

## Population stratification assessment

The PCA revealed that the Shanghai Holstein population are subdivided into five differentiated groups by the first two principal components, which explained 17.25 and 10.96% of the genetic variability in the analysis, respectively, and about 28.21% of the variation is explained by the first three PCs together (Figure 3).

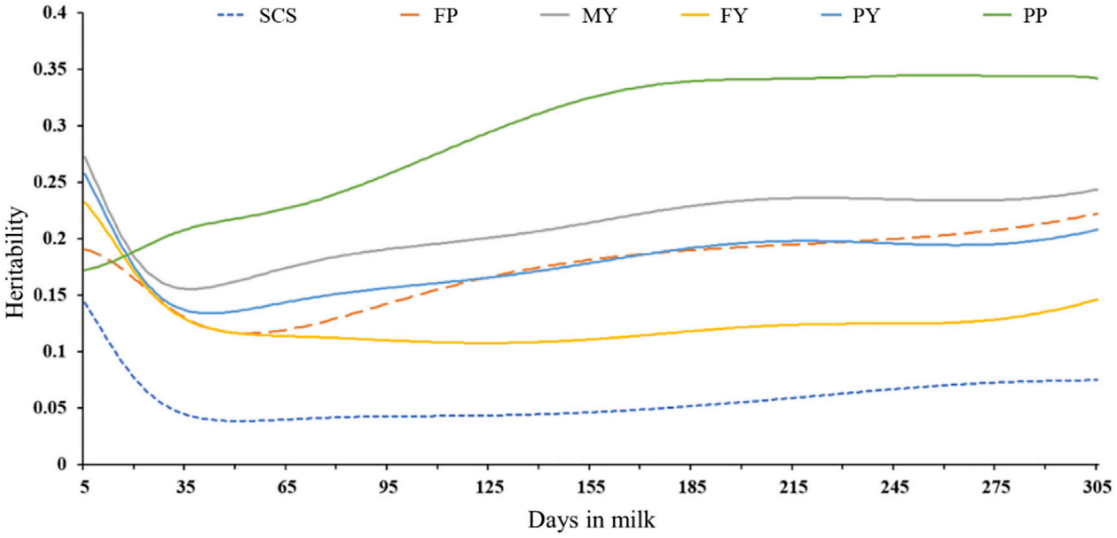


FIGURE 1
Heritabilities for milk yield, fat yield, protein yield, fat percentage, protein percentage, and SCS estimated with the random regression model for DIM.

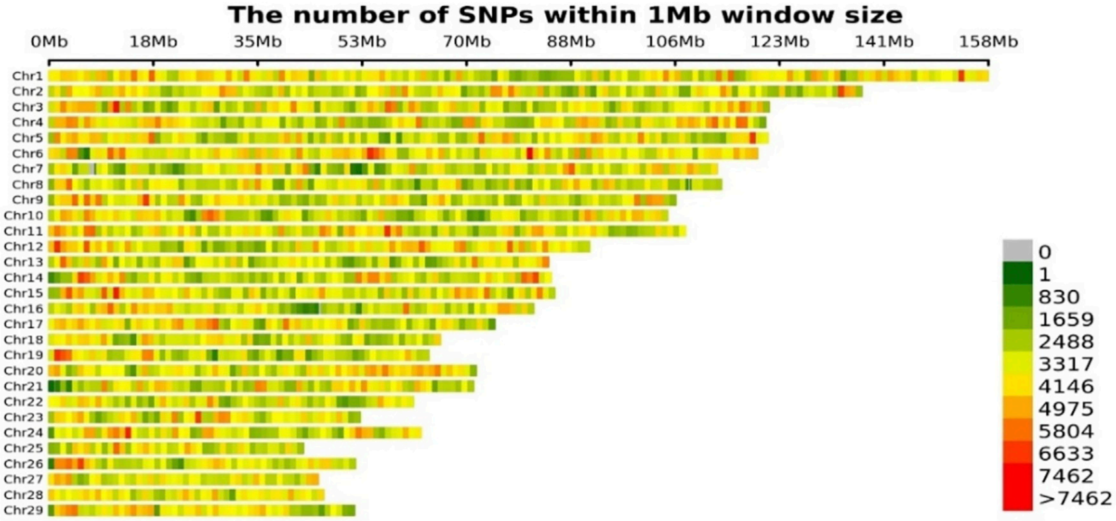


FIGURE 2
Distribution of SNPs in genome.

## GWAS results

Due to the highest genetic correlation estimates being observed between adjacent test days, we only displayed the Manhattan plot of TD7 for all six traits (Figure 4). The Manhattan plot of TD35, TD50, TD140, and TD280 for all six traits are shown in Supplementary Figures S3–S6. The QQ plots are shown in Supplementary Figure S7. The lambda values ranged from 0.921 to 1.042, indicating lower stratification. TD7, TD35, TD50, TD140, and TD280 represented different physiological stages of the mammary gland across lactation (TD7 represented the early, TD50 represented the peak, and TD140 and TD280 represented mid and late lactation, respectively).

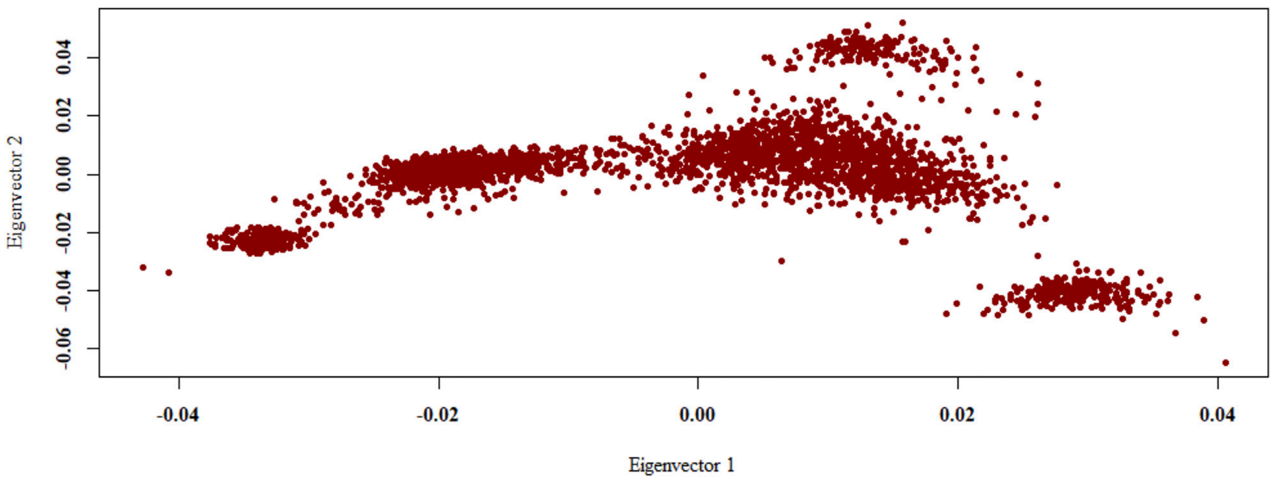


FIGURE 3
Population structure from the principal component analysis. Population structure is shown as a plot of the first two principal components (PCs). PCA was conducted with the 8,686,483 loci for 3,016 cows.

We used $p$-value < $5 \times 10^{-8}$ as the threshold, a total of 984, 1,150, 1,291, 1,229, 1,018, and 1,477 significant SNPs detected by FarmCPU for all DIM of MY, FP, FY, PP, PY, and SCS, respectively. 279, 429, 36, 175, 85, and 42 SNPs were identified as significant by MLM for all DIM of MY, FP, FY, PP, PY, and SCS, respectively (Supplementary Table S2 and S3). There are 44, 30, 20, 16, 26, and 41 significant SNPs are both detected by FarmCPU and MLM for all DIM of MY, FP, FY, PP, PY, and SCS, respectively (Supplementary Table S4). We combined the significant SNPs identified by FarmCPU and MLM. Finally, we obtained a total of 1,241, 1,568, 1,316, 1,399, 1,087, and 1,503 significant SNPs (Supplementary Table S5). These findings are consistent with the results of genetic correlation, a large number of the same SNPs were found in mid and late lactation, while the SNPs found in early and peak lactation were mostly specific. The genes that were located within 150 Kb near the significant SNPs were identified as potential candidate genes for the traits investigated. The number of candidate genes identified is listed in Table 2.

We paid more attention to the candidate genes which contained or were near to the most significant SNPs associated with different milk production in five lactation stages (TD7, TD35, TD50, TD140, and TD280). For MY, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *GRM4*, *VEPH1*, *SCRIB*, *PLBD1,* and *LAMA3*, respectively. For FP, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *ATP2B2*, *NRP1*, *BOP1*, *DGAT1,* and *DGAT1*, respectively. The most significant SNP associated with FP at early lactation was BTA22:55263235 ($p$-value = 2.37E-18). The most significant SNPs for FP at mid and late lactation both were BTA14:1801116 (for TD140: $p$-value = 6.96E-56; for TD280:$p$-value = 7.47E-59). For FY, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *DSP, MAML3, PRKG1, WDR34,* and *SLC1A3,* respectively. For PP, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *DCLK2, AHCTF1, OCLN, MROH1,* and *HSF1*, respectively. The most significant SNP associated with PP at late lactation was BTA14:1807140 ($p$-value = 1.26E-17). For PY, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *CTNND2*, *CSMD3*, *WWOX*, *ARHGAP10,* and *LMAN2L*, respectively. For SCS, the candidate genes contained the most significant SNPs for TD7, TD35, TD50, TD140, and TD280 were *NFKBIE*, *ABCF1*, *MYZAP*, *TTLL7,* and *DNAH9*, respectively.

We further identified the genes which were candidate genes for more than two lactation stages or traits. For MY, there were 18 candidate genes for at least two lactation stages, including *NDST4*, *ICAM2*, *KCNMA1*, *LRP5*, *KALRN*, *IQCA1*, *MANBA*, *SCRIB*, *COL22A1*, *MORN1*, *APBA2*, *ZMYND8*, *WWOX*, *BFAR*, *CECR2*, *GALNT16*, *SPOP,* and *CPEB3*. For FP, a total of 20 candidate genes for at least two lactation stages contained significant SNPs, including *DGAT1, ADAMTS3, ZKSCAN7, CTNNA3, CDH23, ELM O 1, SLC15A5, ESR1, NRP1, BOP1, RPH3A, ATRNL1, FAM21A, MGST1, USH2A, WDR87, SYNRG, RANBP17, ANKRD55,* and *PRIM2*. For FY, 11 genes associated with at least two lactation stages. For PP, 18 genes involved in at least two lactation stages, including *ZMYND8, AHCTF1, TSHR, RALYL, RYR2, ORC2, MAP1S, MT O 1, NRP1, TECPR2, LRP5, NADSYN1, SMC5, KCNQ5, MAP2K6, OCLN, PBX1,* and *PRKG1*. For PY, 15 candidate genes involved in at least two lactation stages, including *WWOX, TMEM132C, NDST4, GUCY1A2,*

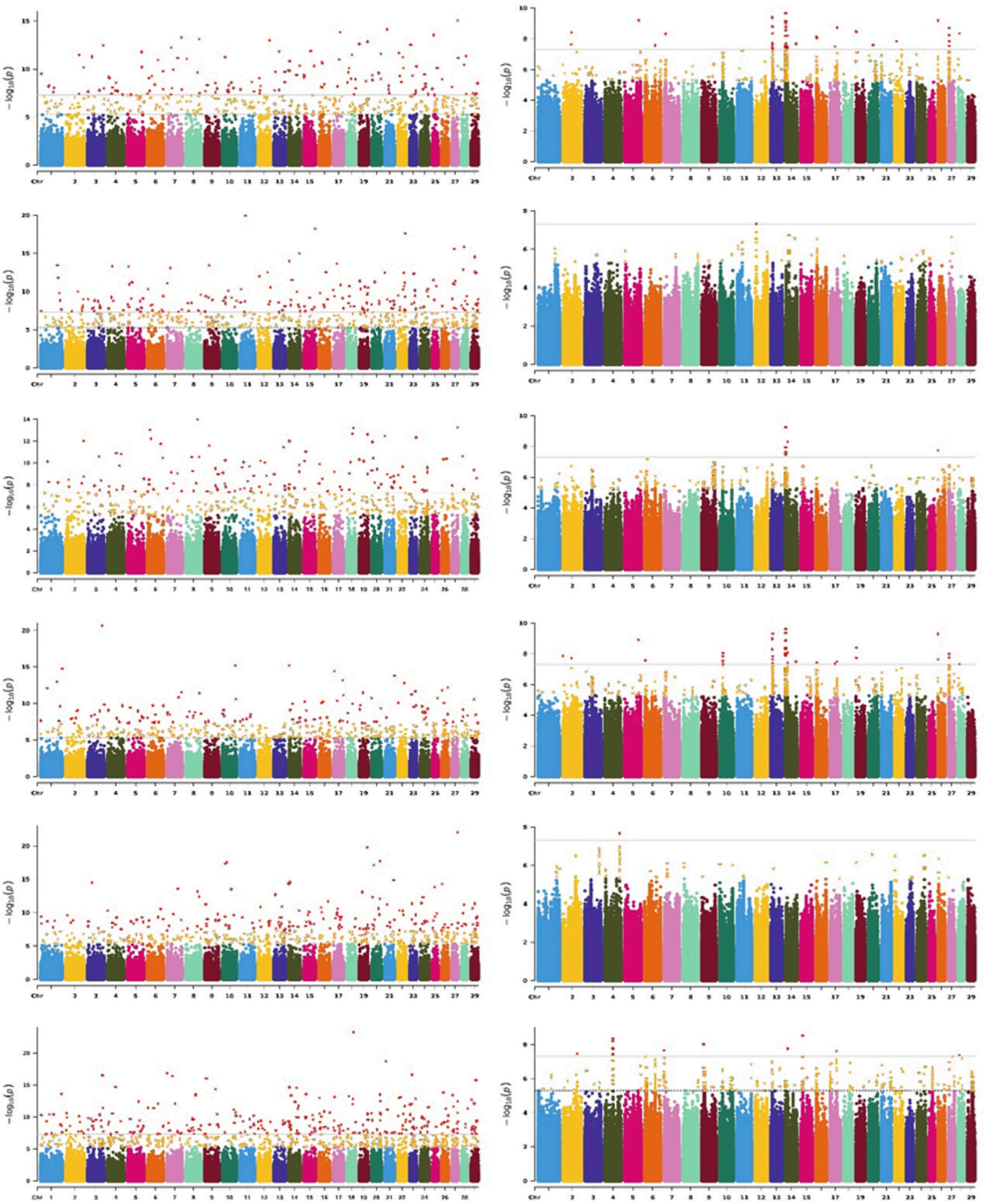


FIGURE 4
Significance [$-\log_{10}$ (*Pvalues*)] of the association of WGS based on analyses using FarmCPU (left) and MLM (right) with the TD7 of six traits, MY, FP, FY, PP, PY, and SCS (top to down) across 29 autosomes. The grey solid line indicates the Bonferroni multiple test threshold at $p = 5 \times 10^{-8}$.

*CTNND2, MANBA, MCC, KCNIP1, ITGA2, CTNNA3, SCRIB, CCDC33, MACROD2, PITPNB,* and *FDXR*. For SCS, 27 candidate genes involved in at least two lactation stages, including *PCDH15, ELM O 1, LDB2, SH3GL2, COL22A1, NUDCD1, HMCN1, CCDC63, GALNS, ADTRP, C1QTNF7, LPAR1, MYZAP, PLCB1, SLC38A9, LANCL2, SLC35F3, DKK2, KCNIP4, TRIM11, RERG, ACOXL, DDX54, DNAH9, ERICH1, MTA1,* and *B3GALNT2*.

TABLE 2 Summary statistics for GO and KEGG associated with milk production in Shanghai Holstein population.

| Traits | DIM | No. Genes | No. GO | No. KEGG |
|---|---|---|---|---|
| Milk yield (MY, kg/d) | D7 | 296 | 7 | 0 |
| | D35 | 339 | 7 | 3 |
| | D50 | 278 | 11 | 1 |
| | D140 | 245 | 5 | 1 |
| | D280 | 179 | 4 | 0 |
| Fat yield (FY, kg/d) | D7 | 232 | 9 | 19 |
| | D35 | 268 | 8 | 1 |
| | D50 | 281 | 17 | 7 |
| | D140 | 336 | 7 | 3 |
| | D280 | 157 | 11 | 2 |
| Protein yield (PY, kg/d) | D7 | 438 | 6 | 5 |
| | D35 | 266 | 5 | 7 |
| | D50 | 280 | 5 | 5 |
| | D140 | 173 | 3 | 2 |
| | D280 | 122 | 6 | 5 |
| Fat (FP, %) | D7 | 450 | 15 | 16 |
| | D35 | 263 | 18 | 8 |
| | D50 | 290 | 7 | 2 |
| | D140 | 308 | 13 | 1 |
| | D280 | 311 | 3 | 0 |
| Protein (PP, %) | D7 | 268 | 11 | 11 |
| | D35 | 208 | 17 | 2 |
| | D50 | 303 | 10 | 2 |
| | D140 | 268 | 11 | 17 |
| | D280 | 325 | 30 | 5 |
| SCS | D7 | 559 | 7 | 6 |
| | D35 | 389 | 5 | 11 |
| | D50 | 373 | 13 | 4 |
| | D140 | 180 | 15 | 1 |
| | D280 | 283 | 3 | 5 |

## Functional annotation of candidate genes

The *p*-value adjusted using the Bonferroni approach (*p*-value < 0.05) was considered to be the threshold value for significantly enriched GO terms and pathways. As shown in Table 2, the number of GO terms and KEGG pathways were significantly enriched for six milk production traits across lactation in Shanghai Holstein. In the current study, gene set enrichment analyses revealed that several terms, such as response to external stimulus (GO:0048870), detection of stimulus (GO: 0051606), negative regulation of response to stimulus (GO: 0048585), and development process were found for almost all traits in almost all lactation stages (Supplementary Table S6–S11). It is interesting that feeding behavior (GO:0007631) was identified for milk yield. For FP, the GO term analysis identified the immune effector process (GO:0002252) and immune response (GO:0003823) in peak lactation. In addition, sexual reproduction (GO:0019953) and reproductive process (GO:0022414) were identified in mid lactation (Supplementary Table S6). For MY, the GO terms were most involved in the biological process and cellular component, such as intracellular (GO:0005622), regulation of signaling (GO: 0023051), and plasma membrane part (GO:0044459) (Supplementary Table S8). For PP, several GO terms related to the development and growth process were identified several in peak and late lactation. 2 GO terms related to reproduction were identified in late lactation (Supplementary Table S9). For SCS, response to chemical (GO:0042221) was identified in peak and mid lactation. In late lactation, the GO terms were related to growth (Supplementary Table S11). The pathways significantly enriched are listed in Supplementary Table S12–S17, of which several pathways were implicated in signal transduction,

TABLE 3 Number of significant SNPs for QTL annotation with different DIM of milk production.

| Traits | DIM | Exterior | Health | Milk | Production | Reproduction |
|---|---|---|---|---|---|---|
| Milk yield (MY, kg/d) | D7 | 23 | 25 | 54 | 60 | 38 |
| | D35 | 27 | 16 | 60 | 71 | 43 |
| | D50 | 28 | 28 | 61 | 71 | 28 |
| | D140 | 49 | 54 | 122 | 116 | 80 |
| | D280 | 166 | 172 | 298 | 244 | 178 |
| Fat yield (FY, kg/d) | D7 | 24 | 18 | 57 | 76 | 39 |
| | D35 | 26 | 28 | 61 | 75 | 37 |
| | D50 | 34 | 23 | 66 | 69 | 48 |
| | D140 | 18 | 32 | 60 | 63 | 40 |
| | D280 | 17 | 12 | 34 | 43 | 28 |
| Protein yield (PY, kg/d) | D7 | 38 | 41 | 86 | 91 | 57 |
| | D35 | 15 | 23 | 66 | 57 | 37 |
| | D50 | 22 | 21 | 54 | 59 | 33 |
| | D140 | 19 | 19 | 35 | 41 | 20 |
| | D280 | 11 | 9 | 23 | 26 | 19 |
| Fat (FP, %) | D7 | 30 | 45 | 80 | 88 | 46 |
| | D35 | 217 | 238 | 385 | 327 | 231 |
| | D50 | 256 | 351 | 534 | 505 | 275 |
| | D140 | 261 | 401 | 590 | 591 | 293 |
| | D280 | 265 | 435 | 649 | 664 | 297 |
| Protein (PP, %) | D7 | 23 | 23 | 70 | 65 | 38 |
| | D35 | 19 | 13 | 38 | 40 | 26 |
| | D50 | 17 | 24 | 53 | 53 | 36 |
| | D140 | 147 | 171 | 316 | 252 | 177 |
| | D280 | 151 | 173 | 333 | 274 | 184 |
| SCS | D7 | 59 | 31 | 115 | 137 | 82 |
| | D35 | 32 | 42 | 87 | 94 | 58 |
| | D50 | 35 | 27 | 85 | 73 | 51 |
| | D140 | 24 | 18 | 57 | 50 | 33 |
| | D280 | 13 | 17 | 39 | 46 | 26 |

including the MAPK signaling pathway (bta04010), Rap1 signaling pathway (bta04015), Ras signaling pathway (bta04014), chemokine signaling pathway (bta04062), Jak-STAT signaling pathway (bta04630), oxytocin signaling pathway (bta04921), and sphingolipid signaling pathway (bta04071); one pathway, olfactory transduction (bta04740), was identified in PP in early lactation and SCS in peak and mid lactation and MY in mid lactation. One pathway was associated with PY, namely, inflammatory mediator regulation of TRP channels (bta04750).

The number and function of variants identified using QTL annotation are listed in Table 3. The significant SNPs associated with MY in late lactation and PP in mid and late lactation were mainly overlapped with milk-related and production-related QTL regions. The SNPs were identified variants and were used for subsequent analyses.

## Discussion

In this research, we estimated various genetic parameters in a large population of Shanghai Holstein that had been regularly measured for six major dairy traits throughout lactation since 1995. This estimation was performed by using a random regression model for the first time in Shanghai. Currently, there are many studies for different Holstein populations (Buaban et al., 2021; Salimiyekta et al., 2021; Fathoni et al., 2022; Sungkhapreecha et al., 2022). We found that the genetic correlation between different test days for milk production was less than one, implying that the different test days had a different additive genetic variance. Oliveira H. et al. (2019) demonstrated that distinct genomic regions affect milk production traits across test days in a whole lactation (Oliveira H. R. et al., 2019). Compared with the genetic correlation estimated in this study, the genetic correlations between TD5 and TD7, TD95 to TD185, and TD245 to TD305 were all extremely high. This means that genetic improvement of one test day of milk production traits could result in a correlated response in the correlated traits. Although there have been many GWAS analyses of milk production traits, elucidating the molecular mechanisms of these traits in other populations can provide new insights into understanding the genetic basis of these traits in dairy cows. Our study subdivided milk production traits during lactation and, more precisely, found significant SNPs that affected different test days.

Currently, there are many studies on the submodels in the random regression test day model. The results of these studies showed that the lactation curves of milk production traits obtained by different researchers were also quite different (El Faro et al., 2008; Zhou and Zhang, 2021; Paiva et al., 2022). Since 1994, with the application of Legendre polynomials in the random regression test day model, research on its order has continued. Li J. et al. (2020) found that for local Chinese Holstein populations, models with third-, fourth-, and fifth-order of Legendre polynomials (LP) led to similar estimates of genetic parameters and predictive ability. Models with higher order obtained lower Akaike information criterion (AIC) and Bayesian information criterion (BIC) values, which was in line with previous studies (Pereira et al., 2013). This means models with LP5 fit data best regardless of complexity. Costa et al. (2008) used fifth-order Legendre polynomial to fit two random effects. Also, RRM based on Legendre polynomials is sensitive to too few records per cow, especially for estimating extreme values of the lactation curve. At the same time, to avoid non-convergence in the RRM due to too few records per cow, we eliminated individuals with fewer than three records when filtering the data.

In our study, except for FP, other traits showed that heritability reached its maximum in early lactation. The heritability of MY varied from 0.16 to 0.27, with the lowest value in peak lactation. In general, the trend for MY heritabilities was like the trend found by Kheirabadi (2019) and Jamrozik and Schaeffer (2012). Kheirabadi (2019) reported that the heritabilities of MY increased with stage of lactation from 0.05 to 0.09 for DIM 5 to 0.24 to 0.25 for DIM 305 for the Iran Holstein population. Jamrozik and Schaeffer (2012) reported that the heritabilities expressed daily were relatively uniform across DIM, except for DIM ranging from 5 to 25. Several studies have reported that the heritabilities of MY in early and late lactation were larger than the value in peak lactation, which is consistent with our results. SCS can reflect the health of the mammary glands, but the low heritability of SCS is an important factor limiting mastitis-resistant breeding. In our research, the heritabilities for DIM ranged from 0.04 (TD51) to 0.14 (TD5). Jamrozik and Schaeffer (2012) found that SCS reached a maximum value in the early lactation, then gradually decreased, and reached a minimum at the peak lactation, then increased steadily and slowly across the lactation. Zakizadeh and Jafari (2014) reported that the heritabilities varied from 0.04 (in early lactation) to 0.136 (in late lactation) for SCS.

We analyzed the genetic correlation between different test days and found that it was highest (close to 1) on adjacent test days but gradually decreased with increasing DIM intervals, which was consistent with previous studies. Jakobsen et al. (2002a) found a genetic correlation between different test days greater than 0.4 (Jakobsen et al., 2002b). Elahi Torshizi et al. (2016) reported that the genetic correlation between different test days varied from 0.47 to 0.98 (Elahi Torshizi, 2016). There was a significant negative genetic correlation between milk production traits in early and late lactation. These negative genetic correlations may be due to difficulties in modeling milk production traits in early lactation when cows are experiencing postpartum stress and lack of energy. Soumri et al. (2020) found the genetic correlation for SCS between test days from -0.11 to 0.99 by using fifth-order Legendre polynomial to fit random effects, which is like the findings in our research (-0.02 to 0.99 across the whole lactation for SCS) (Soumri et al., 2020). The genetic correlation for different DIM is not 1, which means that the additive genetic variance in different DIM is different, which also means that the RRM is used to analyze longitudinal data (e.g., milk production traits). Also, the extremely high genetic correlation between TD95 to TD185 and TD245 to TD305 can explain why the measurement and recording of milk production traits during some test days can be simplified without compromising the reliability of parameter estimates using the RRM.

SNP chips are customized chips based on existing SNP information, and new SNP cannot be found. The coverage of genotyping-by-sequencing accounts for only about 5% of the whole genome and many SNPs are missed. WGS can find SNPs on a genome-wide scale without causing the omission of SNPs (Ye et al., 2018). Compared with SNP chips and GBS, GWAS based on WGS has significant advantages, including that WGS is based on the entire genome to scan and detect SNPs, and the mapping is more accurate (Wu et al., 2019). Therefore, the use of WGS data is expected to improve the detection of QTL, such as the GWAS by using 234 bulls' WGS data in the 1000 Bull Genomes Project (Daetwyler et al., 2014). Although the cost of WGS has decreased, sequencing a large number of individuals for WGS data is still exorbitant. With the development of genotype imputation software, a low-cost method to increase the number of animals with WGS data has been proposed by imputing the lower-density microarray data to the WGS level. Recently, GWAS using imputed WGS data has been widely used in different livestock, such as pigs (Li X. et al., 2020), chickens (Ni et al., 2017; Visscher et al., 2017), cattle (Van Binsbergen et al., 2015; Zhang et al., 2016), and horses (Asadollahpour Nanaei et al., 2020). Especially for cattle, many studies have detected significant important candidate genes by using imputed WGS data in GWAS (Chen N. et al., 2018). In our research, we imputed low- and medium-density SNP chips and GGRS by using a high-coverage WGS-based imputation reference panel (222 bulls from Run 2 of the 1000 Bull Genome project) to WGS data, which is consistent with imputation strategies used in other studies. It has been shown that the use of imputed WGS data in cattle is effective in detecting significant SNPs peaks that were not previously found when using high-density SNP chips in GWAS (Yoshida and Yáñez, 2022). Simultaneously, some authors detected significant SNPs in almost all autosomes by using the imputed WGS data to conduct GWAS on milk production traits, which is in line with our results. In this study, these SNPs identified on different DIM partially overlapped (Sanchez et al., 2017). At the same time, we used a very strict significance threshold (Bonferroni correction treats all variants as independent) that may reduce detection power but minimizes the risk of false positive QTLs.

The genes found in at least two lactation stages or traits and contained or near the most significant SNPs associated with milk production traits were the most important candidate genes in our study. For all six traits studied, there are many common candidate genes detected in TD35 and TD50, such as seven genes among 20 candidate genes for FP, which may be due to the relatively close lactation interval of TD35 and TD50, and the high genetic correlation (greater than 0.9); thus, the mechanisms affecting the traits are similar. *NDST4* is associated with milk fever in the U.S. Holstein cattle (Cavani et al., 2022). In a previous study of milk production traits in Canadian Holstein at different lactation stages, *SCRIB* on BTA14 was a candidate gene for MY and was associated with TD95 to TD215 of PY (Oliveira et al., 2018). Jiang et al. (2010) found that *COL22A1* was an important candidate gene for MY, FP, and PY by conducting GWAS in Chinese Holstein cattle (Jiang et al., 2010). *DGAT1* was detected in the mid and late lactation of FP, which mainly had positive effects on FY and negative effects on MY and PY. Studies have reported that *MGST1* and *SLC15A5* are associated with FY (Jiang et al., 2019). *ADAMTS3* was detected in early, mid, and late lactation, and *ADAMTS3* has been reported to be associated with MY and PY. It is worth noting that *ADAMTS3* is also significantly associated with the longevity of cows (Mészáros et al., 2014). *TECPR2* is related to the heat resistance traits of Chinese cattle, and SNPs located in the gene can be used as molecular markers for Chinese cattle breeding (Ma et al., 2021). Also, *TECPR2* was found to be a candidate gene for SCS in Thai Holstein cattle (Buaban et al., 2022). *PRKG1* plays a key role in lipolysis and is an important candidate gene for fatty acids in milk (Shi et al., 2019, 1). Meanwhile, *PRKG1* was associated with tick resistance in cattle. Our study further supports the importance of this gene in disease resistance traits (Alshawi et al., 2019). *TMEM132C, CTNND2*, and *PCDH15* have been found to be associated with milk production traits (Yodklaew et al., 2017, 2017; Gan et al., 2020). *HMCN1* is known to be associated with age-related macular degeneration, and polymorphisms within the *HMCN1* gene are associated with diabetes in humans (Fisher et al., 2007). This reflects a consistent increase in SCS with age and the progression of lactation, which is consistent with the findings of this study. *DKK2* is involved in adipocyte lipogenesis, which may play a role in fat secretion in milk (Li et al., 2010). *ACOXL* is associated with lipid metabolism and glucose pathways (Klein et al., 2020). *PLCB1* plays multiple biological roles in human diseases, such as inflammation, cell proliferation, and schizophrenia. *DNAH9* affects milk's volatile fatty acid content (Nakamura et al., 2018). *B3GALNT2* was found in a GWAS study of milk production traits in Danish Jersey and Holstein cattle by Poulsen et al. (Buitenhuis et al., 2014). A previous study showed that *ATP2B2* is associated with milk production traits and mastitis (Ogorevc et al., 2009), and the most significant SNP (BTA22:55263235, *p*-value = 2.37E-18) in the GWAS of TD7 FP is located in the intronic region of *ATP2B2*. *PLBD1* is an important candidate gene for fatty acid composition in milk (Atashi et al., 2020). The most significant SNP for PP of TD280 was BTA14:1807140 with a *p*-value = 1.26E-17, which was located on *HSF1*, and *HSF1* plays a crucial role in heat stress response. A previous study found an SNP in the 3′-UTR (g.4693G>T) of *HSF1* that was related to thermo tolerance in Chinese Holstein cattle through association analysis (Li et al., 2011). *NFKBIE* may control the response to several bacterial and viral pathogens and vaccine responses (Lundbo et al., 2016).

Only a few studies have focused on time-dependent genetic associations in livestock to date, but the investigation of the association at certain lactation stages seems to be a promising approach to detect loci associated with milk production (Strucken et al., 2012). Thus, we analyzed how the genetic influence of genomic regions changes during the most critical stages of lactation in our study. We found that the genetic influence on milk production traits varies throughout lactation, which is crucial to enable more efficient genetic selection for these traits and for better management practices, especially for farms or breeders to select high-yielding or milk long-lasting dairy cows. Milk production is related to the stage of lactation, including early lactation, peak lactation, mid lactation, and late lactation. Early lactation is known to be a critical period, especially in high-yielding dairy cows (Deng et al., 2019). Selecting for maximum milk production during lactation early in lactation would improve persistency by lowering the rate of decrease after peak yield (Ferris et al., 1985). Peak milk yield plays a decisive role during the whole lactation period. Zhang et al. (2019) reported that for every 1 kg increase in peak milk production, the yield per primiparous cow increases by about 400 kg. The effect of heat stress on milk yield has been shown to be highest in mid or late lactation. Different genes may be involved in handling different disturbances, explaining the genetic difference among the milk production traits in different lactation stages (Poppe et al., 2021). Candidate genes were only detected at the beginning of lactation showed that the impact on milk production traits must be diminishing in late lactation and suggested that these genes are associated with lactogenesis at the onset of lactation. Candidate genes were detected for all stages of lactation, which could therefore play a role in the immune response of the mammary gland and prevents inflammation during lactation (Strucken et al., 2012). We can use a genomic selection model that combines with markers (significantly associated with different stages of lactation) fit as fixed effects selected from the results of a GWAS (Yin et al., 2020). For example, *MROH1*, an important candidate gene for milk protein composition, is located in a 1.85–2.11 Mb region on BTA14 that has been shown to be associated with 305-days and peak milk production in cows. In addition, the model for selecting is also important. RRM is a feasible alternative to yield more accurate selection and culling decisions. RRM provides information about the temporal variation of biological processes underlying the studied traits to exploit for management and breeding purposes (Oliveira H. et al., 2019).

## Conclusion

In our study, an RRM with fifth-order of Legendre polynomials was an appropriate model for genetic evaluation of six milk production traits in Shanghai Holstein populations. The main results showed that genetic parameters and breeding values were successfully estimated. The results of genetic correlations demonstrated that combining the milk production traits tested on different lactation into a single trait can lead to inaccurate estimates of the genetic value of dairy cows. At the same time, the measurement and recording of milk for some adjacent lactation periods can be simplified without affecting the reliability of parameter estimation using RRM. Then, we detected significant SNPs and candidate genes associated with different traits in different lactation stages, mainly including milk-related genes (*DGAT1, MGST1, PTK2, SCRIB, PRKG1, CTNND2, MROH1, ATP2B2,* and *DNAH9*), disease-related genes (*LY6K, COL22A1, TECPR2, KALRN, CYP7B1, HMCN1,* and *PLCB1*), heat stress–related genes (*ITGA9, NDST4, TECPR2,* and *HSF1*), and reproduction-related genes (*7SK* and *DOCK2*). The genes and QTLs related to heat stress are important to investigate the mechanism of response to heat stress, such as *ITGA9*, which can act as an important gene for heat-resistant breeding of Shanghai Holstein.

## Data availability statement

The data presented in the study are deposited in the Alphaindex repository (http://alphaindex.zju.edu.cn/alphaindex/index.php).

## Ethics statement

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of Shanghai Jiao Tong University.

## Author contributions

Conceptualization: DL and PM; Data curation: DL, SW, and TL; Formal analysis: DL; Methodology: DL, PM, and WZ; Software: DL; Supervision: YP; Writing—original draft: DL; Writing—review & editing: QW, YP, KZ, GL, XZ, and ZX.

## Funding

The project was supported by the National Natural Science Foundation (Grant Nos. 31872321, 31941007, 31972534, and 31701077), the Multidisciplinary Cross Research Foundation of Shanghai Jiao Tong University (Grant No. ZH2018QNA42), and the Fundamental Research Funds for the Central Universities (No. 2662020DKPY005).

## Conflict of interest

Authors KZ, GL and XZ were employed by the company Shanghai Dairy Cattle Breeding Centre Co, Ltd.

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.940650/full#supplementary-material

# References

Aerts, J., Piwczyński, D., Ghiasi, H., Sitkowska, B., Kolenda, M., and Önder, H. (2021). Genetic parameters estimation of milking traits in polish holstein-friesians based on automatic milking system data. *Animals.* 11, 1943. doi:10.3390/ani11071943

Akanno, E., and Ibe, S. (2005). Estimates of genetic parameters for growth traits of domestic rabbits in the humid tropics. *Livest. Res. Rural Dev.* 17, 86.

Alshawi, A., Essa, A., Al-Bayatti, S., and Hanotte, O. (2019). Genome analysis reveals genetic admixture and signature of selection for productivity and environmental traits in Iraqi cattle. *Front. Genet.* 10, 609. doi:10.3389/fgene.2019.00609

Asadollahpour Nanaei, H., Ayatollahi Mehrgardi, A., and Esmailizadeh, A. (2020). Whole-genome sequence analysis reveals candidate genomic footprints and genes associated with reproductive traits in Thoroughbred horse. *Reproduction Domest. Animals* 55, 200–208. doi:10.1111/rda.13608

Atashi, H., Salavati, M., De Koster, J., Ehrlich, J., Crowe, M., Opsomer, G., et al. (2020). Genome-wide association for milk production and lactation curve parameters in Holstein dairy cows. *J. Animal Breed. Genet.* 137, 292–304. doi:10.1111/jbg.12442

Bignardi, A. B., El Faro, L., Cardoso, V. L., Machado, P. F., and de Albuquerque, L. G. (2009). Random regression models to estimate test-day milk yield genetic parameters Holstein cows in Southeastern Brazil. *Livest. Sci.* 123, 1–7. doi:10.1016/j.livsci.2008.09.021

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi:10.1016/j.ajhg.2009.01.005

Buaban, S., Lengnudum, K., Boonkum, W., and Phakdeedindan, P. (2022). Genome-wide association study on milk production and somatic cell score for Thai dairy cattle using weighted single-step approach with random regression test-day model. *J. Dairy Sci.* 105, 468–494. doi:10.3168/jds.2020-19826

Buaban, S., Prempree, S., Sumreddee, P., Duangjinda, M., and Masuda, Y. (2021). Genomic prediction of milk-production traits and somatic cell score using single-step genomic best linear unbiased predictor with random regression test-day model in Thai dairy cattle. *J. Dairy Sci.* 104, 12713–12723. doi:10.3168/jds.2021-20263

Buitenhuis, B., Janss, L. L., Poulsen, N. A., Larsen, L. B., Larsen, M. K., and Sørensen, P. (2014). Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics* 15, 1112. doi:10.1186/1471-2164-15-1112

Cavani, L., Poindexter, M. B., Nelson, C. D., Santos, J. E. P., and Peñagaricano, F. (2022). Gene mapping, gene-set analysis, and genomic prediction of postpartum blood calcium in Holstein cows. *J. Dairy Sci.* 105, 525–534. doi:10.3168/jds.2021-20872

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018a). 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, S., Zhou, Y., Chen, Y., and Gu, J. (2018b). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi:10.1093/bioinformatics/bty560

Chen, Z., Yao, Y., Ma, P., Wang, Q., and Pan, Y. (2018c). Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins. *PLoS One* 13, e0192695. doi:10.1371/journal.pone.0192695

Costa, A., Schwarzenbacher, H., Mészáros, G., Fuerst-Waltl, B., Fuerst, C., Sölkner, J., et al. (2019). On the genomic regions associated with milk lactose in Fleckvieh cattle. *J. Dairy Sci.* 102, 10088–10099. doi:10.3168/jds.2019-16663

Costa, C. N., Melo, C. M. R. de, Packer, I. U., Freitas, A. F. de, Teixeira, N. M., and Cobuci, J. A. (2008). Genetic parameters for test day milk yield of first lactation Holstein cows estimated by random regression using Legendre polynomials. *R. Bras. Zootec.* 37, 602–608. doi:10.1590/s1516-35982008000400003

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46, 858–865. doi:10.1038/ng.3034

Davies, R. W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* 48, 965–969. doi:10.1038/ng.3594

Deng, T., Liang, A., Liang, S., Ma, X., Lu, X., Duan, A., et al. (2019). Integrative analysis of transcriptome and GWAS data to identify the hub genes associated with milk yield trait in buffalo. *Front. Genet.* 10, 36. doi:10.3389/fgene.2019.00036

El Faro, L., Cardoso, V. L., and Albuquerque, L. G. de (2008). Variance component estimates applying random regression models for test-day milk yield in Caracu heifers (*Bos taurus* Artiodactyla, Bovidae). *Genet. Mol. Biol.* 31, 665–673. doi:10.1590/s1415-47572008000400011

Elahi Torshizi, M. (2016). Effects of season and age at first calving on genetic and phenotypic characteristics of lactation curve parameters in Holstein cows. *J. Anim. Sci. Technol.* 58, 8. doi:10.1186/s40781-016-0089-1

Fathoni, A., Boonkum, W., Chankitisakul, V., and Duangjinda, M. (2022). An appropriate genetic approach for improving reproductive traits in crossbred Thai–Holstein cattle under heat stress conditions. *Vet. Sci.* 9, 163. doi:10.3390/vetsci9040163

Ferreri, M., Gao, J., Wang, Z., Chen, L., Su, J., and Han, B. (2011). Chinese Holstein cattle shows a genetic contribution from native asian cattle breeds: A study of shared haplotypes and demographic history. *Asian-Australas. J. Anim. Sci.* 24, 1048–1052. doi:10.5713/ajas.2011.10461

Ferris, T. A., Mao, I. L., and Anderson, C. R. (1985). Selecting for lactation curve and milk yield in dairy cattle. *J. Dairy Sci.* 68, 1438–1448. doi:10.3168/jds.S0022-0302(85)80981-4

Fisher, S. A., Rivera, A., Fritsche, L. G., Keilhauer, C. N., Lichtner, P., Meitinger, T., et al. (2007). Case–control genetic association study of fibulin-6 (FBLN6 or HMCN1) variants in age-related macular degeneration (AMD). *Hum. Mutat.* 28, 406–413. doi:10.1002/humu.20464

Gan, Q., Li, Y., Liu, Q., Lund, M., Su, GuoSheng., and Liang, XueWu. (2020). Genome-wide association studies for the concentrations of insulin, triiodothyronine, and thyroxine in Chinese Holstein cattle. *Trop. Anim. Health Prod.* 52, 1655–1660. doi:10.1007/s11250-019-02170-z

Huang, J., Wang, H., Wang, C., Li, J., Li, Q., Hou, M., et al. (2010). Single nucleotide polymorphisms, haplotypes and combined genotypes of lactoferrin gene and their associations with mastitis in Chinese Holstein cattle. *Mol. Biol. Rep.* 37, 477–483. doi:10.1007/s11033-009-9669-1

Iheshiulor, O., Woolliams, J. A., Yu, X., Wellmann, R., and Meuwissen, T. (2016). Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet. Sel. Evol.* 48, 15. doi:10.1186/s12711-016-0193-1

Jakobsen, J. H., Madsen, P., Jensen, J., Pedersen, J., Christensen, L. G., and Sorensen, D. A. (2002a). Genetic parameters for milk production and persistency for Danish holsteins estimated in random regression models using REML. *J. Dairy Sci.* 85, 1607–1616. doi:10.3168/jds.S0022-0302(02)74231-8

Jakobsen, J. H., Madsen, P., Jensen, J., Pedersen, J., Christensen, L. G., and Sorensen, D. A. (2002b). Genetic parameters for milk production and persistency for Danish holsteins estimated in random regression models using REML. *J. Dairy Sci.* 85, 1607–1616. doi:10.3168/jds.S0022-0302(02)74231-8

Jamrozik, J., and Schaeffer, L. R. (2012). Test-day somatic cell score, fat-to-protein ratio and milk yield as indicator traits for sub-clinical mastitis in dairy cattle: Indicator traits for sub-clinical mastitis. *J. Animal Breed. Genet.* 129, 11–19. doi:10.1111/j.1439-0388.2011.00929.x

Ji, J., Yan, G., Chen, D., Xiao, S., Gao, J., and Zhang, Z. (2019). An association study using imputed whole-genome sequence data identifies novel significant loci for growth-related traits in a Duroc × Erhualian $F_2$ population. *J. Animal Breed. Genet.* 136, 217–228. doi:10.1111/jbg.12389

Jiang, J., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., et al. (2010). Genome wide association studies for milk production traits in Chinese holstein population. *PLoS One* 5, e13661. doi:10.1371/journal.pone.0013661

Jiang, J., Ma, L., Prakapenka, D., VanRaden, P. M., Cole, J. B., and Da, Y. (2019). A large-scale genome-wide association study in U.S. Holstein cattle. *Front. Genet.* 10, 412. doi:10.3389/fgene.2019.00412

Jiang, L., Jiang, J., Wang, J., Ding, X., Liu, J., and Zhang, Q. (2012). Genome-wide identification of copy number variations in Chinese Holstein. *PLOS ONE* 7, e48732. doi:10.1371/journal.pone.0048732

Khanzadeh, H., Hossein-Zadeh, N. G., and Naserani, M. (2013). Estimation of genetic parameters and trends for milk fat and protein percentages in Iranian Holsteins using random regression test day model. *Arch. Anim. Breed.* 56, 487–496. doi:10.7482/0003-9438-56-047

Kheirabadi, K. (2019). Estimates of genetic trends for daily milk yield and somatic cell score of primiparous Holstein cattle in Iran. *J. Appl. Animal Res.* 47, 467–473. doi:10.1080/09712119.2019.1663741

Klein, S.-L., Scheper, C., May, K., and König, S. (2020). Genetic and nongenetic profiling of milk β-hydroxybutyrate and acetone and their associations with ketosis in Holstein cows. *J. Dairy Sci.* 103, 10332–10346. doi:10.3168/jds.2020-18339

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Wang, Z., Moore, S., Schenkel, F., and Stothard, P. (2010). "Genome-wide scan for positional and functional candidate genes affecting milk production traits in Canadian Holstein cattle," in *Proc 9th WCGALP* (Leipzig, Germany: Springer), 26.

Li, J., Gao, H., Madsen, P., Li, R., Liu, W., Bao, P., et al. (2020a). Impact of the order of Legendre polynomials in random regression model on genetic evaluation for milk yield in dairy cattle population. *Front. Genet.* 11, 586155. doi:10.3389/fgene.2020.586155

Li, Q.-L., Ju, Z.-H., Huang, J.-M., Li, J.-B., Li, R.-L., Hou, M.-H., et al. (2011). Two novel SNPs in HSF1 gene are associated with thermal tolerance traits in Chinese Holstein cattle. *DNA Cell Biol.* 30, 247–254. doi:10.1089/dna.2010.1133

Li, X., Ye, J., Han, X., Qiao, R., Li, X., Lv, G., et al. (2020b). Whole-genome sequencing identifies potential candidate genes for reproductive traits in pigs. *Genomics* 112, 199–206. doi:10.1016/j.ygeno.2019.01.014

Liu, D., Chen, Z., Zhang, Z., Sun, H., Ma, P., Zhu, K., et al. (2019). Detection of genome-wide structural variations in the Shanghai Holstein cattle population using next-generation sequencing. *Asian-Australas. J. Anim. Sci.* 32, 320–333. doi:10.5713/ajas.18.0204

Liu, D., Chen, Z., Zhao, W., Guo, L., Sun, H., Zhu, K., et al. (2021). Genome-wide selection signatures detection in Shanghai Holstein cattle population identified genes related to adaption, health and reproduction traits. *BMC Genomics* 22, 747. doi:10.1186/s12864-021-08042-x

Liu, L., Zhou, J., Chen, C. J., Zhang, J., Wen, W., Tian, J., et al. (2020). GWAS-based identification of new loci for milk yield, fat, and protein in Holstein cattle. *Animals.* 10, 2048. doi:10.3390/ani10112048

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12, e1005767. doi:10.1371/journal.pgen.1005767

López-Romero, P., and Carabaño, M. J. (2003). Comparing alternative random regression models to analyse first lactation daily milk yield data in Holstein–Friesian cattle. *Livest. Prod. Sci.* 82, 81–96. doi:10.1016/s0301-6226(03)00003-4

Lundbo, L. F., Harboe, Z. B., Clausen, L. N., Hollegaard, M. V., Sørensen, H. T., Hougaard, D. M., et al. (2016). Genetic variation in NFKBIE is associated with increased risk of pneumococcal meningitis in children. *EBioMedicine* 3, 93–99. doi:10.1016/j.ebiom.2015.11.048

Ma, X., Liu, Y., Sun, L., Hanif, Q., Qu, K., Liu, J., et al. (2021). A novel SNP of TECPR2 gene associated with heat tolerance in Chinese cattle. *Anim. Biotechnol.* 0, 1–8. doi:10.1080/10495398.2021.2011305

Madsen, P. (2012). "DMU trace, a program to trace the pedigree for a subset of animals from a large pedigree file," in *Center for quantitative genetics and genomics, department of molecular biology and genetics, aarhus university* (Aarhus, Denmark: DMU).

Mao, J. (2015). *Genetic analysis between type traits, milk production traits, SCS and longevity traits of Holstein cattle in Shanghai*. Nangjing: Nangjing Agriculture University.

Mbuthia, J. M., Mayer, M., and Reinsch, N. (2021). Modeling heat stress effects on dairy cattle milk production in a tropical environment using test-day records and random regression models. *Animal* 15, 100222. doi:10.1016/j.animal.2021.100222

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

Mészáros, G., Eaglen, S., Waldmann, P., and S, J. (2014). A genome wide association study for longevity in cattle. *Open J. Genet.* 04, 46–55. doi:10.4236/ojgen.2014.41007

Meuwissen, T., van den Berg, I., and Goddard, M. (2021). On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. *Genet. Sel. Evol.* 53, 19. doi:10.1186/s12711-021-00607-4

Meyer, K. (1989). Estimation of genetic parameters. *Rev. Mol. Quantitative Genet. Aproaches Honor Alan Robertson* 1, 159–167.

Nakamura, H., Takahashi-Jitsuki, A., Makihara, H., Asano, T., Kimura, Y., Nakabayashi, J., et al. (2018). Proteome and behavioral alterations in phosphorylation-deficient mutant Collapsin Response Mediator Protein2 knock-in mice. *Neurochem. Int.* 119, 207–217. doi:10.1016/j.neuint.2018.04.009

Ni, G., Cavero, D., Fangmann, A., Erbe, M., and Simianer, H. (2017). Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* 49, 8–14. doi:10.1186/s12711-016-0277-y

Ogorevc, J., Kunej, T., Razpet, A., and Dovc, P. (2009). Database of cattle candidate genes and genetic markers for milk production and mastitis. *Anim. Genet.* 40, 832–851. doi:10.1111/j.1365-2052.2009.01921.x

Oliveira, H., Brito, L., Lourenco, D., Silva, F., Jamrozik, J., Schaeffer, L., et al. (2019a). Invited review: Advances and applications of random regression models: From quantitative genetics to genomics. *J. Dairy Sci.* 102, 7664–7683. doi:10.3168/jds.2019-16265

Oliveira, H. R., Cant, J. P., Brito, L. F., Feitosa, F. L. B., Chud, T. C. S., Fonseca, P. A. S., et al. (2019b). Genome-wide association for milk production traits and somatic cell score in different lactation stages of Ayrshire, Holstein, and Jersey dairy cattle. *J. Dairy Sci.* 102, 8159–8174. doi:10.3168/jds.2019-16451

Oliveira, H., Silva, F., Brito, L., Jamrozik, J., Lourenco, D., and Schenkel, F. (2018). Genome-wide association study for milk, fat and protein yields in different lactation stages in Canadian Holstein and Jersey cattle. *J. Dairy Sci.* 102 (9), 8159–8174. doi:10.3168/jds.2019-16451

Paiva, J. T., Mota, R. R., Lopes, P. S., Hammami, H., Vanderick, S., Oliveira, H. R., et al. (2022). Random regression test-day models to describe milk production and fatty acid traits in first lactation Walloon Holstein cows. *J. Animal Breed. Genet.* 139, 398–413. doi:10.1111/jbg.12673

Pereira, R. J., Bignardi, A. B., El Faro, L., Verneque, R. S., Vercesi Filho, A. E., and Albuquerque, L. G. (2013). Random regression models using Legendre polynomials or linear splines for test-day milk yield of dairy Gyr (*Bos indicus*) cattle. *J. Dairy Sci.* 96, 565–574. doi:10.3168/jds.2011-5051

Pérez-Cabal, M. A., Vazquez, A. I., Gianola, D., Rosa, G. J. M., and Weigel, K. A. (2012). Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front. Genet.* 3, 27. doi:10.3389/fgene.2012.00027

Poppe, M., Bonekamp, G., van Pelt, M. L., and Mulder, H. A. (2021). Genetic analysis of resilience indicators based on milk yield records in different lactations and at different lactation stages. *J. Dairy Sci.* 104, 1967–1981. doi:10.3168/jds.2020-19245

Salimiyekta, Y., Vaez-Torshizi, R., Abbasi, M. A., Emmamjome-Kashan, N., Amin-Afshar, M., Guo, X., et al. (2021). Random regression model for genetic evaluation and early selection in the Iranian Holstein population. *Animals.* 11, 3492. doi:10.3390/ani11123492

Sanchez, M.-P., Govignon-Gion, A., Croiseau, P., Fritz, S., Hozé, C., Miranda, G., et al. (2017). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet. Sel. Evol.* 49, 68. doi:10.1186/s12711-017-0344-z

Schaeffer, L. R. (2004). Application of random regression models in animal breeding. *Livest. Prod. Sci.* 86, 35–45. doi:10.1016/S0301-6226(03)00151-9

Shao, B., Sun, H., Ahmad, M. J., Ghanem, N., Abdel-Shafy, H., Du, C., et al. (2021). Genetic features of reproductive traits in bovine and buffalo: Lessons from bovine to buffalo. *Front. Genet.* 12, 617128. doi:10.3389/fgene.2021.617128

Shi, L., Lv, X., Liu, L., Yang, Y., Ma, Z., Han, B., et al. (2019). A post-GWAS confirming effects of PRKG1 gene on milk fatty acids in a Chinese Holstein dairy population. *BMC Genet.* 20, 53. doi:10.1186/s12863-019-0755-7

Silva, D. A., Costa, C. N., Silva, A. A., Silva, H. T., Lopes, P. S., Silva, F. F., et al. (2020). Autoregressive and random regression test-day models for multiple lactations in genetic evaluation of Brazilian Holstein cattle. *J. Anim. Breed. Genet.* 137, 305–315. doi:10.1111/jbg.12459

Soumri, N., Carabaño, M. J., González-Recio, O., and Bedhiaf-Romdhani, S. (2020). Genetic parameters of somatic cell scores using random regression test-day models with Legendre polynomials in Tunisian dairy cattle. *Livest. Sci.* 241, 104178. doi:10.1016/j.livsci.2020.104178

Strucken, E. M., Bortfeldt, R. H., Tetens, J., Thaller, G., and Brockmann, G. A. (2012). Genetic effects and correlations between production and fertility traits and their dependency on the lactation-stage in Holstein Friesians. *BMC Genet.* 13, 108. doi:10.1186/1471-2156-13-108

Sun, X., Yang, J., and Luo, X. (2008). Correlation analysis of several milk producing traits measured by 305-day milk yield and production performance of Holstein cattle in Beijing. *China Dairy Cattle* 01, 24–26.

Sungkhapreecha, P., Chankitisakul, V., Duangjinda, M., Buaban, S., and Boonkum, W. (2022). Determining heat stress effects of multiple genetic traits in tropical dairy cattle using single-step genomic BLUP. *Vet. Sci.* 9, 66. doi:10.3390/vetsci9020066

Van Binsbergen, R., Calus, M. P., Bink, M. C., van Eeuwijk, F. A., Schrooten, C., and Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 47, 71–13. doi:10.1186/s12711-015-0149-x

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., Mccarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* 101 (1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Wolak, M. E. (2012). Nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods Ecol. Evol.* 3, 792–796. doi:10.1111/j.2041-210x.2012.00213.x

Wu, P., Wang, K., Zhou, J., Chen, D., Tang, G., Yang, X., et al. (2019). GWAS on imputed whole-genome resequencing from genotyping-by-sequencing data for farrowing interval of different parities in pigs. *Front. Genet.* 10, 1012. doi:10.3389/fgene.2019.01012

Yan, G., Liu, X., Xiao, S., Xin, W., Xu, W., Li, Y., et al. (2021). An imputed whole-genome sequence-based GWAS approach pinpoints causal mutations for complex traits in a specific swine population. *Sci. China. Life Sci.* 65, 781–794. doi:10.1007/s11427-020-1960-9

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011

Yang, L., Yang, Q., Yi, M., Pang, Z. H., and Xiong, B. H. (2013). Effects of seasonal change and parity on raw milk composition and related indices in Chinese Holstein cows in northern China. *J. Dairy Sci.* 96, 6863–6869. doi:10.3168/jds.2013-6846

Ye, S., Chen, Z.-T., Zheng, R., Diao, S., Teng, J., Yuan, X., et al. (2020). New insights from imputed whole-genome sequence-based genome-wide association analysis and transcriptome analysis: The genetic mechanisms underlying residual feed intake in chickens. *Front. Genet.* 11, 243. doi:10.3389/fgene.2020.00243

Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., et al. (2018). Imputation from SNP chip to sequence: A case study in a Chinese indigenous chicken population. *J. Anim. Sci. Biotechnol.* 9, 30–12. doi:10.1186/s40104-018-0241-5

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al. (2021). rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteomics Bioinforma.* 19, 619–628. doi:10.1016/j.gpb.2020.10.007

Yin, L., Zhang, H., Zhou, X., Yuan, X., Zhao, S., Li, X., et al. (2020). Kaml: Improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biol.* 21, 146. doi:10.1186/s13059-020-02052-w

Yodklaew, P., Koonawootrittriron, S., Elzo, M. A., Suwanasopee, T., and Laodim, T. (2017). Genome-wide association study for lactation characteristics, milk yield and age at first calving in a Thai multibreed dairy cattle population. *Agric. Nat. Resour.* 51, 223–230. doi:10.1016/j.anres.2017.04.002

Yoshida, G. M., and Yáñez, J. M. (2022). Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. *Evol. Appl.* 15, 537–552. doi:10.1111/eva.13240

Zakizadeh, S., and Jafari, M. (2014). Genetic parameters for somatic cell count and clinical mastitis in the first lactation of Iranian Holstein cattle. *J. Dairy Sci.* 82, 2198–2204. doi:10.3168/jds.S0022-0302(99)75465-2

Zhang, Q., Guldbrandtsen, B., Thomasen, J. R., Lund, M. S., and Sahana, G. (2016). Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. *J. Dairy Sci.* 99, 7289–7298. doi:10.3168/jds.2015-10697

Zhang, S., and Sun, D. (2021). Past, now and future of dairy breeding industry. *China Anim. Ind.* 15, 22–26.

Zhang, S., Yan, T., Yu, S., and Guo, K. (2019). Research on lactation curve of dairy farms in Shanghai. *Feed. Manag.* 3, 43–48.

Zhou, X., and Zhang, J. (2021). Comparison and estimation of different linear and nonlinear lactation curve submodels in random regression analyses on dairy cattle. *Can. J. Anim. Sci.* 101, 567–576. doi:10.1139/cjas-2020-0085

TYPE Original Research
PUBLISHED 07 September 2022
DOI 10.3389/fgene.2022.980463

OPEN ACCESS

EDITED BY
Xiangdong Ding,
China Agricultural University, China

REVIEWED BY
Huifen Xu,
Henan Agricultural University, China
Fengguang Guo,
Texas A&M University, United States
Zhibin Ji,
Shandong Agricultural University, China

*CORRESPONDENCE
Jun Luo,
luojun@nwsuaf.edu.cn
Cong Li,
congl@nwafu.edu.cn

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

RECEIVED 28 June 2022
ACCEPTED 17 August 2022
PUBLISHED 07 September 2022

CITATION
Zhang F, Luo J, Shi C, Zhu L, He Q, Tian H, Wu J, Zhao J and Li C (2022), Genome-wide analysis of the acyl-coenzyme A synthetase family and their association with the formation of goat milk flavour.
*Front. Genet.* 13:980463.
doi: 10.3389/fgene.2022.980463

# Genome-wide analysis of the acyl-coenzyme A synthetase family and their association with the formation of goat milk flavour

Fuhong Zhang, Jun Luo*, Chenbo Shi, Lu Zhu, Qiuya He, Huibin Tian, Jiao Wu, Jianqing Zhao and Cong Li*

Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, China

Goat milk is rich in fat and protein, thus, has high nutritional values and benefits human health. However, goaty flavour is a major concern that interferes with consumer acceptability of goat milk and the 4-alkyl-branched-chain fatty acids (vBCFAs) are the major substances relevant to the goaty flavour in goat milk. Previous research reported that the acyl-coenzyme A synthetases (ACSs) play a key role in the activation of fatty acids, which is a prerequisite for fatty acids entering anabolic and catabolic processes and highly involved in the regulation of vBCFAs metabolism. Although ACS genes have been identified in humans and mice, they have not been systematically characterized in goats. In this research, we performed genome-wide characterization of the ACS genes in goats, identifying that a total of 25 ACS genes (without *ACSM2A*) were obtained in the *Capra hircus* and each ACS protein contained the conserved AMP-binding domain. Phylogenetic analysis showed that out of the 25 genes, 21 belonged to the ACSS, ACSM, ACSL, ACSVL, and ACSBG subfamilies. However, *AACS*, *AASDH*, *ACSF*, and *ACSF3* genes were not classified in the common evolutionary branch and belonged to the ACS superfamily. The genes in the same clade had similar conserved structures, motifs and protein domains. The expression analysis showed that the majority of ACS genes were expressed in multi tissues. The comparative analysis of expression patterns in non-lactation and lactation mammary glands of goat, sheep and cow indicated that *ACSS2* and *ACSF3* genes may participate in the formation mechanisms of goaty flavour in goat milk. In conclusion, current research provides important genomic

**Abbreviations:** AACS, Acetoacetyl-CoA synthetase; AASDH, 2-Aminoadipic 6-semialdehyde dehydrogenase; ACAS_N, Acetyl-coenzyme A synthetase N-terminus; ACSs, Acyl-coenzyme A synthetase genes; ACSBG, Bubblegum acyl-coenzyme A synthetase; ACSF2, Acyl-CoA synthetase family member 2; ACSF3, Acyl-CoA synthetase family member 3; ACSL, Long chain acyl-coenzyme A synthase; ACSM Medium-chain acyl-coenzyme A synthetase; ACSS, Short-chain acyl-coenzyme A synthetase; ACSVL, Very long chain acyl-coenzyme A synthase; AMP-binding_C, AMP-binding enzyme C-terminal; Bt, *Bos taurus*; CDS, Coding sequence; Ch, *Capra hircus*; DHA, Omega-3 docosahexaenoic acid; ER, Endoplasmic reticulum; FAs, Fatty acids; HMM, Hidden Markov Model; Hs, *Homo sapiens*; MEME, Multiple expectation maximization for motif elicitation; MW, Molecular weight; NCBI, National Center for Biotechnology Information; Oa, *Ovis aries*; pI, Theoretical isoelectric point; PP-binding, phosphopantetheine attachment site; RPKM, Reads per kilobase per million mapped reads; UTRs, untranslated regions; vBCFAs, 4-alkyl-branched-chain fatty acids.

resources and expression information for ACSs in goats, which will support further research on investigating the formation mechanisms of the goaty flavour in goat milk.

KEYWORDS

**acyl-coenzyme A synthetase (ACSs), genome-wide, phylogenetic analysis, dairy goat, goaty flavour**

## Introduction

The ACS family comprises a large and diverse group of enzymes. Each member of the ACS family contains a highly conserved amino acid sequence motif, an ATP/AMP binding domain (Black et al., 1997; Mashek et al., 2004). This motif locates at 200–300 amino acids from the N-terminus and is the marker of adenylate-forming enzymes (Watkins et al., 2007; Watkins and Ellis, 2012). In humans, the ACS gene family contains 26 members. Of these members, 22 are subdivided into five subfamilies based on their discrepancies in AMP/ATP and fatty acid-binding motifs (Steinberg et al., 2000). The five subfamilies are the short-chain acyl-CoA synthetase (ACSS), the medium-chain acyl-CoA synthetase (ACSM), the long-chain acyl-CoA synthase (ACSL), the very long-chain synthetase (ACSVL) and the bubblegum ACS synthetase (ACSBG) subfamilies (Grevengoed et al., 2014). Due to their structural features, the acetoacetyl-CoA synthetase (*AACS*), acyl-CoA synthetase family member 2 (*ACSF2*), acyl-CoA synthetase family member 3 (*ACSF3*), and 2-Aminoadipic 6-semialdehyde dehydrogenase (*AASDH*) genes are not classified into any subfamilies and are independent members of the ACS superfamily (Watkins et al., 2007). HUGO nomenclature advisors suggested name these four genes using the interim designation ACSF (ACSF1–4) family (Watkins et al., 2007). In mammals, the ACS family has been characterized in humans and mice, but not in goats.

Different ACS subfamilies exhibit their own preferences for different length of fatty acids (Rossi Sebastiano and Konstantinidou, 2019), with members of each subfamily showing tissue different expression profiles and subcellular locations (Grevengoed et al., 2014). Fatty acids with less than 6 carbons are typically catalyzed by ACSSs, C6—C10 fatty acids are catalyzed by ACSMs, while C12—C20 fatty acids and very long chain fatty acids (>20 carbons) are preferred by ACSLs and ACSVLs, respectively (Grevengoed et al., 2014; Tang et al., 2018). The long chain fatty acids are predominant fats and fulfill essential physiological functions in living organisms (O'Brien et al., 2020). ACSL family has been widely researched in the past (Ansari et al., 2017). ACSL family members, including *ACSL1*, *ACSL3*, *ACSL4*, *ACSL5*, and *ACSL6*, exhibit distinct substrate preferences (Rossi Sebastiano and Konstantinidou, 2019). *ACSL1*, highly expressed in multi tissues such as liver, kidney heart and muscle (Grevengoed et al., 2014), typically prefers oleate and linoleate (Kanter et al., 2012). The absence of *ACSL1* inhibits the sensitivity of macrophages to oleic- and linoleic-mediated degradation of ABCA1 (ATP binding cassette transporter A1), and increase cholesterol spillage (Kanter et al., 2012). *ACSL3* has been found in the endoplasmic reticulum and lipid droplets (Grevengoed et al., 2014), which preferentially catalyzes palmitic and arachidonic fatty acids (Ndiaye et al., 2020). *ACSL4* has also prominent expressed in multi tissues, mainly adrenal gland, brain, ovary, and testis (Grevengoed et al., 2014), and prefers arachidonic acid (Ndiaye et al., 2020). Moreover, previous research suggests the dysregulated expression of both *ACSL3* and *ACSL4* is linked to several diseases, especially cancer (Tang et al., 2018). *ACSL5* has the highest expression in intestinal mucosa relative to other tissues (Meller et al., 2013). The splice variants in *ACSL5* is associated with several types of cancer (Perez-Nunez et al., 2019), and a deletion of *ACSL5* can lead to intestinal lipid malabsorption (O'Brien et al., 2020). The expression of *ACSL6* is large in brain (Takahiro and Tokuo, 1992; Grevengoed et al., 2014), and the absence of *ACLS6* is the most likely cause of the omega-3 docosahexaenoic acid (DHA) deficiency in the brain and spine (Fernandez et al., 2018). Taken altogether, ACSs fulfil distinct roles in fatty acid metabolism.

Goat milk is rich in milk fat and protein, which has high nutritional value and is beneficial to human health (Teng et al., 2018). However, the consumer's acceptance of goat milk and dairy products is restricted because of the perceived characteristic goaty flavour (Kaffarnik et al., 2014). Previous studies have confirmed that 4-alkyl-branched-chain fatty acids (vBCFAs) are the main substances relevant to the goaty flavour in goat milk (Brennand et al., 2010; Salles et al., 2010; Teng et al., 2018). In addition, the concentration of vBCFAs, including 4-ethyloctanoic acid (4-Et-8:0), 4-methyloctanoic acid (4-Me-8:0) and 4-methylnonanoic acid (4-Me-9:0) in goat milk is much higher than in cow milk (Ha and Lindsay, 1993; Kaffarnik et al., 2014). Branched-chain fatty acids (BCFAs) are more likely to be synthesized in goat tissues rather than by rumen microbes (Berthelot et al., 2001). Fatty acids, including vBCFAs, are catalyzed to fatty acyl-CoAs by ACSs before involving in both anabolic and catabolic processes (Steinberg et al., 2000; Watkins et al., 2007; Grevengoed et al., 2014). However, research on the potential role of ACSs in the regulation of vBCFAs metabolism was limited.

Thus, the present study performed a genome-wide characterization of the ACS genes in goat and analyzed their expression profiles in dairy goat, and compared their expression patterns in the non-lactation and lactation mammary glands of dairy goat, sheep, and cow. The major objective of this study was to investigate the function of ACSs in the formation of goaty flavour in goat milk, thus, to provide gene resources for the genetic improvement of goaty flavour by regulating the ACS-mediated vBCFAs metabolism.

# Materials and methods

## Genome-wide identification of acyl-coenzyme A synthetase genes

The goat (*Capra hircus*) reference genome (Accession NO. GCA_001704415.1) was downloaded from the Ensemble database (http://ftp.ensembl.org/pub/release-104/fasta/capra_hircus/dna). The ACS protein sequences of *Homo sapiens* were downloaded from the GenBank (https://www.ncbi.nlm.nih.gov/). To identify ACS genes of goat, we used human ACS protein sequences as queries to carry out local BLASTP ($p = 0.001$) searches against the goat genome database. To further improve the accuracy, we acquired the HMM (hidden Markov model) profile of the AMP-binding domain (PF00501.31) from the Pfam database (http://pfam.xfam.org) (Kochan et al., 2009), then searched the sequence of candidate genes identified by blastp, using HMMER 3.3 (http://eddylab.org/software/hmmer/hmmer-3.3.tar.gz) (Finn et al., 2011). Using the same criterion, ACS genes protein sequences of sheep and cow were obtained from *Ovis aries* (Accession NO, GCA_002742125.1) and *Bos taurus* reference genomes (Accession NO. GCA_002263795.2).

Each potential ACS genes of goat was further analyzed by the programs Pfam to identify the location of domains. The “Compute pI/Mw” tool was used to obtain the pI (theoretical isoelectric point) and MW (molecular weight) of identified ACS proteins (https://web.expasy.org/compute_pi/). Subcellular localizations of identified ACS genes were predicted, using WoLF PSORT website (https://wolfpsort.hgc.jp/). The location of ACS genes on goat genome was mapped using an online website (http://mg2c.iask.in/mg2c_v2.1/). Motif prediction of identified ACS proteins was analyzed using the MEME suite with a maximum number of 15 motifs (http://meme-suite.org/tools/meme). The gene structures and motif sequences were drawn using the EvolView online tool (https://www.evolgenius.info/evolview/#login) (Balakrishnan et al., 2019).

## Multiple sequence alignment and phylogenetic tree construction

The ACS protein sequences from goat, human, sheep, and cow were analyzed together. Multiple sequence alignment was conducted using the MAFFT-7.429 (Katoh and Standley, 2013). The results of multiple sequence alignments were used to predict conserved domains and motifs, and to construct phylogenetic trees. The phylogenetic tree was constructed using the IQ-TREE with a ultrafast bootstrap value of 1,000 (Lam-Tung et al., 2015). The phylogenetic tree and locations of conserved domains of ACSs were drawn using the EvolView online tool.

## Expression analysis of acyl-coenzyme A synthetase genes

The expression levels of the ACS genes of dairy goat at non-lactation and lactation stages were analyzed based on the transcriptome data that were obtained by our laboratory (NCBI SRA accession: PRJNA637690) (Li et al., 2020). To understand the expression profiles of distinct tissues in dairy goat, transcriptome datasets of three organs (heart, kidney, and liver) and skeletal muscle tissue were downloaded, with 3 biological replicates (NCBI SRA accession: PRJNA309284 and NCBI SRA accession: PRJNA309345) from SRA database (https://www.ncbi.nlm.nih.gov/sra/). To further investigate the distinction in transcriptional responses of ACSs in the mammary glands of goat, sheep and cow, the sheep (NCBI SRA accession: PRJNA309284) and cow (NCBI SRA accession: PRJNA482783) transcriptome datasets of mammary gland in the non-lactation and lactation period were downloaded from SRA database (https://www.ncbi.nlm.nih.gov/sra/). The analysis procedures were described as bellow. The transcriptome datasets were converted to fastq format using fastq-dump. All clean reads were mapped to the respective reference genome sequence using Hisat2-2.1.0 (https://github.com/infphilo/hisat2/), and the transcription-level expression was calculated using StringTie (https://ccb.jhu.edu/software/stringtie/). The FPKM values were $\log_2$ transformed, and the heat map of gene expression levels was plotted using the EvolView online tool.

# Results

## Characterization of acyl-coenzyme A synthetase genes

To identify ACS genes in goat, ACS protein sequences of human were used as queries to search the goat genome. A total of 25 ACS genes (*ChACSs*) were identified after an analysis of conserved domains in goat. The full-length ChACS protein sequences were given in the Supplementary Sheet S1. We found that the *ACSM2A* gene was absent in goat reference genome, nor in sheep and cows’ reference genome. The *ChACSs* shared homology to human ACSs (*HsACSs*), with the amino acid identity of 71.2%–96.4% (Supplementary Sheet S2).

TABLE 1 Information of ACS family genes in goat.

| Subfamily | Gene | CDS (bp) | Exon | Intron | Amino acid (aa) | MW (Da) | pI | Localization |
|---|---|---|---|---|---|---|---|---|
| ACSBG | *ChACSBG1* | 2,298 | 17 | 16 | 766 | 85,738.59 | 6.54 | Mitochondrion |
| | *ChACSBG2* | 1866 | 14 | 13 | 622 | 69,142.06 | 7.87 | Plasma membrane |
| ACSL | *ChACSL1* | 2,100 | 21 | 20 | 700 | 78,128.85 | 8.00 | Cytoplasm |
| | *ChACSL3* | 2,163 | 15 | 14 | 721 | 80,127.67 | 8.65 | Cytoplasm |
| | *ChACSL4* | 2013 | 16 | 15 | 671 | 74,477.15 | 8.41 | Mitochondrion |
| | *ChACSL5* | 2052 | 21 | 20 | 684 | 75,532.52 | 7.50 | Endoplasmic reticulum |
| | *ChACSL6* | 2,169 | 21 | 20 | 723 | 80,846.28 | 6.14 | Endoplasmic reticulum |
| ACSM | *ChACSM1* | 1734 | 14 | 13 | 578 | 64,819.43 | 7.84 | Mitochondrion |
| | *ChACSM2B* | 1734 | 15 | 14 | 578 | 64,636.62 | 7.21 | Mitochondrion |
| | *ChACSM3* | 1743 | 14 | 13 | 581 | 65,712.44 | 8.86 | Mitochondrion |
| | *ChACSM4* | 1743 | 13 | 12 | 581 | 64,965.59 | 8.64 | Mitochondrion |
| | *ChACSM5* | 1773 | 14 | 13 | 591 | 66,162.68 | 7.56 | Mitochondrion |
| ACSS | *ChACSS1* | 2028 | 14 | 13 | 676 | 74,352.33 | 6.59 | Mitochondrion |
| | *ChACSS2* | 2,145 | 19 | 18 | 715 | 80,245.85 | 6.34 | Cytoplasm |
| | *ChACSS3* | 2061 | 16 | 15 | 687 | 74,761.3 | 8.86 | Mitochondrion |
| ACSVL | *ChSLC27A1* | 1941 | 13 | 12 | 647 | 71,003.95 | 8.83 | Peroxisome |
| | *ChSLC27A2* | 1863 | 10 | 9 | 621 | 70,319.22 | 8.72 | Endoplasmic reticulum |
| | *ChSLC27A3* | 2,211 | 11 | 10 | 737 | 79,697.36 | 8.73 | Plasma membrane |
| | *ChSLC27A4* | 1932 | 13 | 12 | 644 | 72,242.34 | 8.92 | Peroxisome |
| | *ChSLC27A5* | 2073 | 10 | 9 | 691 | 75,683.17 | 8.70 | Plasma membrane |
| | *ChSLC27A6* | 1905 | 10 | 9 | 635 | 71,710.21 | 8.71 | Endoplasmic reticulum |

TABLE 2 Information of the other ACS genes.

| Gene | CDS (bp) | Exon | Intron | Amino acid (aa) | MW (Da) | pI | Localization |
|---|---|---|---|---|---|---|---|
| *ChAACS* | 2019 | 18 | 17 | 673 | 74,924.77 | 6.14 | cytoplasm |
| *ChACSF2* | 1848 | 16 | 15 | 616 | 68,226.71 | 8.13 | mitochondrion |
| *ChACSF3* | 1761 | 10 | 9 | 587 | 65,045.76 | 6.75 | mitochondrion |
| *ChAASDH* | 3,321 | 15 | 14 | 1,107 | 123,626.02 | 6.24 | plasma membrane |

Among 25 identified genes, 22 were divided into five groups based on the sequence similarity and the principle of the human ACS family nomenclature. There were three genes in ACSS subfamily, five genes in the ACSM subfamily, five genes in the ACSL subfamily, six genes in the ACSVL subfamily, and two genes in the ACSBG subfamily (Table 1). However, *ChAACS*, *ChACSF2*, *ChACSF3*, and *ChAASDH* were not classified in any subfamilies and they were independent members of the ACS superfamily (Table 2).

Subcellular localizations of *ChACSs* were predicted using wolfpsort software. The results revealed that the *ChACSMs*, *ChACSF2*, *ChACSF3*, *ChACSBG1*, *ChACSL4*, *ChACSS1*, and *ChACSS3* genes were located in the mitochondrion, the *ChAACS*, *ChACSL1*, *ChACSL3*, and *ChACSS2* genes in the cytoplasm, the *ChACSL5*, *ChACSL6*, *ChSLC27A2*, and *ChSLC27A6* genes in endoplasmic reticulum, the *ChAASDH*, *ChACSBG2*, *ChSLC27A3*, and *ChSLC27A5* genes in plasma membrance and others in peroxisome. The lengths, molecular weight and theoretical isoelectric point of ACS proteins exhibited substantial variation. The lengths of 26 ACS proteins ranged from 578 to 1,107 amino acids, the MW varied from 64,636.62 to 123,626.02 Da and the pI value changed from 6.14 to 8.92 (Tables 1, 2).

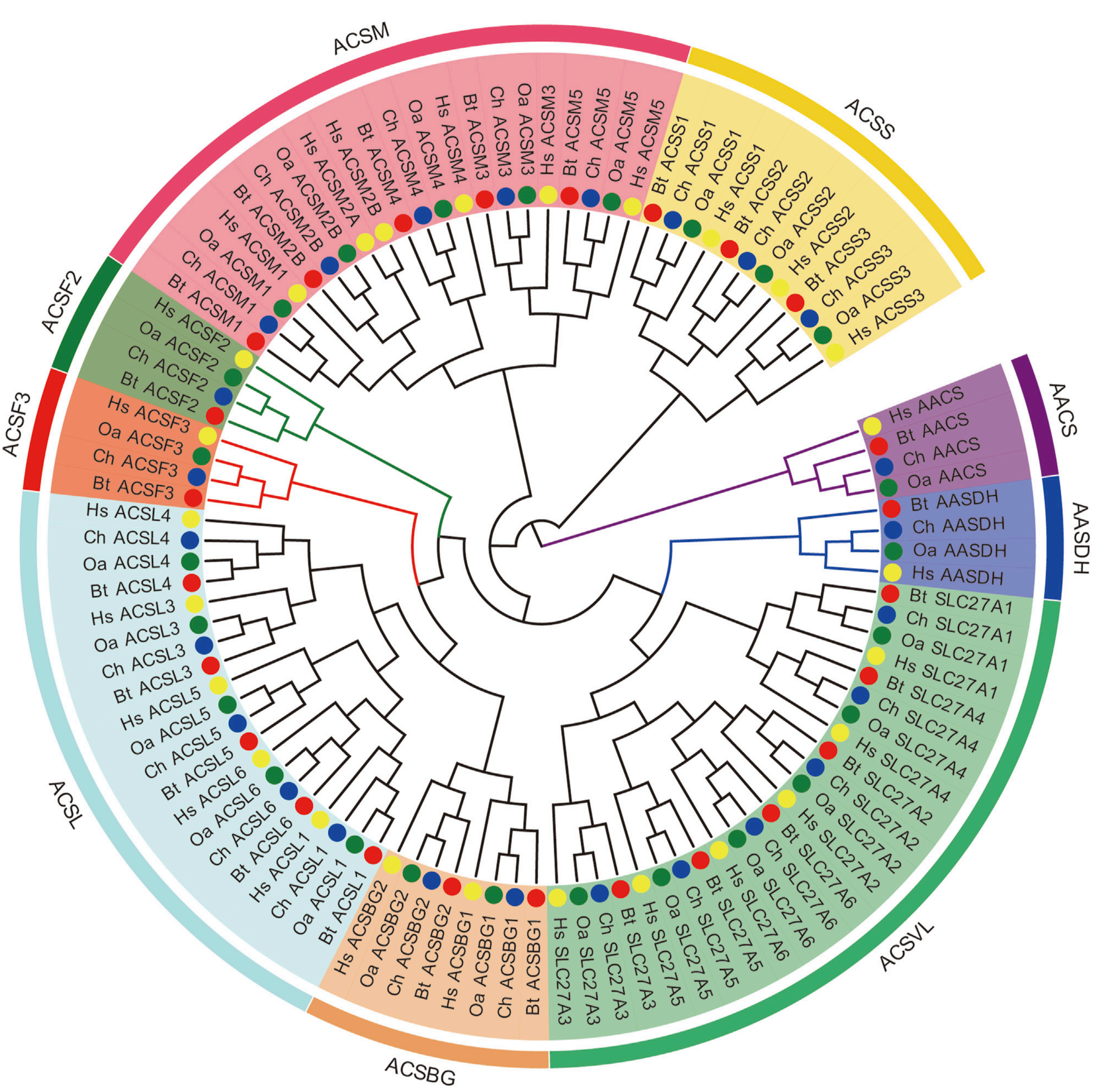


FIGURE 1
Phylogenetic tree of ACS proteins from *Capra hircus*, *Homo sapiens*, *Ovis aries*, and *Bos taurus*. Legends: The tree was generated using the IQ-TREE with an ultrafast bootstrap value of 1,000. Each ACS subfamiliy is marked by a different colour. Note: Ch, *Capra hircus*; *HS, Homo sapiens*; Oa, *Ovis aries*; Bt, *Bos taurus*.

## Phylogenetic analysis and multiple alignments

In order to investigate the evolutionary relationships of the ACS proteins among human, goat, sheep, and cow, a phylogenetic tree was constructed based on their full-length protein sequences, using IQ-TREE software. According to the topological structure of the phylogenetic tree, all the ACS proteins are clustered into nine distinct clades, including the ACSS, ACSM, ACSL, ACSVL, and ACSBG families (Figure 1). The other four clades (ACSS, ACSF1, ACSF2, and AASDH) were not classified in the common evolutionary branch, and belonged to the greater ACS family (Watkins et al., 2007). Hereafter, we characterized these four genes using the ACSF group for facilitating description. Three species (goat, sheep, and cow) had the same amount of

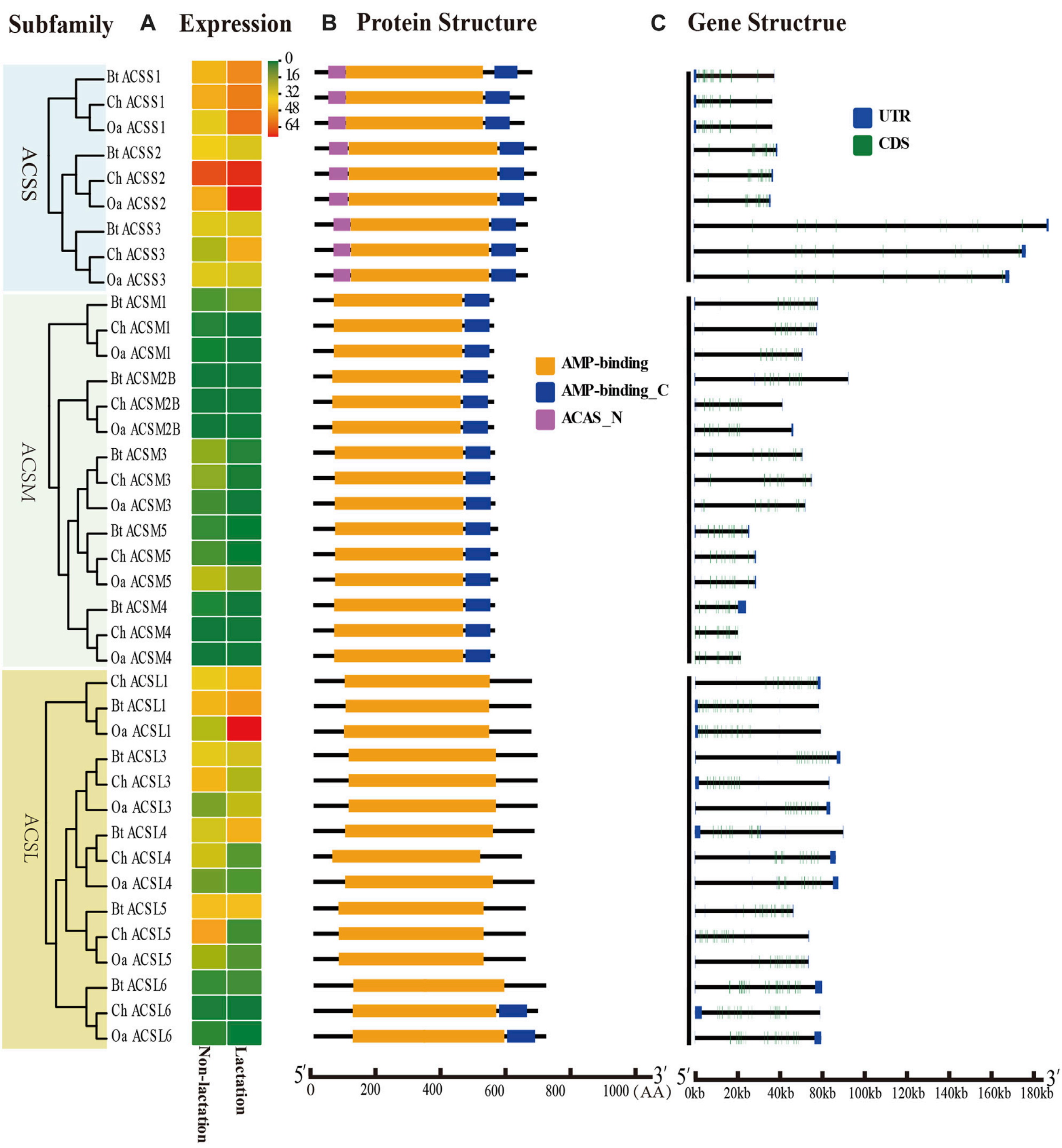


FIGURE 2
Expression profiles, gene structure and protein structure of ACS genes from the ACSS, ACSM, and ACSL subfamilies. Legends: **(A)** Heatmap showing the expression profiles of ACS genes in the non-lactation and lactation period of three species (dairy goat, sheep, and cow). Color gradient from green-to-red indicates expression values change from low to high. **(B)** Structures of ACS proteins with the AMP-binding domain represented by orange boxes, the AMP_C domain in red and ACAS_N domain in purple boxes. **(C)** Structure of ACS genes with exons in green, UTR regions in blue, and solid lines between the colored boxes representing introns.

ACS proteins and the same gene name, except that the *ACSM2A* protein in the ACSM group. The ACS proteins of goat and sheep showed closer evolutionary relationships among three species. The ACSVL group was the largest branch of the ChACS phylogeny and contained six ChACS proteins.

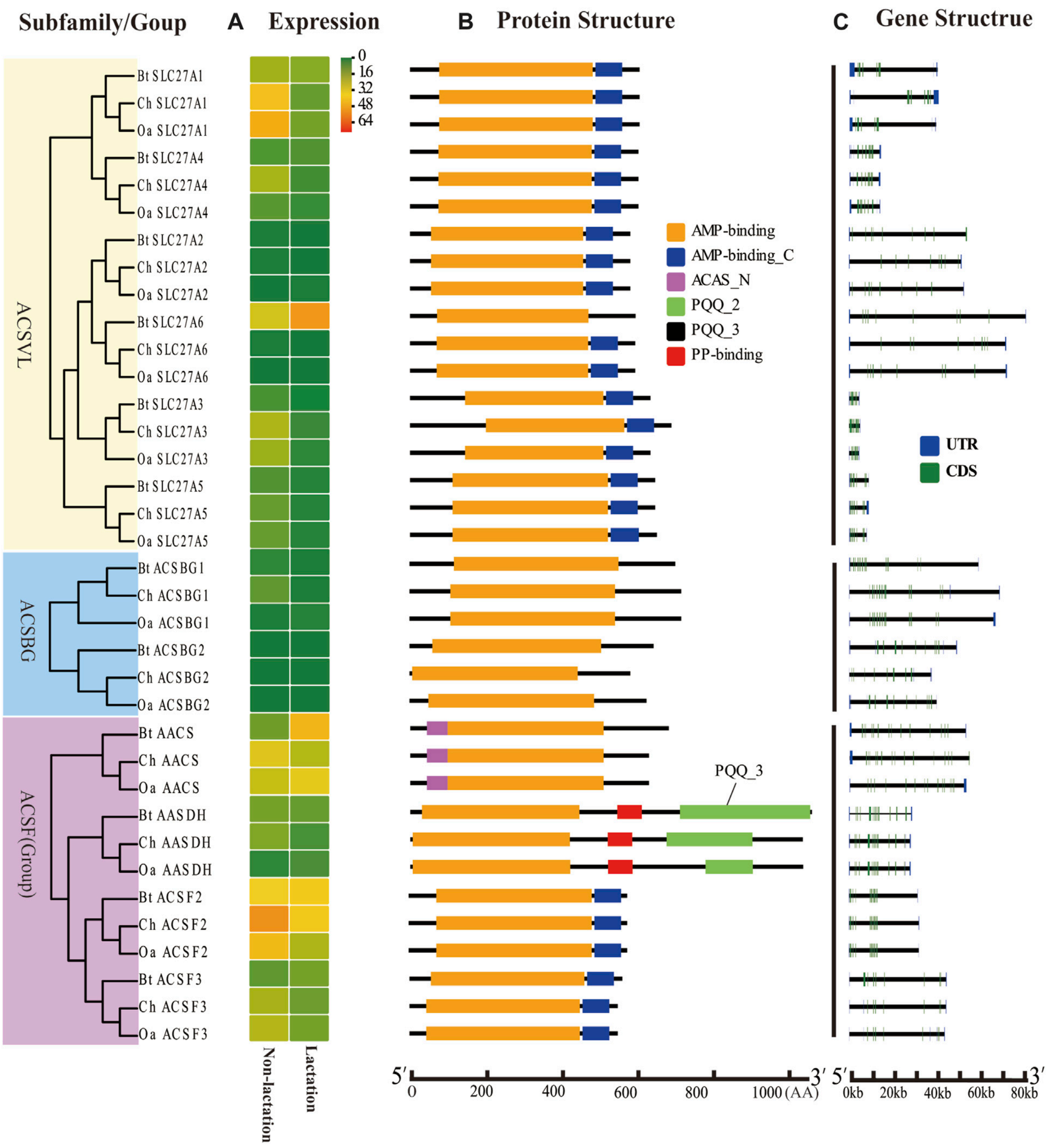


FIGURE 3
Expression profiles, gene structure, and protein structure of ACS genes from the ACSVL and ACSBG subfamilies and ACSF group. Legends: **(A)** Heatmap showing the expression profiles of ACS genes in the non-lactation and lactation period of three species (dairy goat, sheep, and cow). Color gradient from green-to-red indicates expression values change from low to high. **(B)** Structures of ACS proteins with the AMP-binding domain represented by orange boxes, the AMP_C domain in red, ACAS_N domain in purple, PQQ_2 domain in green, PQQ_3 domain in black and PP-binding domain in red. **(C)** Structure of ACS genes with exons in green, UTR regions in blue, and solid lines between the colored boxes representing introns.

To further characterize the ACS genes in goat, the protein sequences were aligned using the MAFFT-7.429. The conserved sequences of goat, sheep and cow ACS proteins were predicted by Pfam, and all proteins were found to contain AMP-binding domains (Figures 2B, 3B). In addition, the distribution of other protein domains was generally group specific. For instance, the

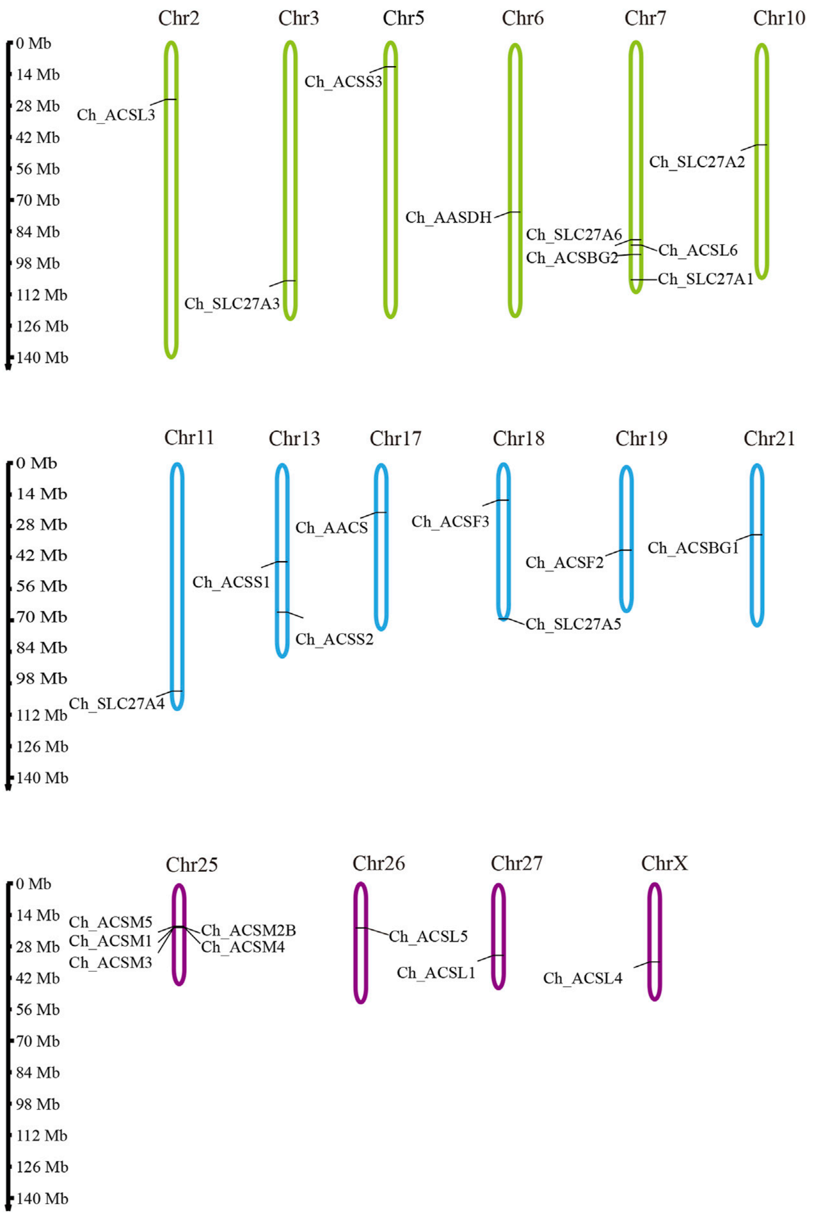


FIGURE 4
The chromosomal distribution of ChACS genes. Legends: The chromosomal position of each ACS gene was mapped to the goat genome. The chromosome number is indicated at the top of each chromosome.

acetyl-coenzyme A synthetase N-terminus (ACAS_N) domain was shared by all proteins in AACS subfamily. The AMP-binding enzyme C-terminal (AMP-binding_C) domain appeared in ACSS, ACSM, and ACSVL subfamily. Other domains, including phosphopantetheine attachment site (PP-binding), PQQ_2 and PQQ_3, were exclusive to AASDH proteins. Notably, *ChSLC27A6* and *ChACSL6* contained the AMP-binding_C domain, while *BtSLC27A6* and *BtACSL6* did not, indicating the function of *SLC27A6* and *ACSL6* genes might have changed in goat and cow (Figure 3B).

## acyl-coenzyme A synthetase gene structure, chromosomal location, and conserved motif analysis

The structural diversity of exon-introns is considered to play an important role in genetic evolution. Therefore, we performed exon–intron structure analysis to explore the structural evolution mechanism of the ACS genes using the GSDS tool. The result showed that the number of introns in the ChACS genes contained from 9 to 21 introns (Tables 1, 2), which was

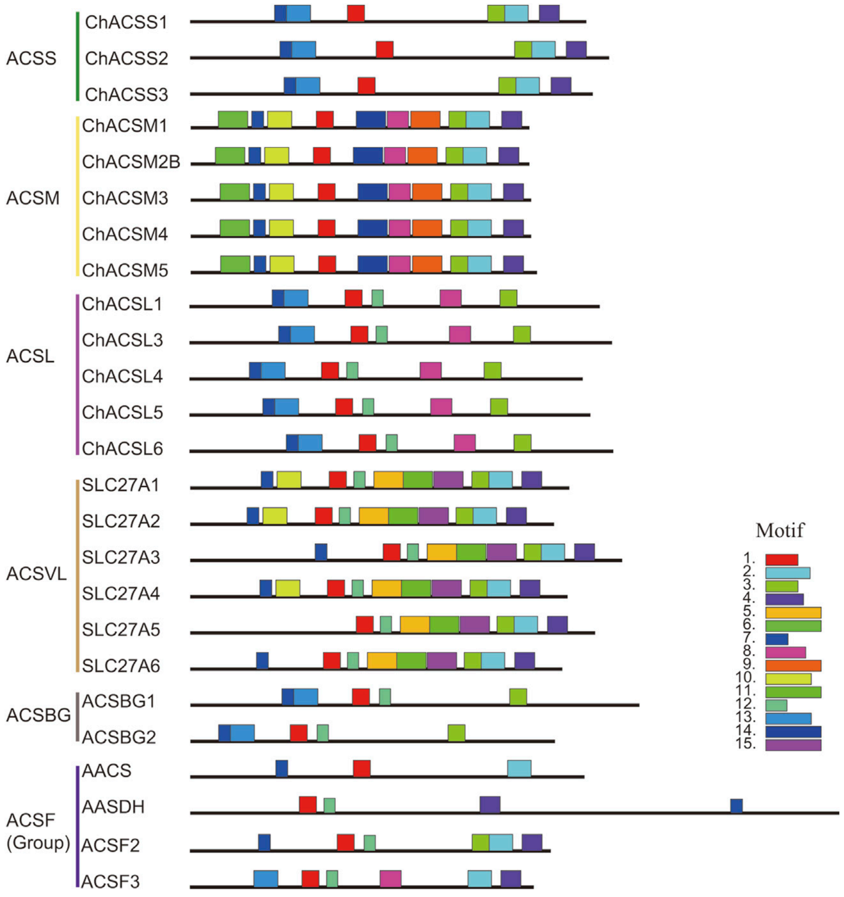


FIGURE 5
Motif distribution in ACS genes from goat. Legends: Motifs were predicted using the MEME web server (https://www.swissmodel.expasy.org/). The motifs are represented by different colors. The length of each box in the figure does not represent the actual motif size.

essentially consistent with what was found for the BtACS (Figures 2C, 3C). The exon-intron structure data also supported the phylogenetic tree topological structure. For example, the ACSVL subfamily contained 9–12 introns, and the ACSM subfamily had 12–14 introns (Table 1). The majority of ChACS genes had both 5′-and 3′-untranslated regions (UTRs), the *ChSLC27A3* contained a 5′ -UTR only, and the *ChACM4* contained no UTR region (Figures 2C, 3C).

To investigate the distribution of ChACS genes in the goat genome, the ChACS genes were mapped to individual chromosomes. The 25 ChACSs were distributed on 16 chromosomes, with five *ChACSs* on chromosome 25, four *ChACSs* on chromosome 7, two genes on chromosomes 25 and 28, and one *ChACS* on each of chromosomes 2, 3, 5, 6, 10, 11, 17, 19, 21, 26, 27, and X (Figure 4).

In this study, the MEME tool was used to identify conserved motifs. A total of 15 different conserved motifs were predicted in the ChACS proteins (Figure 5). All of the 25 ACS proteins contained Motif 1, *ChACSF3* did not contain Motif 7, while three proteins (*ChAACS*, *ChAACDH*, and *ChACSF3*) did not contain Motif 3. The closest ACS proteins in the phylogenetic tree had similar motifs. For example, each ACSM group member contained 10 motifs with similar motif composition pattern, which was very distinct from that of proteins in the other groups. Notably, more than 22 ACS proteins contained motif 1 and motif 3 (Figure 6), which were components of the AMP-binding domain and played vital roles in substrate binding and/or catalysis (Watkins et al., 2007).

## Expression analysis of acyl-coenzyme A synthetase genes

To analyze the expression profiles of ACS genes in distinct tissues of dairy goat, we investigated the FPKM values of the ACS genes in heart, kidney, liver, mammary gland, and muscle tissues. The majority of the 25 ACS genes were expressed in the test tissues, and the expression of the *ACSBG2* was not detected in

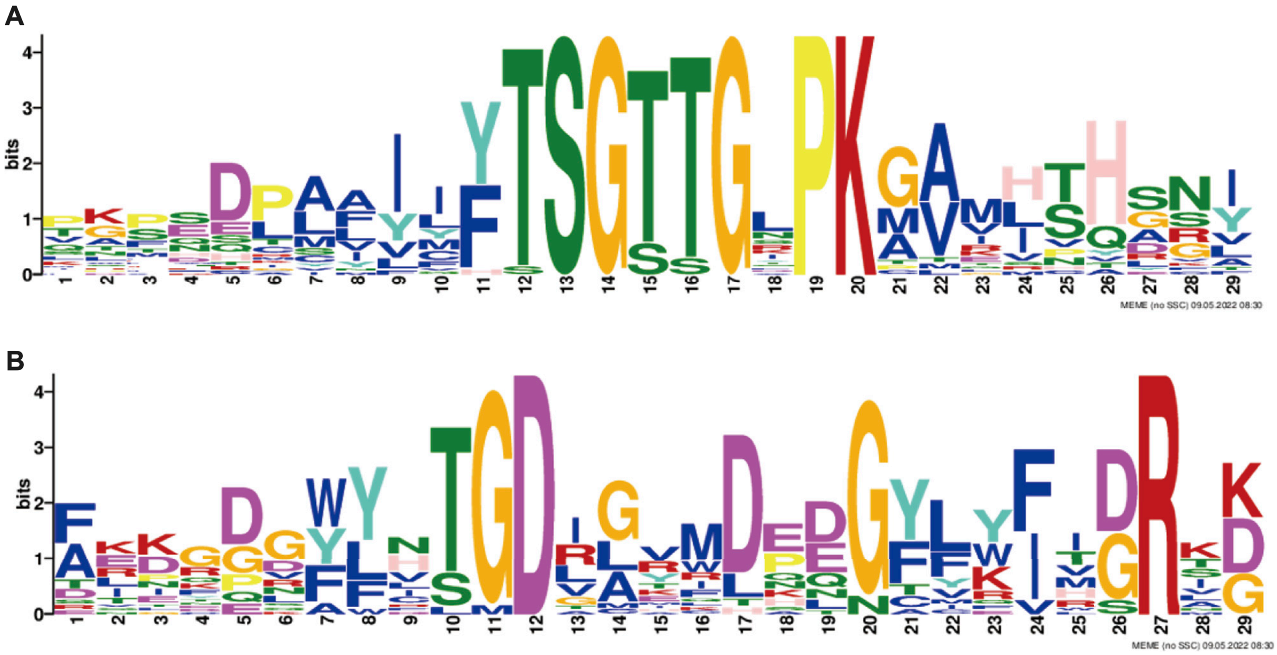


FIGURE 6
The conserved **(A)** motif 1 and **(B)** motif 3. Legends: These logos are graphic representations of amino acid sequences obtained from multiple sequence alignments. The motif 1 and motif 3, both of which were all components of the AMP-binding domain and played vital roles in substrate binding and/or catalysis (Watkins et al., 2007). The larger the fonts, the more conserved the motifs.

multi tissues. Only *ACSL1* was highly expressed in all tissues. In general, the expression profiles analysis suggested that the expression of ACS genes of the dairy goat was related to distinct tissues and the expression patterns were also distinct among each ACS subfamily (Figure 7; Supplementary Sheet S3). For example, the members of the ACSM subfamily, except for *ACSM4*, exhibited higher expression level in kidney and liver than in other tissues. In heart, *ACSS1* was most highly expressed, and the ACS genes, including *ACSL1*, *ACSL4*, *ACSF2*, *ACSS1*, and *ACSS2*, were highly expressed. Most of ACS genes were highly expressed in kidney with *ACSM1* expression was prominent. Similarly, *ACSM1* was also highly expressed in liver. The majority of ACS genes exhibited lower gene expression in mammary gland and muscle tissues. The highest expression was observed in mammary gland and muscle for *ACSS2* and *ACSL1*, respectively. These finding suggested ACS genes had different functions in distinct tissues.

To explore the ACS genes expression profiles in mammary glands among dairy goat, sheep and cow, we compared their transcript abundance in the non-lactation and lactation period (Figures 2A, 3A). The expression pattern indicated that the expression levels of ACS genes was associated with each ACS protein group. For instance, the proteins of ACSM and ACSBG groups showed lower expression in three species, and members of ACSS and ACSL groups exhibited high expressions, except for *ACSL6*. Some ACS genes showed identical expression trend in dairy goat and sheep, while exhibited distinct expression patterns in cow (Supplementary Sheet S4). For example, the expression of *ChACSS2* and *OaACSS2* was the most highly expressed relative to other members in lactation stage, while *BtACSS2* was not preferentially expressed. In addition, the *ACSS2* exhibited higher transcript abundance level in lactation stage than in non-lactation stage in dairy goats and sheep, while the opposite was observed in cow. Whereas, the transcript abundance of *ACSL4* and *ACSF3* was significantly lower in lactation stage than in non-lactation stage in dairy goat and sheep, which showed opposite result in cow. Some genes such as *ACSM3*, *ACSM5*, and *ACSS1* exhibited the same expression trend in dairy goat and cow, which were lower expression level in non-lactation period than in lactation period. These differences in expression profiles of goat and cow suggested that the ASC genes might have an effect on composition of fatty acids in goat milk.

## Discussion

Previous studies have shown that fatty acids are activated to fatty acyl-CoAs by ACSs before involving in both anabolic and catabolic processes. Processes such as the synthesis of acylated protein and comple lipids, fatty acid extension or unsaturation, and fatty acid oxidation all require activated fatty acid substrates (Steinberg et al., 2000). This study found that a total of 25 full-length ACSs (without

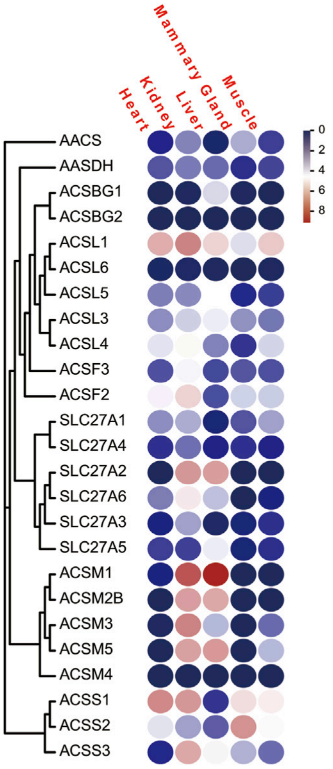


FIGURE 7
The FPKM of the of ACS genes of the dairy goat in five different tissues. Legends: The heat map was created with the log-transformed values of the FPKM values of ACS genes. Color gradient from navy-white-red indicates expression values change from low to high.

*ACSM2A*), representing five subfamilies (ACSS, ACSM, ACSL, ACSVL, ACSBG) and one group (ACSF) in the *Capra hircus* reference genome. In addition, we performed analyses of their phylogeny relationships, gene structures, conserved domains and motifs, expression profiling in multi tissues, and expressed difference in the non-lactation and lactation mammary glands of three species (goat, sheep, and cow).

There are 26 members of the ACS gene family reported in mammals (Watkins et al., 2007). Surprisingly, *ACSM2A* is missing in goat, sheep and cow genomes in the present study. The biological function of *ACSM2A* has rarely been reported in the current literature (Rencia and Erasmus, 2016; van der sluis et al., 2018). Several reports have shown that the *ACSM2A* and *ACSM2B* are nearly identical, with nucleotide homology of 98.8% and an amino acid identity of 97.1%. However, evidence also suggests that they are distinct genes (Rencia and Erasmus, 2016). Our findings showed the *ChACSM2B* was homologous to the *HsACSM2A* and *HsACSM2B* (Watanabe et al., 2020), with the amino acid identity of 80.6% and 80.4% respectively. Thus, *ChACSM2B* is highly conserved in the evolutionary process and might have similar function to *HsACSM2A*. Phylogenetic analysis exhibited that each evolutionary branch of the ACS family contained goat, sheep, cow and human proteins, suggesting that possible functions may be conserved among species (Liu et al., 2020). The four members of ACSF group, including *AACS*, *ACSF2*, *ACSF3*, and *AASDH*, clustered into four distinct clades, which was in consistent with the previous studies (Watkins et al., 2007). All the coding sequence of *ACSs* are disrupted by 9–21 introns, and the intron positions of goat and cow are distinct, suggesting that intron insertion may be result from independent events (Karan et al., 2003). All of ACS members contain related AMP-binding domains and FA binding motifs (Watkins et al., 2007). In this study, we found all ACS proteins shared a similar AMP-binding domain, which also illustrated that the AMP-binding functional domain was a conserved sequence and directly participated in the catalytic reaction (Wang et al., 2022). In addition to AMP-binding domain, the distribution of other protein domains was generally group specific. The conserved motif analysis showed that ACSs motifs also shared group specific, and the closest ACS proteins in the phylogenetic tree had similar motifs. These findings suggest that a wide range and diversity of ACSs might be result from suffering selective pressures to adapt to the metabolism of numerous and complex fatty acids in the evolutionary process (Watkins et al., 2007).

Each ACS gene plays a unique role, channeling its CoA derivatives to a specific metabolic pathway (Rencia and Erasmus, 2016). Fatty acyl-CoA molecules are the important regulatory molecules and metabolic intermediates (He et al., 2022), which have a variety of functions in the metabolism. Acyl-CoAs are oxidized to provide cellular energy, and are instrumental in the synthesis of acylated protein, and complex lipids such as triacylglycerols and phospholipids (Wang et al., 2022). Although *ACSL1* is highly expressed in heart, kidney, liver, mammary gland, and skeletal muscle, it has been shown to have different functions in different tissues because of the dual location on both the mitochondria and the endoplasmic reticulum (Grevengoed et al., 2014). The ACSM proteins are considered as liver mitochondrial enzymes (Rencia and Erasmus, 2016), which were also confirmed in this study. We also observed that the ACSM proteins were highly expressed in kidney, indicating that the ACSM family members might also play vital roles in kidney fatty acids metabolism. Several of the ACSVL family members located in the internal cellular membranes may have separate transport and activation functions, while ACSVLs not located on the plasma membrane are thought to enhance cellular uptake of fatty acids (Grevengoed et al., 2014). Multiple ACS proteins were expressed simultaneously in the same tissue, suggesting that these genes may coordinate together to perform a similar function. Several genes in the same subfamily exhibited distinct expression pattern from other members, such as *ACSL1* in ACSL subfamily and *SLC27A6* in ACSVL subfamily, suggesting that they may be involved in different biological functions.

vBCFAs are the cause of the goaty flavour of goat milk (Teng et al., 2018) and are more likely to be synthesized in goat tissues than in rumen microbes (Berthelot et al., 2001). It is believed that methylmalonate formed by carboxylation of propionic acid is an essential substance for the synthesis of vBCFAs (Priolo et al., 2001). Thus, *ACSs* might be involved in the synthesis or regulation of vBCFAs. By investigating the differences of expressed profiles in non-lactation and lactation mammary glands between goat and cow, it was possible to explore which ACS genes participate in formation of the goaty flavour of goat milk. In this study, the analysis of expression profiles suggests that *ACSF3* showed identical expression trend in dairy goat and sheep, while exhibited distinct expression patterns in cow. *ACSF3* is able to activate malonic acid to malonyl CoA (Chen et al., 2011), and also catalyzes methylmalonic acid to methylmalonyl-CoA (Monteuuis et al., 2017). It was reasonable to speculate that *ACSF3* provided the substrate for the synthesis of vBCFAs. ACS catalyzes propionate to propionyl-CoA which is a primer for the synthesis of 4-methyloctanoic acid (4-Et-8:0), and butyryl-CoA is the primer for the synthesis of 4-ethyloctanoic acid (4-Et-8:0). In this study, ACSS subfamily members were highly expressed, and the *ChACSS2* was the gene with the highest expression among all *ChACSs*, but *BtACSS2* was not preferentially expressed in cow. ACSSs typically activate short-chain fatty acids like acetate, propionate, or butyrate, involving in energy metabolism (Watkins et al., 2007). Our subcellular localization results have also proven that *ACSS1* and *ACSS3* were the mitochondrial proteins (Moffett et al., 2020; He et al., 2022), and *ACSS2* was localized in cytoplasm and nucleus (Li et al., 2017). As the fatty acids are activated in cytoplasm, our results suggest that *ACSS2* might participate in the synthesis of vBCFAs.

Goat milk contains high amounts of short-chain and medium-chain fatty acids (Luna et al., 2008). The oxidation of fatty acids is initiated in the cytoplasm by the formation of acyl-CoA by ACSs that are located in the endoplasmic reticulum and mitochondrial outer membrane (Watkins and Ellis, 2012). ACSMs typically activate medium-chain fatty acids, while ACSLs have a preference for long-chain fatty acids (Rencia and Erasmus, 2016). In this study, ACSMs were expressed at relatively low levels in lactation stage. We suggest that low expression of ACSMs is a possible reason for the high concentration of free medium-chain fatty acids in goat milk. We also observed that the genes including *ChACSL4*, *ChACSL5*, *OaACSL4*, and *OaACSL5* exhibited a lower transcript abundance in lactation stage than in non-lactation stage, but *BtACSL4* and *BtACSL5* showed opposite trend. Thus, we propose that *ACSL4* and *ACSL5* may have an effect on the long-chain fatty acid content between goat milk and cow's milk.

ACSVL subfamily members are integral transmembrane proteins, which play a vital role in the absorption of long-chain fatty acids into cells (Gallardo et al., 2013). *ACSF2* is a mitochondrial matrix enzyme involved in the tricarboxylic acid cycle and fatty acid synthesis (Yang et al., 2019). *AASDH* is a protein of unknown function and homologous to bacterial non-ribosomal peptide synthetase (Drozak et al., 2014). Previous studies and the current research suggest these genes may not related to the metabolic process of vBCFAs.

## Conclusion

A total of 25 ACS genes were characterized in goats and subdivided into five subfamilies. The ACS proteins all had the conserved the AMP-binding domain and motif1. The phylogenetic relationships of *ACSs* were also supported by gene structures, motifs and protein domain. The majority of the ACS genes were expressed in the multi tissues, with similar or different expression levels. These findings provide reference information to further understand the classification and putative functions of ACS genes in goats. Two genes, *ACSS2* and *ACSF3*, may take part in the synthesis of vBCFAs. This study also provides genomic and expression information for *ACSs* in goat, and the findings may be useful for further research on the formation mechanisms of the goaty flavour in goat milk.

## Data availability statement

The sequences information analyzed in this study are available in the GenBank (https://www.ncbi.nlm.nih.gov/), and the Ensemble database (http://ftp.ensembl.org/pub/release-104/). The RNA-seq data was from the NCBI's SRA database (Accession: PRJNA309284, PRJNA309345, PRJNA339650, PRJNA637690, and PRJNA482783).

## Ethics statement

The animal study was reviewed and approved by the Animal Care and Use Committee of the Northwest A&F University.

## Author contributions

JL was responsible for experimental design. FZ, CS, LZ, JW, and JZ performed data curation. FZ wrote the manuscript. FZ, JL, CL, HT, and QH participated in result interpretation and manuscript revision. All authors read and approved the final manuscript.

## Funding

This work was jointly supported by the National Natural Science Foundation of China (No. 31772575), the Key Research and Development Plan of Shaanxi Province, China (No. 2021ZDLNY02-02).

## Acknowledgments

Thanks a lot to all the research workers in the Dairy Goat Lab of Northwest A&F University for their useful discussion.

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.980463/full#supplementary-material

## References

Ansari, I. H., Longacre, M. J., Stoker, S. W., Kendrick, M. A., O'Neill, L. M., Zitur, L. J., et al. (2017). Characterization of Acyl-CoA synthetase isoforms in pancreatic beta cells: Gene silencing shows participation of ACSL3 and ACSL4 in insulin secretion. *Arch. Biochem. Biophys.* 618, 32–43. doi:10.1016/j.abb.2017.02.001

Balakrishnan, S., Gao, S., Lercher, M. J., Hu, S., and Chen, W. H. J. N. A. R. (2019). Evolview v3: A webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47, W270–W275. doi:10.1093/nar/gkz357

Berthelot, V., Bas, P., Schmidely, P., and Duvaux-Ponter, C. J. S. R. R. (2001). Effect of dietary propionate on intake patterns and fatty acid composition of adipose tissues in lambs. *Small Rumin. Res.* 40 (1), 29–39. doi:10.1016/s0921-4488(00)00217-0

Black, P. N., Zhang, Q., Weimar, J. D., and DiRusso, C. C. (1997). Mutational analysis of a fatty acyl-coenzyme A synthetase signature motif identifies seven amino acid residues that modulate fatty acid substrate specificity. *J. Biol. Chem.* 272 (8), 4896–4903. doi:10.1074/jbc.272.8.4896

Brennand, C. P., Kim, H. J., and Lindsay, R. C. J. (2010). Aroma properties and thresholds of some branched-chain and other minor volatile fatty acids occurring in milkfat and meat lipids. *J. Sens. Stud.* 4 (2), 105–120. doi:10.1111/j.1745-459X.1989.tb00461.x

Chen, H., Kim, H. U., Weng, H., and Browse, J. J. P. C. (2011). Malonyl-CoA synthetase, encoded by ACYL ACTIVATING ENZYME13, is essential for growth and development of arabidopsis. *Plant Cell* 23 (6), 2247–2262. doi:10.1105/tpc.111.086140

Drozak, J., Veiga-Da-Cunha, M., Kadziolka, B., and Van Schaftingen, E. J. F. J. (2014). Vertebrate AcylCoA synthetase family member4 (ACSF4-U26) is a β-alanine-activating enzyme homologous to bacterial non-ribosomal peptide synthetase. *FEBS J.* 281 (6), 1585–1597. doi:10.1111/febs.12725

Fernandez, R. F., Kim, S. Q., Zhao, Y., Foguth, R. M., Weera, M. M., Counihan, J. L., et al. (2018). Acyl-CoA synthetase 6 enriches the neuroprotective omega-3 fatty acid DHA in the brain. *Proc. Natl. Acad. Sci. U. S. A.* 115 (49), 12525–12530. doi:10.1073/pnas.1807958115

Finn, R. D., Clements, J., and Eddy, S. R. J. N. A. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39, 29–37. doi:10.1093/nar/gkr367

Gallardo, D., Amills, M., Quintanilla, R., and Pena, R. N. (2013). Mapping and tissue mRNA expression analysis of the pig solute carrier 27A (SLC27A) multigene family. *Gene* 515 (1), 220–223. doi:10.1016/j.gene.2012.11.029

Grevengoed, T. J., Klett, E. L., and Coleman, R. A. (2014). Acyl-CoA metabolism and partitioning. *Annu. Rev. Nutr.* 34, 1–30. doi:10.1146/annurev-nutr-071813-105541

Ha, J. K., and Lindsay, R. C. (1993). Release of volatile branched-chain and other fatty acids from ruminant milk fats by various lipases. *J. Dairy Sci.* 76 (3), 677–690. doi:10.3168/jds.s0022-0302(93)77391-9

He, W., Zhou, X., Wu, Q., Zhou, L., Zhang, Z., Zhang, R., et al. (2022). Acetyl CoA synthase 2 potentiates ATG5-induced autophagy against neuronal apoptosis after subarachnoid hemorrhage. *J. Mol. Histol.* 10, 511–521. doi:10.1007/s10735-022-10057-x

Kaffarnik, S., Kayademir, Y., Heid, C., and Vetter, W. J. J. (2014). Concentrations of volatile 4-alkyl-branched fatty acids in sheep and goat milk and dairy products. *J. Food Sci.* 79 (11), C2209–C2214. doi:10.1111/1750-3841.12673

Kanter, J. E., Tang, C., Oram, J. F., and Bornfeldt, K. E. (2012). Acyl-CoA synthetase 1 is required for oleate and linoleate mediated inhibition of cholesterol efflux through ATP-binding cassette transporter A1 in macrophages. *Biochim. Biophys. Acta* 1821 (3), 358–364. doi:10.1016/j.bbalip.2011.10.008

Katoh, K., and Standley, D. M (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi:10.1093/molbev/mst010

Karan, D., Lesbats, M., David, J. R., and Capy, P. J. J. (2003). Evolution of the AMP-forming acetyl-CoA synthetase gene in the Drosophilidae family. *J. Mol. Evol.* 57 (1), S297–S303. doi:10.1007/s00239-003-0040-1

Kochan, G., Pilka, E. S., Delft, F. V., Oppermann, U., and Yue, W. W. (2009). Structural snapshots for the conformation-dependent catalysis by human medium-chain acyl-coenzyme A synthetase ACSM2A. *J. Mol. Biol.* 388 (5), 997–1008. doi:10.1016/j.jmb.2009.03.064

Lam-Tung, N., Schmidt, H. A., Arndt, V. H., and Quang, M. B. J. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 1, 268–274. doi:10.1093/molbev/msu300

Li, C., Zhu, J., Shi, H., Luo, J., and Loor, J. J. (2020). Comprehensive transcriptome profiling of dairy goat mammary gland identifies genes and networks crucial for lactation and fatty acid metabolism. *Front. Genet.* 11, 878. doi:10.3389/fgene.2020.00878

Li, X., Yu, W., Qian, X., Xia, Y., Zheng, Y., Lee, J. H., et al. (2017). Nucleus-translocated ACSS2 promotes gene transcription for lysosomal biogenesis and autophagy. *Mol. Cell* 66 (5), 684–697. doi:10.1016/j.molcel.2017.04.026

Liu, Z., Ge, X. X., Wu, X. M., Xu, Q., and Guo, W. W. J. B. G. (2020). Genome-wide analysis of the citrus B3 superfamily and their association with somatic embryogenesis. *BMC Genomics* 21 (1), 305. doi:10.1186/s12864-020-6715-9

Luna, P., Bach, A., Juárez, M., and de la Fuente, M. A. (2008). Effect of a diet enriched in whole linseed and sunflower oil on goat milk fatty acid composition and conjugated linoleic acid isomer profile. *J. Dairy Sci.* 91 (1), 20–28. doi:10.3168/jds.2007-0447

Mashek, D. G., Bornfeldt, K. E., Coleman, R. A., Berger, J., Bernlohr, D. A., Black, P., et al. (2004). Revised nomenclature for the mammalian long-chain acyl-CoA synthetase gene family. *J. Lipid Res.* 45 (10), 1958–1961. doi:10.1194/jlr.E400002-JLR200

Meller, N., Morgan, M. E., Wong, W. P., Altemus, J. B., and Sehayek, E. J. L. (2013). Targeting of Acyl-CoA synthetase 5 decreases jejunal fatty acid activation with no effect on dietary long-chain fatty acid absorption. *Lipids Health Dis.* 12 (1), 88–94. doi:10.1186/1476-511X-12-88

Moffett, J. R., Puthillathu, N., Vengilote, R., Jaworski, D. M., and Namboodiri, A. M. (2020). Acetate revisited: A key biomolecule at the nexus of metabolism, epigenetics, and oncogenesis - Part 2: Acetate and ACSS2 in health and disease. *Front. Physiol.* 11, 580171. doi:10.3389/fphys.2020.580171

Monteuuis, G., Suomi, F., Kertr, J. M., Masud, A. J., and Kastaniotis, A. J. J. B. J. (2017). A conserved mammalian mitochondrial isoform of acetyl-CoA carboxylase ACC1 provides the malonyl-CoA essential for mitochondrial biogenesis in tandem with ACSF3. *Biochem. J.* 474 (22), 3783–3797. doi:10.1042/BCJ20170416

Ndiaye, H., Liu, J. Y., Hall, A., Minogue, S., Morgan, M. Y., and Waugh, M. G. (2020). Immunohistochemical staining reveals differential expression of ACSL3 and ACSL4 in hepatocellular carcinoma and hepatic gastrointestinal metastases. *Biosci. Rep.* 40 (4), BSR20200219. doi:10.1042/BSR20200219

O'Brien, M. J., Beijerink, N. J., Sansom, M., Thornton, S. W., Chew, T., and Wade, C. M. (2020). A large deletion on CFA28 omitting ACSL5 gene is associated with intestinal lipid malabsorption in the Australian Kelpie dog breed. *Sci. Rep.* 10 (1), 18223. doi:10.1038/s41598-020-75243-x

Perez-Nunez, I., Karaky, M., Fedetz, M., Barrionuevo, C., Izquierdo, G., Matesanz, F., et al. (2019). Splice-site variant in ACSL5: A marker promoting opposing effect on cell viability and protein expression. *Eur. J. Hum. Genet.* 27 (12), 1836–1844. doi:10.1038/s41431-019-0414-5

Priolo, A., Micol, D., and Agabriel, J. J. A. R. (2001). Effects of grass feeding systems on ruminant meat colour and flavour. A review. *Anim. Res.* 50 (3), 185–200. doi:10.1051/animres:2001125

Rencia, V., and Erasmus, E. J. E. O. D. M. T. (2016). Xenobiotic/medium chain fatty acid: CoA ligase – a critical review on its role in fatty acid metabolism and the detoxification of benzoic acid and aspirin. *Expert Opin. Drug Metab. Toxicol.* 12 (10), 1169–1179. doi:10.1080/17425255.2016.1206888

Rossi Sebastiano, M., and Konstantinidou, G. (2019). Targeting long chain acyl-CoA synthetases for cancer therapy. *Int. J. Mol. Sci.* 20 (15), E3624. doi:10.3390/ijms20153624

Salles, C., Sommerer, N., Septier, C., Issanchou, S., Science, J. L. Q., Garem, A., et al. (2010). Goat cheese flavor: Sensory evaluation of branched-chain fatty acids and small peptides. *J. Food Sci.* 67 (2), 835–841. doi:10.1111/j.1365-2621.2002.tb10686.x

Steinberg, S. J., Morgenthaler, J., Heinzer, A. K., Smith, K. D., and Watkins, P. A. (2000). Very long-chain acyl-CoA synthetases. Human "bubblegum" represents a new family of proteins capable of activating very long-chain fatty acids. *J. Biol. Chem.* 275 (45), 35162–35169. doi:10.1074/jbc.M006403200

Takahiro, F., and Tokuo, Y. J. J. o. B. (1992). Cloning and functional expression of a novel long-chain acyl-CoA synthetase expressed in brain. *J. Biochem.* 111 (2), 197–203. doi:10.1093/oxfordjournals.jbchem.a123737

Tang, Y., Zhou, J., Hooi, S. C., Jiang, Y. M., and Lu, G. D. (2018). Fatty acid activation in carcinogenesis and cancer development: Essential roles of long-chain acyl-CoA synthetases. *Oncol. Lett.* 16 (2), 1390–1396. doi:10.3892/ol.2018.8843

Teng, F., Reis, M. G., Ma, Y., and Day, L. (2018). Effects of season and industrial processes on volatile 4-alkyl-branched chain fatty acids in sheep milk. *Food Chem.* 260, 327–335. doi:10.1016/j.foodchem.2018.04.011

van der Sluis, R., Genetics, R. J. M., and Mgg, G. (2018). Analyses of the genetic diversity and protein expression variation of the acyl: CoA medium-chain ligases, ACSM2A and ACSM2B. *Mol. Genet. Genomics* 293 (5), 1279–1292. doi:10.1007/s00438-018-1460-3

Wang, Y., Guo, L., Zhao, Y., Zhao, X., and Yuan, Z. (2022). Systematic analysis and expression profiles of the 4-coumarate: CoA ligase (4CL) gene family in pomegranate (punica granatum L.). *Int. J. Mol. Sci.* 23 (7), 3509. doi:10.3390/ijms23073509

Watanabe, H., Paxton, R. L., Tolerico, M. R., Nagalakshmi, V. K., Gomez, R. A., Okusa, M. D., et al. (2020). Expression of Acsm2, a kidney-specific gene, parallels the function and maturation of proximal tubular cells. *Am. J. Physiol. Ren. Physiol.* 319 (4), F603–F611. doi:10.1152/ajprenal.00348.2020

Watkins, P. A., and Ellis, J. M. (2012). Peroxisomal acyl-CoA synthetases. *Biochim. Biophys. Acta* 1822 (9), 1411–1420. doi:10.1016/j.bbadis.2012.02.010

Watkins, P. A., Maiguel, D., Jia, Z., and Pevsner, J. (2007). Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome. *J. Lipid Res.* 48 (12), 2736–2750. doi:10.1194/jlr.M700378-JLR200

Yang, Y., An, C., Yao, Y., Cao, Z., Gu, T., Xu, Q., et al. (2019). Intron polymorphisms of MAGI-1 and ACSF2 and effects on their expression in different goose breeds. *Gene* 701, 82–88. doi:10.1016/j.gene.2019.02.102

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/fvets.2022.932034

Check for updates

OPEN ACCESS

EDITED BY
Xiangdong Ding,
China Agricultural University, China

REVIEWED BY
Yao Jiang,
China Agricultural University, China
Yuchun Pan,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Zhipeng Wang
wangzhipeng@neau.edu.cn

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

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

RECEIVED 29 April 2022
ACCEPTED 09 September 2022
PUBLISHED 04 October 2022

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

# Genome-wide association analysis of milk production, somatic cell score, and body conformation traits in Holstein cows

Peng Wang[1†], Xue Li[2,3†], Yihao Zhu[1], Jiani Wei[4], Chaoxin Zhang[2,3], Qingfang Kong[1], Xu Nie[1], Qi Zhang[5] and Zhipeng Wang[2,3*]

[1]Heilongjiang Animal Husbandry Service, Harbin, China, [2]College of Animal Science and Technology, Northeast Agricultural University, Harbin, China, [3]Bioinformatics Center, Northeast Agricultural University, Harbin, China, [4]School of mathematics, University of Edinburgh, Edinburgh, United Kingdom, [5]College of Animal Science and Technology, China Agricultural University, Beijing, China

Milk production and body conformation traits are critical economic traits for dairy cows. To understand the basic genetic structure for those traits, a genome wide association study was performed on milk yield, milk fat yield, milk fat percentage, milk protein yield, milk protein percentage, somatic cell score, body form composite index, daily capacity composite index, feed, and leg conformation traits, based on the Illumina Bovine HD100k BeadChip. A total of 57, 12 and 26 SNPs were found to be related to the milk production, somatic cell score and body conformation traits in the Holstein cattle. Genes with pleiotropic effect were also found in this study. Seven significant SNPs were associated with multi-traits and were located on the *PLEC, PLEKHA5, TONSL, PTGER4,* and *LCORL* genes. In addition, some important candidate genes, like *GPAT3, CEBPB, AGO2, SLC37A1*, and *FNDC3B,* were found to participate in fat metabolism or mammary gland development. These results can be used as candidate genes for milk production, somatic cell score, and body conformation traits of Holstein cows, and are helpful for further gene function analysis to improve milk production and quality.

KEYWORDS
**milk production traits, body conformation traits, pleiotropic effect, genome-wide association study, Holstein cattle**

## Introduction

Milk is a source of nutrients essential for human growth and development. The milk production traits are important for the dairy industry. Body conformation traits have been applied in several countries with the development of dairy cattle breeding since they are closely related to the health (1), productivity (2), lifetime (3), and calving ease (4) of cows. Some studies have identified the genetic correlation between body conformation traits and first lactation milk yield to be between 0.48 and 0.54 (5). These correlations are therefore very important for the dairy industry to improve the milk production traits and body conformation traits.

The rapid development of sequencing technology has revealed the cause variants of complex traits using genome-wide association analysis (GWAS). A study by Schennink et al. (6) has revealed *DGAT1* and *SCD1* to be highly associated with the composition of milk-fat (long-chain fatty acid). Kiser et al. (7) verified the TFAP2A gene to be related to the production of colostrum in Jersey cattle. It reported the genes *CDH2* and *GABRG2* to be related to the milk fat percentage and milk protein traits, respectively, in dual-purpose Xinjiang brown cattle (8). Bouwman et al. (9) and Vanvanhossou et al. (10) have reported the VEPH1 gene to be associated with conformation. However, the identified genes have not explained all genetic variances. There is a need to continue the search for novel genes related to some quantitative traits.

This study conducted GWAS using the Illumina Bovine HD100k(100k) BeadChip, for identifying important candidate genes or variants related to milk production, somatic cell score, and body conformation traits. There was an expectation for discovering novel genetic variations or candidate genes.

# Materials and methods

## Animal population

This experiment involved 1,313 cows from 7 different pastures in Heilongjiang Province. The use and care of the animals in this study were approved by the Animal Care Advisory Committee, Northeast Agricultural University (Harbin, China), and all the experimental procedures were according to the university's guidelines for animal research.

## Genotypes data

The samples were collected from the tail roots near the hips of the cows. The DNA in the hair was extracted and genotyped using Illumina Bovine HD100k BeadChip, containing 95,256 SNPs. The markers with minor allele frequencies < 0.05 and call rates < 0.90 were filtered out and individuals with a call rate of 0.80 or greater were selected. These SNPs were distributed across 29 chromosomes.

## Population stratification

The SNP genotypes of these individuals were used to estimate the population stratification based on principal component analysis (PCA), and Plink (version 1.9) (11) was used to analyze a total of 1,310 cows with 86,645 markers covering the whole genome to study the population structure (12). The software uses the default matrix construction method to construct G matrix and get the PCA results. We used R language (version 4.1.2)—ggplot 2 package to draw pictures. The PCA scatterplots (Figure 1) illustrate a clear population structure for the 1,310 individuals in the seven pastures cattle herds that comprised our study population.

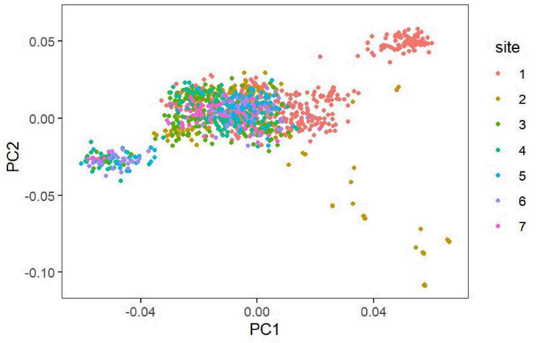


FIGURE 1
Population structure demonstrated by principal component analysis.

## Genome-wide association analysis

Combination with dairy herd improvement data of National Holstein cows in China, this study estimated the genomic estimated breeding values (GEBVs) of all animal milk production traits, somatic cell score, and body conformation traits, using single-step genomic best linear unbiased prediction (ssGBLUP). The ssGBLUP was developed to integrate all the information including genotypes, phenotypes, and pedigree information in one step, and each SNP effect was calculated using the FarmCPU method (13) based on the predicted GEBVs. The ssGBLUP method is an improvement of BLUP, in which the pedigree relationship matrix $a^{-1}$ matrix must be replaced by $H^{-1}$ (14). The specific model is as follows:

$$y = Xb + Zu + e$$

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}$$

Where y was each phenotypic value vector; b is the fixed effect of the field and the PCA effect to explain the population stratification, and u is a vector of animal effects. The e was a vector of random residual effects with e∼N(0,I), and X, Z were incidence matrices for b and u, respectively.

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau[(1-\mathrm{w})(\alpha + \mathrm{b} * \mathrm{G}) + \mathrm{w} * \mathrm{A}_{22}]^{-1} - \omega \mathrm{A}_{22}^{-1} \end{bmatrix}$$

Where, the A matrix is pedigree relationship matrix, $A_{22}$ is a numerator relationship matrix for genotyped animals, and G is a genomic relationship matrix (15). $G^{-1}$ was obtained as the inverse of a combination of the G matrix and the corresponding A matrix. The w is the weight of $A_{22}$ in the matrix, the default value is 0.05. The $\tau$ and $\omega$ are 1. We use DMU software to calculate the GEBV value. Both G and H matrices were derived using software default parameter setting by DMU software. G was calulated as:

$$G = \frac{WDW'}{2\sum_{i=1}^{n} p_i(1-p_i)}$$

Where $p_i$ is the allele frequency at locus i in all genotyped animals, is a normalizing constant (16) that sums expected variances across markers scaling G toward the A matrix (17), D is weight for each locus(I if same variance assumed), W is a design matrix as follows):

$$w_{ii} \begin{cases} 0 - 2p_i, homozygous \\ 1 - 2p_i, heterozygous \\ 2 - 2p_i, homozygous \end{cases}$$

Each SNP effect was calculated using the FarmCPU method (13) based on the predicted GEBVs. The FarmCPU method (13) in this study can be written as two models.

$$y = SNP_{\mathrm{i}} + K + e$$
$$y = pseudQTN + SNP_{\mathrm{i}} + e$$

The y is the GEBV value. The pseudoQTN is significant marker from previous loops that is null when the model begins. $\mathrm{SNP_i}$ is testing marker in each loop. The K is the kinship between each individuals. The e is residual vector.

For each trait, the threshold *P*-value for genome-wide significance was $5.99 \times 10^{-7} = 0.05/83446$ using the Bonferroni multiple test method.

## QTLs annotation analysis

The cattle QTL data were downloaded from the Cattle QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index) referred to as the ARS-UCD1.2 assembly. The square of the correlation coefficient ($r^2$) between the two loci is used to evaluate the range of LD measurement, because $r^2$ is considered to be more robust and not affected by changes in allele frequency and population size (18). Haploview software was used to calculate the genotype correlation coefficient ($r^2$) between all SNP pairs in the cow population to estimate the LD of the whole genome, and the LD decay map with distance of the cow population was visualized.

# Results

## Population stratification

The phenomenon of group stratification is an important research problem in the study of group association (19). In order to determine the population stratification level, we drew the population structure by principal component analysis (PCA). The PCA scatterplots shows the population structure of a 1,300 individual composed of seven pastures (Figure 1). Different colors represent different pastures. It can be seen that it is mainly divided into three clusters, but most of the cows in the seven pastures are gathered together, and only a few cows are separated. These clusters indicate that, although individuals may come from different ranches, they still retain close genetic relationships.

## The genome-wide association study

Basic descriptive statistics of milk production traits, somatic cell score and body conformation traits (see Table 1). A total of 86,645 SNPs were retained after quality control for the GWAS (Table 2). The average physical distance between the adjacent SNP markers was approximately 29.58 kb, ranging between 26.37 kb (BTA19) and 32.02 kb (BTA8).

The *p*-value profiles of all the SNP markers associated with each trait are represented in Figures 3, 4 and included the Manhattan and Quantile-Quantile plots. In total, 95 genome-wide significant SNPs were detected for the milk production traits, such as milk yield (MY), milk fat yield (FY), milk fat percentage (FP), milk protein yield (PY), milk protein percentage (PP), somatic cell score (SCS), and body conformation traits (body form composite index, BFCI; daily capacity composite index, DCCI; feed and leg conformation, FTLEG). There were 57, 12 and 26 SNPs related to milk production, somatic cell score and body conformation traits, respectively. Among them, we mainly focused on the first few significant SNPs in each trait. In addition, we also found seven SNPs that overlap with multiple traits, such as *PLEC* is related to MY, FP and PP, *PLEKHA5* is related to FP and FY, *TONSL* is connected with FY and SCS, *LCORL* is correlated with DCCI and FTLEG, *PYGB* is related to BFCI and FTLEG, and *PTGER4* is related to BFCI, and DCCI (see Table 3).

As shown in Tables 4, 5, 12, 11, 15, 11, 12, and 17 genome-wide significant SNPs were detected for MY, FY, FP, PY, PP, and

TABLE 1 Descriptive statistics of milk procuction trants and body conformation traits.

| Statistic | MY (kg) | FP (%) | PP (%) | SCS | BFCI | FTLEG | DCCI |
|---|---|---|---|---|---|---|---|
| Mean | 8382.99 | 3.85 | 3.31 | 4.02 | 85.68 | 85.46 | 85.91 |
| Standard Deviation | 1950.68 | 0.50 | 0.26 | 1.44 | 4.88 | 4.16 | 7.68 |
| Minimum | 1505.00 | 2 | 2.17 | 1.00 | 65.25 | 65.80 | 56.18 |
| Maximum | 15983.00 | 6.20 | 5.00 | 9.00 | 98.36 | 99.00 | 99.95 |
| Coefficient of Variation | 0.23 | 0.13 | 0.08 | 0.36 | 0.06 | 0.05 | 0.09 |

TABLE 2 Distribution of SNPs after quality control.

| BTA | Length (Mb) | No. SNP (Chip data) | No. SNP (after QC) | Density length/SNP(kb) |
|---|---|---|---|---|
| 1 | 158.53 | 5556 | 5188 | 30.56 |
| 2 | 136.23 | 4688 | 4367 | 31.20 |
| 3 | 121.01 | 4508 | 4158 | 29.10 |
| 4 | 120.00 | 4049 | 3760 | 31.92 |
| 5 | 120.09 | 4523 | 4083 | 29.41 |
| 6 | 117.81 | 4364 | 3977 | 29.62 |
| 7 | 110.68 | 3903 | 3551 | 31.17 |
| 8 | 113.32 | 3805 | 3539 | 32.02 |
| 9 | 105.45 | 3695 | 3469 | 30.40 |
| 10 | 103.31 | 3626 | 3376 | 30.60 |
| 11 | 106.98 | 3801 | 3522 | 30.38 |
| 12 | 87.22 | 3044 | 2842 | 30.69 |
| 13 | 83.47 | 3064 | 2822 | 29.58 |
| 14 | 82.40 | 3045 | 2796 | 29.47 |
| 15 | 85.01 | 3119 | 2885 | 29.47 |
| 16 | 81.01 | 2826 | 2586 | 31.33 |
| 17 | 73.17 | 2668 | 2506 | 29.20 |
| 18 | 65.82 | 2605 | 2389 | 27.55 |
| 19 | 63.45 | 2726 | 2406 | 26.37 |
| 20 | 71.97 | 2737 | 2498 | 28.81 |
| 21 | 69.86 | 2573 | 2374 | 29.43 |
| 22 | 60.77 | 2201 | 2038 | 29.82 |
| 23 | 52.50 | 2110 | 1951 | 26.91 |
| 24 | 62.32 | 2259 | 2081 | 29.95 |
| 25 | 42.35 | 1726 | 1589 | 26.65 |
| 26 | 51.99 | 1823 | 1708 | 30.44 |
| 27 | 45.61 | 1699 | 1624 | 28.09 |
| 28 | 45.94 | 1735 | 1630 | 28.18 |
| 29 | 51.10 | 1871 | 1731 | 29.52 |
| Total | 2489.37 | 90349 | 83446 | 29.58 |

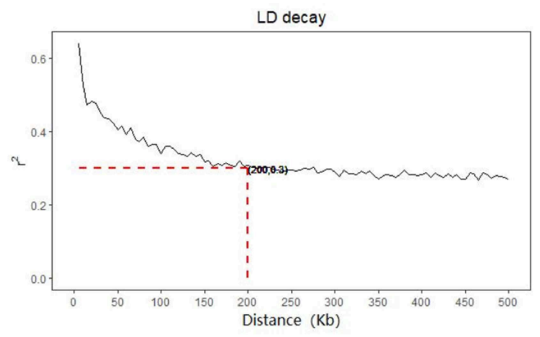


FIGURE 2
LD decay of cow.

SCS, respectively. These significant SNPs are mainly distributed in BTA 1, BTA 2, BTA 5, BTA 6, BTA 11, BTA 14, and BTA 20, with as many as 9 SNPs on BTA 14.

In addition, this study reported an interesting phenomenon where four SNPs were found to be related to multi-traits, including BovineHD0500025853 (BTA 5:90.66 Mb), BovineHD1400000206 (BTA 14:0.49 Mb), BovineHD1400000287 (BTA 14:0.88 Mb), and BovineHD1400011649 (BTA 14:38.57 Mb) (see Table 6). The bovinehd1400000287 SNP located in the 58th intron of the *PLEC* gene was found to be associated with MY, FP, and PP. The fat yield and the somatic cell score trait shared one SNP bovinehd1400000206 located 1.46 kb away from *TONSL* on BTA 14.

This study detected 10, 7, and 11 significant SNPs related to BFCI, DCCI, and FTLEG, respectively. There were 4 SNPs distributed on BTA 16. Three SNPs were found to be possibly as pleiotropism SNPs, including BovineHD4100004660 (BTA 6:38.22 Mb), BovineHD1300012605 (BTA 13:42.81 Mb) and BTA-50244-no-rs (BTA 20:34.30 Mb), respectively. Of these significant SNPs, the BTA-50244-no-rs SNP related to BFCI ($P$ = 5.84E-13) was located downstream of the *PTGER4* gene.

## QTL annotation analysis

The LD of cows decreases with the increase of distance, when the distance is extended to 200 Kb, the

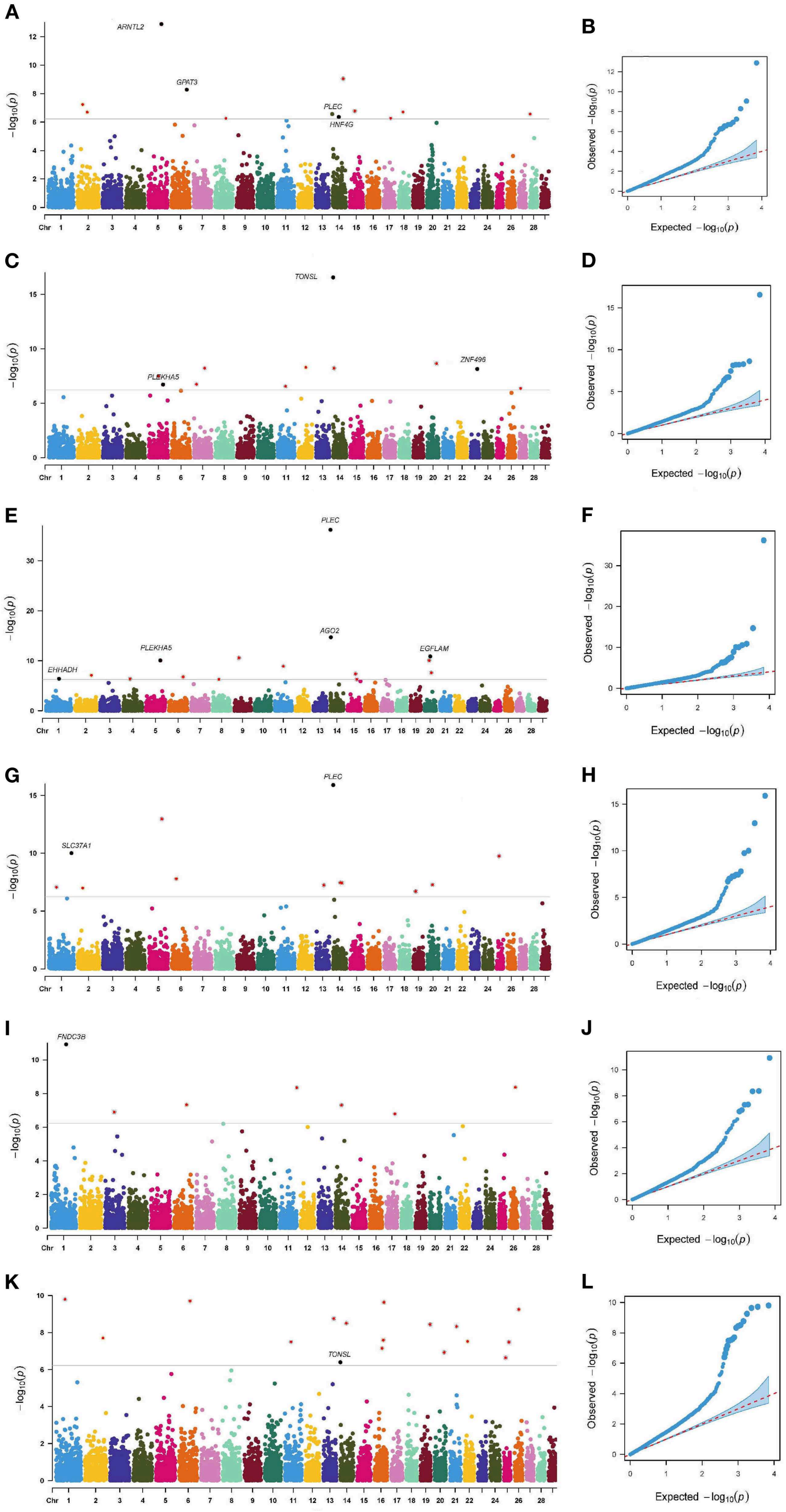


FIGURE 3
Manhattan plots and Quantile-Quantile plots for the milk production and somatic cell score traits. MY **(A,B)**, FY **(C,D)**, FP **(E,F)**, PP **(G,H)**, PY **(I,J)** and SCS **(K,L)**.

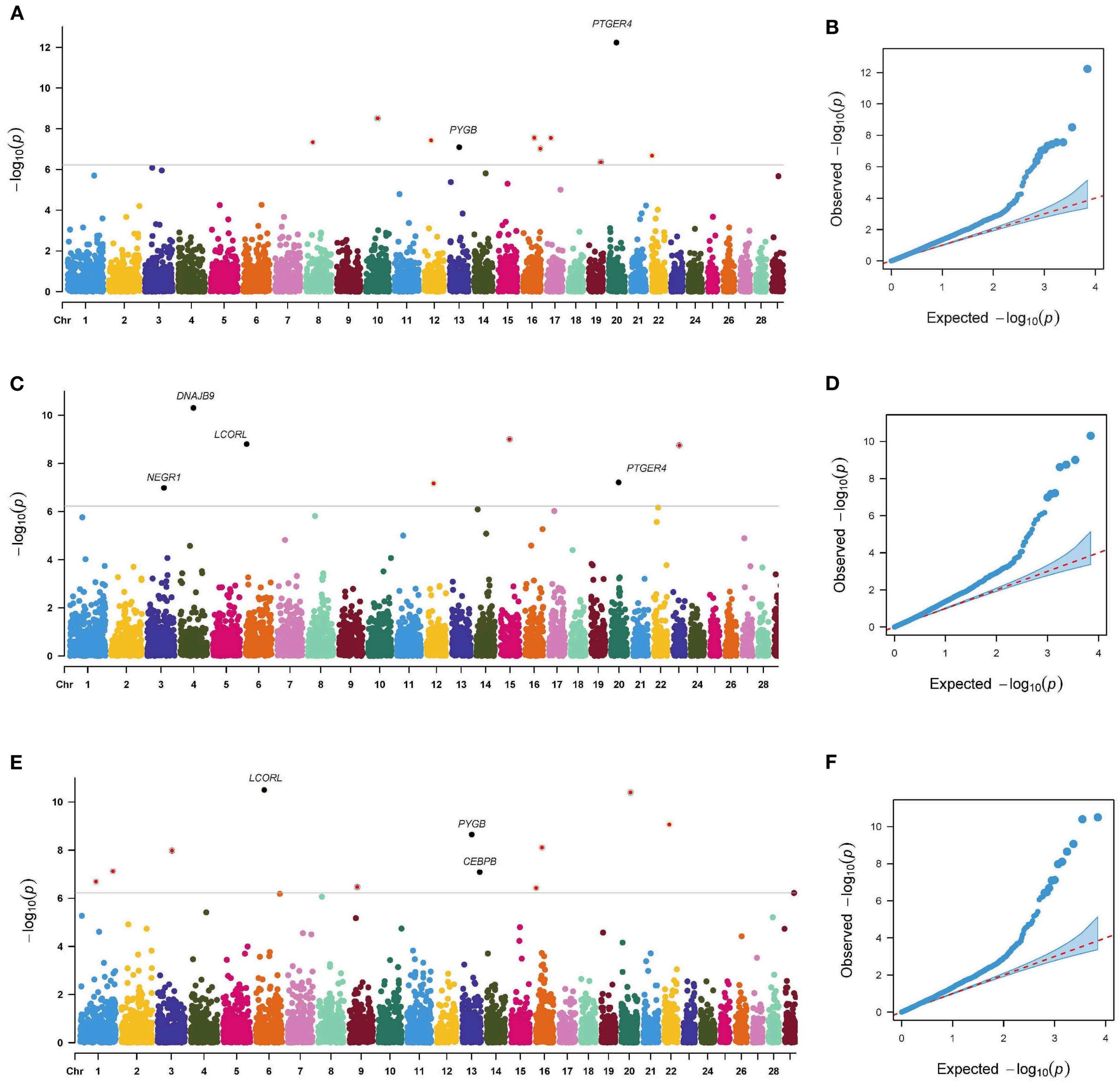


FIGURE 4
Manhattan plots and Quantile-Quantile plots for the body conformation traits. BFCI **(A,B)**, DCCI **(C,D)** and FTLEG **(E,F)**.

decline rate of LD of cows tends to be gentle, and the average $r^2$ value of cows is 0.3 at this time (show as Figure 2). The 100 Kb range of SNP upstream and downstream of significant trait association obtained from genome-wide association analysis is compared with the data that has been verified in the current cattle QTL database. Our significant SNPs associated with MY, FY, FP, PP and SCS overlapped with 1332, 1177, 3042, 1288, 24 QTLs, respectively. But there are also very few QTLs about body conformation traits overlapped with significant SNPs.

# Discussion

## Comparison with the other GWAS studies

In this study, FarmCPU was applied for screening the QTLs related to the milk production traits, health traits, and body conformation traits. A total of 95 significant SNPs were detected, located on the 93 candidate genes. Of these genes, *EHHADH*, *SLC37A1*, *PLEKHA5*, *TONSL*, *PLEC*, and *IL5RA* were reportedly related to milk production traits in other studies (15, 20, 21, 24, 39). However, this study did not detect some important

TABLE 3 The SNPs and candidate genes with pleiotropic effect in this study.

| SNP name | Traits | Gene | Distance(kb)* | Gene full name | Gene function |
|---|---|---|---|---|---|
| BovineHD1400000287 | MY, FP, PP | PLEC | Intron | Plectin | Related to the MY, FP, and PP traits in Chinese Holsteins (20). |
| BovineHD0500025853 | FP, FY | PLEKHA5 | Intron | Pleckstrin homology domain containing, family A member 5 | Significantly associated with FP (21). |
| BovineHD1400011649 | MY, PY | HNF4G | Intron | Hepatocyte nuclear factor 4 gamma | Associated with childhood obesity (22). Key regulators of beef cattle carcass IMF (23). |
| BovineHD1400000206 | FY, SCS | TONSL | 1.65 (U) | Tonsoku like, DNA repair protein | Related to milk yield (24, 25) and affect the gamma–linolenic acid, long–chain saturated fatty acids and milk fat percent of the Canadian Holstein cows (26). |
| BovineHD4100004660 | DCCI, FTLEG | LCORL | 665.01 (D) | Ligand dependent nuclear receptor corepressor like | Affect human height (27), pig body length (28), horse height (29), chicken carcass weight (30), and the growth and development of cattle (31). Associated with the human skeletal frame size (32). |
| BovineHD1300012605 | BFCI, FTLEG | PYGB | 0.61 (D) | Glycogen phosphorylase B | Inhibition of glycogen utilization (33) |
| BTA−50244–no–rs | BFCI, DCCI | PTGER4 | 541.63 (D) | Prostaglandin E receptor 4 | Relaxation to the smooth muscle (34), leading to the phosphorylation of glycogen synthase kinase−3 (35), involving in osteoporosis (36), and regulating lipid droplet size and mitochondrial activity in the white adipose tissue (37, 38). |

*U, Upstream; D, Downstream.

candidate genes, such as *DGAT1*. Because in this study, the closest SNP on both flanks of *DGAT1* are BovineHD1400000206 (109.2 kb) and ARS-BFGL-NGS-55227 (50.8 Mb), respectively. Of these, BovineHD1400000206 associated with fat yield (*P* value = 2.76E−17). But the nearest gene on this significantly SNP is the TONSL gene (1.65 kb), which is a neighboring gene to *DGAT1*. So, the *DGAT1* gene was not detected in this study. The study by Ning et al. (40) used two models and a 70k SNP chip based on the Chinese Holsteins population and identified the *DGAT1* gene to be related to milk (40). Kim et al. (41) also obtained *DGAT1* affecting MY and FY in the Korean cattle population (41). Cole et al. (42) identified the *PHKA2* gene to be highly significant for four body size traits (stature, strength, body depth, rump width) (42). The 770k BeadChip was used by An et al. (43) to identify five candidate genes (*CSMD3*, *LAP3, SYN3, FAM19A5,* and *TIMP3*) related to the body conformation traits. This study did not detect the above genes to be associated with body conformation traits.

These inconsistencies might be due to differences in the detection platforms or algorithms used in the corresponding analysis, changes in the genetic background of the analyzed cattle, differences in the size and structure of the study population, or random or technical errors in some analyses. This also indicated that there are many important genetic markers or candidate genes in the bovine genome that are yet to be discovered.

## Genetic analysis of pleiotropic genes

Organisms have hundreds of thousands of genes and tens of thousands of phenotypes. The relationship between genes and epigenetic factors is complex. There are various associations such as pleiotropism, multigenic effect, polygene effect and so on. Pleiotropy is defined as the phenomenon where a single locus affects two or more distinct phenotypic traits (44, 45). It

TABLE 4 Genome-wide significant SNPs are associated with milk production traits.

| Traits | SNP name | BTA | Postion (Mb) | MAF | Nearest gene | Distance (kb)* | *P*–value | SNP effect |
|---|---|---|---|---|---|---|---|---|
| MY | BTB−00088434 | 2 | 33.86 | 0.0599 | KCNH7 | 39.78 (U) | 5.87E−08 | −295.1002 |
| MY | Hapmap40999–BTA−47831 | 2 | 62.78 | 0.4824 | TMEM163 | 0.81 (D) | 2.01E−07 | −131.0532 |
| MY | 5–82810184–C–T–rs110495697 | 5 | 82.40 | 0.4156 | ARNTL2 | Intron | 1.30E−13 | 206.5174 |
| MY | BovineHD0600027996 | 6 | 98.51 | 0.2000 | GPAT3 | 144.85(D) | 5.23E−09 | −188.5885 |
| MY | BovineHD0800021118 | 8 | 69.70 | 0.2790 | SLC39A14 | Intron | 5.53E−07 | −145.3961 |
| MY | BovineHD1400000287 | 14 | 0.88 | 0.2240 | PLEC | Intron | 2.71E−07 | 163.9574 |
| MY | BovineHD1400011649 | 14 | 38.57 | 0.1725 | HNF4G | Intron | 4.36E−07 | −154.3554 |
| MY | BTA−07375–no–rs | 14 | 66.16 | 0.3824 | ERICH5 | Intron | 8.97E−10 | 194.9580 |
| MY | BovineHD1500014407 | 15 | 49.22 | 0.3053 | OR51L4 | 0.17(U) | 1.23E−07 | −155.6556 |
| MY | BovineHD1700012968 | 17 | 45.45 | 0.2130 | SFSWAP | 90.47(D) | 5.35E−07 | −153.0319 |
| MY | Hapmap55097–rs29010952 | 18 | 26.88 | 0.4260 | GOT2 | 436.24(D) | 1.95E−07 | 136.2285 |
| MY | BovineHD2800000275 | 28 | 1.65 | 0.2076 | URB2 | 45.14(D) | 2.75E−07 | −164.7354 |
| FY | Hapmap24838–BTA−143176 | 5 | 62.15 | 0.4137 | TMPO | 560.66(U) | 3.37E−08 | 3.4575 |
| FY | BovineHD0500025853 | 5 | 90.66 | 0.3145 | PLEKHA5 | Intron | 1.91E−07 | 3.4453 |
| FY | ARS–BFGL–NGS−10921 | 7 | 14.97 | 0.3893 | PDE4A | Intron | 1.80E−07 | 3.3773 |
| FY | BovineHD0700020203 | 7 | 67.04 | 0.3893 | SGCD | 536.03(U) | 6.02E−09 | −3.6651 |
| FY | ARS–BFGL–NGS−75350 | 11 | 49.90 | 0.1893 | TCF7L1 | 19.24(D) | 2.79E−07 | −3.8169 |
| FY | ARS–BFGL–NGS−29737 | 12 | 47.06 | 0.1004 | CXCL14 | 11.84(U) | 5.06E−09 | 6.6175 |
| FY | BovineHD1400000206 | 14 | 0.49 | 0.2172 | TONSL | 1.65(U) | 2.76E−17 | 6.3376 |
| FY | ARS–BFGL–NGS−55227 | 14 | 5.69 | 0.4179 | KHDRBS3 | 711.31(U) | 6.01E−09 | 3.7919 |
| FY | ARS–BFGL–NGS−115233 | 20 | 58.83 | 0.4576 | TRIO | Intron | 2.29E−09 | 3.6663 |
| FY | BovineHD2300011633 | 23 | 40.58 | 0.1126 | ZNF496 | Intron | 7.11E−09 | −5.8266 |
| FY | ARS–BFGL–NGS−95089 | 27 | 8.94 | 0.2202 | AGA | 238.44(D) | 4.30E−07 | −3.7871 |
| FP | BTB−00035766 | 1 | 81.92 | 0.3565 | EHHADH | Intron | 3.93E−07 | 0.0443 |
| FP | Hapmap40546–BTA−48622 | 2 | 104.36 | 0.4168 | MARCHF4 | Intron | 7.92E−08 | −0.0460 |
| FP | BovineHD0400011775 | 4 | 42.79 | 0.4855 | OR2AV14 | 10.61(U) | 4.24E−07 | −0.0432 |
| FP | BovineHD0500025853 | 5 | 90.66 | 0.3145 | PLEKHA5 | Intron | 8.76E−11 | 0.0609 |
| FP | Hapmap35196–BES10_Contig207_566 | 6 | 90.84 | 0.0679 | SDAD1 | Intron | 1.73E−07 | 0.0917 |
| FP | BovineHD0800011917 | 8 | 39.97 | 0.0832 | SLC1A1 | Intron | 5.36E−07 | 0.0705 |
| FP | ARS–BFGL–NGS−15823 | 9 | 28.53 | 0.3385 | PKIB | Intron | 2.70E−11 | 0.0611 |
| FP | BovineHD1100015676 | 11 | 53.86 | 0.4229 | – | – | 1.28E−09 | −0.0479 |
| FP | BovineHD1400000287 | 14 | 0.88 | 0.2240 | PLEC | Intron | 6.48E−37 | 0.1423 |
| FP | UA–IFASA−7269 | 14 | 3.10 | 0.2313 | AGO2 | Intron | 2.01E−15 | −0.1058 |
| FP | BovineHD1500015438 | 15 | 52.73 | 0.1485 | P2RY2 | 0.64(D) | 4.08E−08 | −0.0687 |
| FP | BovineHD1500017563 | 15 | 60.50 | 0.4202 | KCNA4 | 230.62(U) | 5.11E−07 | 0.04461 |
| FP | ARS–BFGL–BAC−27930 | 20 | 29.36 | 0.2080 | DDX6 | Intron | 8.92E−11 | −0.0795 |
| FP | BTA−50420–no–rs | 20 | 36.05 | 0.2107 | EGFLAM | Intron | 1.32E−11 | −0.0872 |
| FP | BTB−01263010 | 20 | 42.72 | 0.1565 | CDH6 | 363.33(D) | 2.61E−08 | 0.0769 |
| PP | BovineHD0100013692 | 1 | 48.12 | 0.2420 | – | – | 8.70E−08 | 0.0222 |
| PP | BovineHD0100041607 | 1 | 142.82 | 0.3676 | SLC37A1 | Intron | 9.80E−11 | 0.0252 |
| PP | ARS–BFGL–NGS−80635 | 2 | 31.59 | 0.4523 | COBLL1 | 6.23(D) | 1.02E−07 | 0.0182 |
| PP | ARS–BFGL–NGS−117881 | 5 | 82.23 | 0.3859 | C5H12orf71 | 21.60(U) | 1.11E−13 | −0.0318 |
| PP | BovineHD0600008707 | 6 | 29.63 | 0.3897 | BMPR1B | 36.41(D) | 1.62E−08 | −0.0220 |
| PP | ARS–BFGL–BAC−15734 | 13 | 48.97 | 0.3725 | BMP2 | 179.62(U) | 5.84E−08 | −0.0210 |
| PP | BovineHD1400000287 | 14 | 0.88 | 0.2240 | PLEC | Intron | 1.27E−16 | 0.0362 |
| PP | BovineHD1400013724 | 14 | 46.18 | 0.4492 | EXT1 | Intron | 3.55E−08 | 0.0211 |
| PP | chr14_57250692 | 14 | 55.09 | 0.1481 | NUDCD1 | Intron | 3.79E−08 | −0.0283 |
| PP | ARS–BFGL–NGS−21921 | 19 | 14.40 | 0.2817 | CCL14 | 59.58(U) | 1.97E−07 | −0.0197 |

*(Continued)*

TABLE 4 (Continued)

| Traits | SNP name | BTA | Postion (Mb) | MAF | Nearest gene | Distance(kb)* | *P*–value | SNP effect |
|---|---|---|---|---|---|---|---|---|
| PP | BovineHD2000009361 | 20 | 32.69 | 0.2844 | OXCT1 | Intron | 5.35E−08 | −0.0294 |
| PP | BovineHD2500004479 | 25 | 15.71 | 0.1275 | XYLT1 | Intron | 1.78E−10 | 0.0365 |
| PY | BovineHD0100027261 | 1 | 95.18 | 0.3279 | FNDC3B | Intron | 1.20E−11 | 3.3681 |
| PY | BovineHD0300017107 | 3 | 56.64 | 0.3195 | LMO4 | 54.14(U) | 1.26E−07 | −2.5817 |
| PY | Hapmap26317–BTC−059618 | 6 | 80.53 | 0.1557 | EPHA5 | 306.96(U) | 4.69E−08 | −3.2564 |
| PY | ARS–BFGL–NGS−4974 | 11 | 106.56 | 0.1069 | ZMYND19 | 5.22(D) | 4.52E−09 | −4.2183 |
| PY | BovineHD1400011649 | 14 | 38.57 | 0.1725 | CRISPLD1 | 226.89(D) | 4.93E−08 | 3.2889 |
| PY | BovineHD1700016449 | 17 | 55.89 | 0.4740 | CCDC60 | Intron | 1.62E−07 | −2.2566 |
| PY | ARS–USMARC–Parent–EF034086–no–rs | 26 | 37.90 | 0.4607 | EMX2 | 63.30(D) | 4.29E−09 | 2.6483 |

*U, Upstream; D, Downstream.

TABLE 5 Genome−wide significant SNPs are associated with somatic cell score.

| Traits | SNP name | BTA | Postion (Mb) | MAF | Nearest gene | Distance(kb) * | P value | SNP effect |
|---|---|---|---|---|---|---|---|---|
| SCS | Hapmap59481–rs29019616 | 1 | 56.94 | 0.1893 | GCSAM | Intron | 1.59E−10 | −0.1547 |
| SCS | BovineHD0200033155 | 2 | 113.94 | 0.2847 | NYAP2 | 90.45(D) | 1.98E−08 | 0.1198 |
| SCS | BovineHD0600020300 | 6 | 71.35 | 0.3580 | CEP135 | 6.10(U) | 1.94E−10 | −0.1304 |
| SCS | BovineHD1100011547 | 11 | 39.19 | 0.2080 | CCDC85A | 118.86(D) | 3.15E−08 | 0.1376 |
| SCS | BovineHD1300019252 | 13 | 67.15 | 0.2282 | KIAA1755 | 21.69(D) | 1.73E−09 | −0.1423 |
| SCS | BovineHD1400000206 | 14 | 0.49 | 0.2172 | TONSL | 1.65(U) | 4.03E−07 | −0.1159 |
| SCS | BovineHD1400011508 | 14 | 38.01 | 0.2939 | PI15 | 147.13(U) | 3.10E−09 | −0.1307 |
| SCS | BovineHD1600013229 | 16 | 47.05 | 0.3996 | ACOT7 | Intron | 7.12E−08 | 0.1065 |
| SCS | BovineHD1600015783 | 16 | 55.27 | 0.0657 | SERPINC1 | Intron | 2.54E−08 | −0.2208 |
| SCS | BTA−65815–no–rs | 16 | 59.73 | 0.2267 | RASAL2 | Intron | 2.31E−10 | −0.1528 |
| SCS | UA–IFASA−5305 | 19 | 59.21 | 0.1271 | SOX9 | 289.75(D) | 3.61E−09 | −0.1751 |
| SCS | BovineHD2000017315 | 20 | 61.61 | 0.4405 | CTNND2 | 11.32(U) | 1.18E−07 | 0.1036 |
| SCS | BTA−52343–no–rs | 21 | 42.73 | 0.1042 | AKAP6 | Intron | 4.65E−09 | −0.1536 |
| SCS | Hapmap46118–BTA−108252 | 22 | 19.45 | 0.4435 | GRM7 | Intron | 3.00E−08 | −0.1121 |
| SCS | ARS–BFGL–NGS−24519 | 25 | 10.59 | 0.1378 | GSPT1 | 0.93(D) | 2.31E−07 | 0.1394 |
| SCS | ARS–BFGL–NGS−37189 | 25 | 32.40 | 0.07786 | RCC1L | 267.69(U) | 3.28E−08 | 0.1948 |
| SCS | Hapmap42542–BTA−40776 | 26 | 27.93 | 0.2504 | SORCS1 | 20.95(D) | 5.60E−10 | 0.1395 |

*U, Upstream; D, Downstream.

is common in nature. For example, the *DGAT1* gene is related to milk yield (40) and fat yield (26, 41). The genes *PIK3R6* and *PIK3R1* showed direct functional associations with height and body size (10). Production and health constitute fundamental dairy functions while body conformation traits are related to the functionality of the cow's body. So, the milk production traits and body conformation traits of dairy cows tend to complement each other. Certain identified regions related to conformation traits overlap with the performance traits such as reproduction (46), and milk production (47). Some genes in these regions were also involved in regulating the cell cycle or cell division, homeostasis, and lipid metabolism (10).

This study also reported this interesting phenomenon where the *PLEC, PLEKHA5,* and *TONSL* genes were found to belong to the pleiotropism gene for milk traits, and the *LCORL*, and *PTGER4* were pleiotropic genes for the body conformation traits. The *PLEC* gene (Plectin) can interlink different elements of the cytoskeleton. The *PLEC* gene was found to be associated with multiple traits, like MY, FP, and PP. Dan Wang et al. (20) also detected *PLEC* to have potential effects on the MY, FP, and PP traits, which could be useful for molecular breeding for milk production in Chinese Holsteins. The *PLEKHA5* gene, located on BTA 5, was predicted to enable the activity of binding phosphatidylinositol phosphate (48). Jiang et al. (21) showed the *PLEKHA5* gene to be significantly associated with FP using two different methods using 294,079 Holstein cows. The TONSL protein was considered to be an NF-κ negative regulator of B mediated transcription. Peters et al. (24), Nayeri et al. (25), and Atashi et al. (49) found this gene to be related to milk yield and the *TONSL* gene was found to reportedly

TABLE 6 Genome−wide significant SNPs are associated with body conformation traits.

| Traits | SNP name | BTA | Position (Mb) | MAF | Nearest gene | Distance (kb)* | *P*-value | SNP effect |
|---|---|---|---|---|---|---|---|---|
| BFCI | ARS–BFGL–NGS−39319 | 8 | 31.33 | 0.3836 | MPDZ | 122.75(D) | 4.59E−08 | −1.3217 |
| BFCI | BovineHD1000015574 | 10 | 52.01 | 0.3450 | AQP9 | 75.39(U) | 3.07E−09 | 1.5722 |
| BFCI | BovineHD1200008803 | 12 | 29.84 | 0.1481 | HSPH1 | 19.83(U) | 3.72E−08 | 1.9433 |
| BFCI | BovineHD1300012605 | 13 | 42.81 | 0.4622 | PYGB | 0.61(D) | 8.06E−08 | 1.2185 |
| BFCI | ARS–BFGL–NGS−66252 | 16 | 50.24 | 0.0805 | MMEL1 | 31.03(U) | 2.79E−08 | 2.4780 |
| BFCI | BovineHD1600023101 | 16 | 77.36 | 0.4538 | ATP6V1G3 | 47.95(U) | 9.44E−08 | 1.1905 |
| BFCI | BovineHD1700005623 | 17 | 19.11 | 0.4050 | SLC7A11 | 307.12(U) | 2.82E−08 | 1.4169 |
| BFCI | BovineHD1900015024 | 19 | 53.08 | 0.3546 | RBFOX3 | Intron | 4.39E−07 | 1.2520 |
| BFCI | BTA−50244–no–rs | 20 | 34.30 | 0.3710 | PTGER4 | 541.63(D) | 5.84E−13 | −1.9115 |
| BFCI | BovineHD2200000513 | 22 | 1.99 | 0.1302 | EOMES | 123.32(U) | 2.09E−07 | 1.8111 |
| DCCI | BovineHD0300021562 | 3 | 73.79 | 0.4351 | NEGR1 | Intron | 1.03E−07 | −1.2304 |
| DCCI | BTB−00190417 | 4 | 59.09 | 0.3496 | DNAJB9 | 493.74(U) | 4.93E−11 | 1.6662 |
| DCCI | BovineHD4100004660 | 6 | 38.22 | 0.4271 | LCORL | 665.01(D) | 2.39E−09 | −1.4509 |
| DCCI | ARS–BFGL–BAC−15023 | 12 | 31.34 | 0.4103 | MTUS2 | Intron | 6.74E−08 | 1.2696 |
| DCCI | BTB−00597065 | 15 | 41.00 | 0.3527 | GALNT18 | 64.26(U) | 9.91E−10 | 1.5227 |
| DCCI | BTA−50244–no–rs | 20 | 34.30 | 0.3710 | PTGER4 | 541.63(D) | 6.11E−08 | −1.2173 |
| DCCI | ARS–BFGL–NGS−97747 | 23 | 28.02 | 0.3840 | CDSN | 4.46(U) | 1.77E−09 | 1.4608 |
| FTLEG | BovineHD0100020157 | 1 | 69.85 | 0.0962 | SNX4 | 37.98(U) | 2.04E−07 | −2.4836 |
| FTLEG | ARS–BFGL–NGS−56584 | 1 | 145.09 | 0.1309 | POFUT2 | Intron | 7.56E−08 | 2.0359 |
| FTLEG | BovineHD0300019080 | 3 | 63.66 | 0.1248 | ADGRL2 | 511.07(D) | 1.06E−08 | 2.5317 |
| FTLEG | BTB−01326707 | 6 | 38.00 | 0.2737 | LCORL | 665.01(D) | 3.16E−11 | −2.0018 |
| FTLEG | BTB−00124923 | 9 | 34.94 | 0.1851 | FRK | 243.36(D) | 3.42E−07 | 1.7253 |
| FTLEG | BovineHD1300012605 | 13 | 42.81 | 0.4622 | PYGB | 0.61(D) | 2.23E−09 | 1.6065 |
| FTLEG | Hapmap50322–BTA−34017 | 13 | 78.20 | 0.1309 | CEBPB | 7.25(U) | 8.11E−08 | −2.2401 |
| FTLEG | BovineHD1600000840 | 16 | 3.12 | 0.1191 | KLHDC8A | 11.96(D) | 3.74E−07 | 2.1807 |
| FTLEG | BovineHD1600008381 | 16 | 28.91 | 0.1683 | TMEM63A | 1.14(D) | 7.79E−09 | −2.1564 |
| FTLEG | BovineHD2000011811 | 20 | 41.04 | 0.3221 | SUB1 | 26.79(U) | 4.00E−11 | −1.9453 |
| FTLEG | BTA−14388–rs29023151 | 22 | 23.20 | 0.4561 | IL5RA | Intron | 8.59E−10 | 1.6383 |

*U, Upstream; D, Downstream.

affect the gamma-linolenic acid, long-chain saturated fatty acids and milk fat percent of the Canadian Holstein cows (26). Interesting, the *TONSL* gene is a neighboring gene to *DGAT1* (flanking < 200 kb), associated with the fat percentage of milk (26).

Some studies on the *LCORL* gene showed it to affect human height (27), pig body length (28), horse height (29), chicken carcass weight (30), and the growth and development of cattle (31). This gene might have been a novel loci associated with the human skeletal frame size (32). *PTGER4* encodes a protein that is one of the members of the G-protein coupled receptor family, which imparts relaxation to the smooth muscle (34), leading to the phosphorylation of glycogen synthase kinase-3 (35), involved in osteoporosis (36), and regulating lipid droplet size and mitochondrial activity in the white adipose tissue (37, 38).

## Important candidate genes related to the fat metabolism or mammary gland development

Fatty acids are essential components of milk with known positive associations with human cardiovascular diseases and so on. This study identified genes such as *GPAT3*, *ARNTL2*, *EHHADH*, *CEBPB*, *DNAJB9*, *ZNF496*, *AGO2*, *GALNT18*, and *NEGR1 as* critical for obesity traits or adipose metabolism (see Table 7).

*GPAT3* is highly expressed in the adipose tissue with an important role in adipogenesis (50). This gene can be regulated by folic acid for controlling lactation and metabolic function of the dairy cows (51) and is also involved in fat and lipid metabolism in the Yunling cattle (52). *EHHADH* involved in fatty acid oxidation is essential for producing medium-chain

TABLE 7 Important candidate genes related to the fat metabolism or mammary gland development.

| Gene name | Location (BTA:Start–End, Mb) | Full name | Gene function |
|---|---|---|---|
| GPAT3 | 6:98.29–98.36 | Glycerol—3–phosphate acyltransferase 3 | Highly expressed in the adipose tissue with an important role in adipogenesis (50). Can be regulated by folic acid for controlling lactation and metabolic function of the dairy cows (51). Involved in fat and lipid metabolism in the Yunling cattle (52). |
| ARNTL2 | 5:82.47–82.55 | Aryl hydrocarbon receptor nuclear translocator like 2 | Influencing Mexican–Mestizo childhood obesity (53). |
| EHHADH | 1:81.88–81.93 | Enoyl–CoA hyd ratase and 3–hydroxyacyl CoA dehydrogenase | Involved in fatty acid oxidation is essential for producing medium–chain dicarboxylic acids (54). Impact on the characteristics of milk fatty acid traits in Chinese Holstein (55). A pivotal gene in the fat–related pathway (56). |
| CEBPB | 13:78.20–78.21 | CCAAT enhancer binding protein beta | Involved in regulating the expression of fatty acid synthase in dairy cow mammary epithelial cells and milk fat synthesis (57). |
| DNAJB9 | 4:59.58–59.59 | DnaJ heat shock protein family (Hsp40) member B9 | The prognostic biomarkers of breast cancer (58). Correlated with the abdominal fat weight (59). |
| ZNF496 | 7:40.57–40.61 | Zinc finger protein 496 | Associated with milk fat and fertility (60). |
| AGO2 | 14:3.06–3.14 | Argonaute RISC catalytic component 2 | Related to mitochondrial oxidation and obesity–associated pathophysiology (61). |
| GALNT18 | 15:41.06–41.42 | Polypeptide N–acetylgalactosaminyltransferase 18 | Associated with milk protein and fat traits (62). |
| NEGR1 | 3:72.81–73.84 | Neuronal growth regulator 1 | Associated with obesity and BMI (body mass index) (63–65). |
| SLC37A1 | 1:142.81–142.87 | Solute carrier family 37 member 1 | Over–expressed in the bovine mammary tissue (66). Increases milk yield, decreases phosphorus concentration (66). |
| FNDC3B | 1:95.12–95.41 | Fibronectin type III domain containing 3B | Biomarker for the bovine mammary stem/progenitor cells, and Essential for the growth and maintenance of the mammary epithelium (67). |

U, Upstream; D, Downstream.

dicarboxylic acids (54). Hence, this gene has a key impact on the characteristics of milk fatty acid traits in Chinese Holstein (55). In porcine adipogenesis, *EHHADH* has been proposed to be a pivotal gene in the fat-related pathway (56). The *DNAJB9* gene is reportedly one of the prognostic biomarkers of breast cancer (58). Interestingly, DNAJB9 and DNAJB6 are members of the DNAJ gene family, with sequence similarity. The expression level of DNAJB6 in the chicken abdominal adipose tissue was significantly negatively correlated with the abdominal fat weight (59). *ZNF496* is reportedly associated with milk concentration (milk fat) and fertility (60). According to Gao et al. (62), the *GALNT18* gene was associated with milk protein and fat traits.

According to the known gene functions, some candidate genes were expressed in the mammary gland, such as the *SLC37A1*, and *FNDC3B* genes (see Table 7). *SLC37A1*, over-expressed in the bovine mammary tissue relative to the 17 other tissue types (66) transports glucose-6-phosphate in one direction and phosphorus in the other (68). Glucose is known to be essential for lactose synthesis in mammary cells. Kemper et al. (66) identified the causative mutation increasing the expression of SLC37A1 leading to an increase in milk yield and decreasing the phosphorus concentration.

## QTLs result overlapped with GWAS

Although many quantitative trait loci (QTLs) related to economically important traits in dairy cows have been identified, due to insufficient sample size and insufficient marker density used in QTL mapping research in history, not all genetic variations of these traits have been captured (69), in the study, we used GWAS to analyze the milk production traits, body conformation traits and somatic cells of dairy cows, and most of the results were also verified in the QTL analysis of dairy cows. Interestingly, our study found many SNP related to pleiotropy, but no repeated QTL regions were found in the QTL analysis (70). Also found the same phenomenon in the study of multiple traits of beef cattle. With these results, we can get some inspiration in verifying QTLs of some characteristics of interest shared among varieties (71).

## Conclusions

A total of 95 significant SNPs were identified to be related to the milk production, somatic cell score, and body conformation traits in Holstein cattle. Among them, 7 significant SNPs located on the *PLEC, PLEKHA5, TONSL, PTGER4,* and *LCORL* genes showed pleiotropic effects on milk production or body conformation traits. In addition, some important candidate genes, including *GPAT3, CEBPB, AGO2, SLC37A1,* and *FNDC3B,* were also found to be related to the fat metabolism or involved in mammary gland development. The above genes however need to be consolidated as new potential genes through future validation.

## Data availability statement

The original contributions presented in the study are included in the article or supplementary material, the variation data reported in this article have been deposited in the Genome Variation Map (GVM) in Big Data Center, Beijing Institute of Genomics (BIG), and Chinese Academy of Sciences, under accession numbers GVM000388 that are publicly accessible at https://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000388. The Bioproject accession number is PRJCA011726. Further inquiries can be directed to the corresponding author.

## Ethics statement

The animal study was reviewed and approved by Animal Care Advisory Committee, Northeast Agricultural University (Harbin, China).

## Author contributions

ZW, PW, and XL conceived the study and participated in its design. YZ, JW, QK, and XN were involved in the acquisition of data. XL, JW and QZ performed all data analysis. XL and ZW drafted the manuscript. ZW, PW, XL, YZ, CZ, QK, and XN contributed to the writing and editing. All authors read and approved the final manuscript.

## Funding

This work was financially supported by the Natural Science Foundation of China (No. 32070571), the Academic Backbone Project of Northeast Agricultural University (No. 15XG14), and the NEAU Research Founding for Excellent Young Teachers (2010RCB29).

## Acknowledgments

The authors are grateful to all the members of the Dairy Breeding Group of the Heilongjiang Provincial Husbandry Department for collecting the data. We appreciate the support by the Project of Science and Education Office in the Department of Agriculture and Rural Affairs of Heilongjiang Province (Application of the genomic breeding technology for modernized large-scale dairy farm in Heilongjiang Province).

## 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.

## References

1. Chapinal N, Koeck A, Sewalem A,Kelton DF, Mason S, Cramer G, et al. Genetic parameters for hoof lesions and their relationship with feet and leg traits in Canadian Holstein cows. *J Dairy Sci.* (2013) 96:2596–04. doi: 10.3168/jds.2012-6071

2. Lund T, Miglior F, Dekkers JCM, Burnside EB. Genetic-relationships between clinical mastitis, somatic-cell count, and udder conformation in Danish Holsteins. *Livest Prod Sci.* (1994) 39:243–51. doi: 10.1016/0301-6226(94)90203-8

3. Vollema AR, Van Der Beek S, Harbers AG, De Jong G. Genetic evaluation for longevity of Dutch dairy bulls. *J Dairy Sci.* (2000):2629–39. doi: 10.3168/jds.S0022-0302(00)75156-3

4. Kadarmideen HN, Wegmann S. Genetic parameters for body condition score and its relationship with type and production traits in Swiss Holsteins. *J Dairy Sci.* (2003) 86:3685–93. doi: 10.3168/jds.S0022-0302(03)73974-5

5. Short TH, Lawlor TJ. Genetic parameters of conformation traits, milk yield, and herd life in Holsteins. *J Dairy Sci.* (1992) 75:1987–8. doi: 10.3168/jds.S0022-0302(92)77958-2

6. Schennink A, Stoop WM, Visker MH, van der Poel JJ, Bovenhuis H, van Arendonk JA. Short communication: genome-wide scan for bovine milk-fat composition. II Quantitative trait loci for long-chain fatty acids. *J Dairy Sci.* (2009) 92:4676–82. doi: 10.3168/jds.2008-1965

7. Kiser JN, Cornmesser MA, Gavin K, Hoffman A, Moore DA, Neibergs HL. Rapid communication: genome-wide association analyses identify loci associated with colostrum production in Jersey cattle1. *J Anim Sci.* (2019) 97:1117–23. doi: 10.1093/jas/sky482

8. Zhou J, Liu L, Chen CJ, Zhang M, Lu X, Zhang Z, et al. Genome-wide association study of milk and reproductive traits in dual-purpose Xinjiang Brown cattle. *BMC Genomics.* (2019) 20:827. doi: 10.1186/s12864-019-6224-x

9. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet.* (2018) 50:362–7. doi: 10.1038/s41588-018-0056-5

10. Vanvanhossou SFU, Scheper C, Dossa LH, Yin T, Brügemann K, König S, et al. Multi-breed GWAS for morphometric traits in four Beninese indigenous cattle breeds reveals loci associated with conformation, carcass and adaptive traits. *BMC Genomics.* (2020) 21:783. doi: 10.1186/s12864-020-07170-0

11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* (2007) 81:559–75. doi: 10.1086/519795

12. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet*. (2008) 40:491–92.

13. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet*. (2016) 12:e1005767. doi: 10.1371/journal.pgen.1005767

14. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S. Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. (2010) 93:743–52. doi: 10.3168/jds.2009-2730

15. Iung LHS, Petrini J, Ramírez-Díaz J, Salvian M, Rovadoscki GA, Pilonetto F,, et al. Genome-wide association study for milk production traits in a Brazilian Holstein population. *J Dairy Sci.* (2019) 102:5305–14. doi: 10.3168/jds.2018-14811

16. Weng Z, Zhang Z. Ding X,Fu W, Ma P, Wang C, et al. Application of imputation methods to genomic selection in Chinese Holstein cattle. *J Anim Sci Biotechnol.* (2012) 3:6. doi: 10.1186/2049-1891-3-6

17. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. (2008) 91:4414–23. doi: 10.3168/jds.2007-0980

18. Zhao H, Nettleton D, Soller M, Dekkers JC. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res*. (2005) 86:77–87. doi: 10.1017/S001667230500769X

19. Wu C, DeWan A, Hoh J, Wang Z. A comparison of association methods correcting for population stratification in case-control studies. *Ann Hum Genet*. (2011)75:418–27. doi: 10.1111/j.1469-1809.2010.00639.x

20. Wang D, Ning C, Liu JF, Zhang Q, Jiang L. Short communication: replication of genome-wide association studies for milk production traits in Chinese Holstein by an efficient rotated linear mixed model. *J Dairy Sci.* (2019) 102:2378–83. doi: 10.3168/jds.2018-15298

21. VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in US Holstein Cattle. *Front Genet.* (2019) 10:412. doi: 10.3389/fgene.2019.00412

22. Selvanayagam T, Walker S, Gazzellone MJ, Kellam B, Cytrynbaum C, Stavropoulos DJ, et al. Genome-wide copy number variation analysis identifies novel candidate loci associated with pediatric obesity. *Eur J Hum Genet.* (2018) 26:1588–96. doi: 10.1038/s41431-018-0189-0

23. Ramayo-Caldas Y, Fortes MR, Hudson NJ, Porto-Neto LR, Bolormaa S. Barendse W,et al. A marker-derived gene network reveals the regulatory role of PPARGC1A, HNF4G, and FOXP3 in intramuscular fat deposition of beef cattle. *J Anim Sci.* (2014) 92:2832–45. doi: 10.2527/jas.2013-7484

24. Peters SO, Kizilkaya K, Ibeagha-Awemu EM, Sinecen M, Zhao X. Comparative accuracies of genetic values predicted for economically important milk traits, genome-wide association, and linkage disequilibrium patterns of Canadian Holstein cows. *J Dairy Sci.* (2021) 104:1900–16. doi: 10.3168/jds.2020-18489

25. Nayeri S, Sargolzaei M, Abo-Ismail MK, May N, Miller SP, Schenkel F, et al. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet.* (2016) 17:75. doi: 10.1186/s12863-016-0386-1

26. Ibeagha-Awemu EM, Peters SO, Akwanji KA, Imumorin IG, Zhao X. High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Sci Rep.* (2016) 6:31109. doi: 10.1038/srep31109

27. Sovio U, Bennett AJ, Millwood IY, Molitor J, O'Reilly PF, Timpson NJ, et al. Genetic determinants of height growth assessed longitudinally from infancy to adulthood in the northern Finland birth cohort 1966. *PLoS Genet.* (2009) 5:e1000409. doi: 10.1371/journal.pgen.1000409

28. Rubin CJ, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci U S A*. (2012) 109:19529–36. doi: 10.1073/pnas.1217149109

29. Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, et al. A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS ONE.* (2012) 7:e37282. doi: 10.1371/journal.pone.0037282

30. Liu R, Sun Y, Zhao G, Wang F, Wu D, Zheng M, et al. Genome-wide association study identifies Loci and candidate genes for body composition and meat quality traits in Beijing-You chickens. *PLoS ONE.* (2013) 8:e61172. doi: 10.1371/journal.pone.0061172

31. Weikard R, Altmaier E, Suhre K, Weinberger KM, Hammon HM, Albrecht E, et al. Metabolomic profiles indicate distinct physiological pathways affected by two loci with major divergent effect on Bos taurus growth and lipid deposition. *Physiol Genomics*. (2010) 42A:79–88. doi: 10.1152/physiolgenomics.00120.2010

32. Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, Hammond N, et al. Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genet*. (2009) 5:e1000445. doi: 10.1371/journal.pgen.1000445

33. Altemus MA, Goo LE, Little AC, Yates JA, Cheriyan HG, Wu ZF, et al. Breast cancers utilize hypoxic glycogen stores via PYGB, the brain isoform of glycogen phosphorylase, to promote metastatic phenotypes. *PLoS ONE.* (2019) 14:e0220973. doi: 10.1371/journal.pone.0220973

34. Tilley SL, Hartney JM, Erikson CJ, Jania C, Nguyen M, Stock J, et al. Receptors and pathways mediating the effects of prostaglandin E2 on airway tone. *Am J Physiol Lung Cell Mol Physiol.* (2003) 284:L599–606.

35. Fujino H, West KA, Regan JW. Phosphorylation of glycogen synthase kinase-3 and stimulation of T-cell factor signaling following activation of EP2 and EP4 prostanoid receptors by prostaglandin E2. *J Biol Chem.* (2002) 277:2614–9. doi: 10.1074/jbc.M109440200

36. Li M, Healy DR, Li Y, Simmons HA, Crawford DT, Ke HZ, et al. Osteopenia and impaired fracture healing in aged EP4 receptor knockout mice. *Bone*. (2005) 37:46–54. doi: 10.1016/j.bone.2005.03.016

37. Ying F, Cai Y, Cai Y, Wang Y, Tang EHC. Prostaglandin E receptor subtype 4 regulates lipid droplet size and mitochondrial activity in murine subcutaneous white adipose tissue. *FASEB J.* (2017) 31:4023–36. doi: 10.1096/fj.201700191R

38. Lv X, Gao F, Li TP, Xue P, Wang X, Wan M, et al. Skeleton interoception regulates bone and fat metabolism through hypothalamic neuroendocrine NPY. *Elife.* (2021) 10:e70324. doi: 10.7554/eLife.70324

39. Zhou C, Li C, Cai W, Liu S, Yin H, Shi S, et al. Genome-wide association study for milk protein composition traits in a Chinese Holstein Population using a single-step approach. *Front Genet*. (2019) 10:72. doi: 10.3389/fgene.2019.00072

40. Ning C, Wang D, Zheng X, Zhang Q, Zhang S. Mrode R, et al. Eigen decomposition expedites longitudinal genome-wide association studies for milk production traits in Chinese Holstein. *Genet Sel Evol.* (2018) 50:12. doi: 10.1186/s12711-018-0383-0

41. Kim S, Lim B, Cho J, Lee S, Dang CG, Jeon JH, et al. Genome-Wide identification of candidate genes for milk production traits in Korean Holstein Cattle. *Animals (Basel).* (2021) 11:1392. doi: 10.3390/ani11051392

42. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ Jr, Crooker BA, et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC Genomics.* (2011) 12:408.

43. An B, Xia J, Chang T, Wang X, Xu L, Zhang L, et al. Genome-wide association study reveals candidate genes associated with body measurement traits in Chinese Wagyu beef cattle. *Anim Genet.* (2019) 50:386–90. doi: 10.1111/age.12805

44. Carter AJ, Nguyen AQ. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med Genet.* (2011) 12:160. doi: 10.1186/1471-2350-12-160

45. Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics*. (2010) 186:767–73. doi: 10.1534/genetics.110.122549

46. Dzidic A, Kuehnl J, Simic M, Bruckmaier RM. Effects of short and long milking intervals on milking characteristics and changes of milk constituents during the course of milking in crossbred Istrian × Awassi × East-Friesian ewes. *J Dairy Res.* (2022) 16:1–6. doi: 10.1017/S0022029922000036

47. Oberbauer AM, Berry SL, Belanger JM, McGoldrick RM, Pinos-Rodriquez JM, Famula TR. Determining the heritable component of dairy cattle foot lesions. *J Dairy Sci.* (2013) 96:605–13. doi: 10.3168/jds.2012-5485

48. Yamada K, Nomura N, Yamano A, Yamada Y, Wakamatsu N. Identification and characterization of splicing variants of PLEKHA5 (Plekha5) during brain development. *Gene*. (2012) 492:270. doi: 10.1016/j.gene.2011.10.018

49. Atashi H, Salavati M, De Koster J, Ehrlich J, Crowe M, Opsomer G, et al. Genome-wide association for milk production and lactation curve parameters in Holstein dairy cows. *J Anim Breed Genet*. (2020) 137:292–304.

50. Shan D, Li JL, Wu L, Li D, Hurov J, Tobin JF, et al. GPAT3 and GPAT4 are regulated by insulin-stimulated phosphorylation and play distinct roles in adipogenesis. *J Lipid Res.* (2010) 1:1971–81. doi: 10.1194/jlr.M006304

51. Khan MZ, Liu L, Zhang Z, Khan A, Wang D. Mi S, et al. Folic acid supplementation regulates milk production variables, metabolic associated genes and pathways in perinatal Holsteins. *J Anim Physiol Anim Nutr (Berl).* (2020) 104:483–92. doi: 10.1111/jpn.13313

52. Zhang F, Hanif Q, Luo X, Jin X, Zhang J, He Z, et al. Muscle transcriptome analysis reveal candidate genes and pathways related to fat and lipid metabolism in Yunling cattle. *Anim Biotechnol*. (2021) 7:1–8. doi: 10.1080/10495398.2021.2009846

53. Costa-Urrutia P, Colistro V, Jiménez-Osorio AS, Cárdenas-Hernández H, Solares-Tlapechco J, Ramirez-Alcántara M, et al. Genome-wide association study of body mass index and body fat in Mexican-Mestizo Children. *Genes (Basel).* (2019) 10:945. doi: 10.3390/genes10110945

54. Houten SM, Denis S, Argmann CA, Jia Y, Ferdinandusse S. Reddy JK, et al. Peroxisomal L-bifunctional enzyme (Ehhadh) is essential for the production of medium-chain dicarboxylic acids. *J Lipid Res.* (2012) 53:1296–303. doi: 10.1194/jlr.M024463

55. Li C, Sun D, Zhang S, Wang S, Wu X. Zhang Q, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS ONE.* (2014) 9:e96186. doi: 10.1371/journal.pone.0096186

56. Zhu J, Yang Z, Hao W, Li J, Wang L, Xia J, et al. Characterization of a read-through fusion transcript, BCL2L2-PABPN1, involved in porcine Adipogenesis. *Genes (Basel).* (2022) 13:445. doi: 10.3390/genes13030445

57. Lv H, Meng Q, Wang N, Duan X, Hou X, Lin Y. Cell death-inducing DNA fragmentation factor-α-like effector C (CIDEC) regulates acetate- and β-hydroxybutyrate-induced milk fat synthesis by increasing FASN expression in mammary epithelial cells of dairy cows. *J Dairy Sci.* (2021) 104:6212–21. doi: 10.3168/jds.2020-18975

58. Ou Yang TH, Cheng WY, Zheng T, Maurer MA, Anastassiou D. Breast cancer prognostic biomarker using attractor metagenes and the FGD3-SUSD3 metagene. *Cancer Epidemiol Biomarkers Prev.* (2014) 23:2850–56. doi: 10.1158/1055-9965.EPI-14-0399

59. Jin P, Wu X, Xu S, Zhang H, Li Y, Cao Z, et al. Differential expression of six genes and correlation with fatness traits in a unique broiler population. *Saudi J Biol Sci*. (2017) 24:945–9. doi: 10.1016/j.sjbs.2015.04.014

60. Golik M, Glick G, Reicher S, Shirak A, Ezra E, Zeron Y, et al. Differential expression of ruminant ZNF496 variants: association with quantitative trait locus affecting bovine milk concentration and fertility. *J Dairy Sci.* (2011) 94:2092–102. doi: 10.3168/jds.2010-3655

61. Bhattacharjee J, Borra VJ, Salem ESB, Zhang C, Murakami K, Gill RK, et al. Hepatic ago2 regulates PPARα for oxidative metabolism linked to glycemic control in obesity and post bariatric surgery. *Endocrinology*. (2021) 162:bqab007. doi: 10.1210/endocr/bqab007

62. Gao Y, Jiang J, Yang S, Hou Y, Liu GE, Zhang S, et al. CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics.* (2017) 18:265. doi: 10.1186/s12864-017-3636-3

63. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, Helgadottir A, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet.* (2009) 41:18–24. doi: 10.1038/ng.274

64. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet.* (2010) 42:937–48. doi: 10.1038/ng.686

65. Mägi R, Manning S, Yousseif A, Pucci A, Santini F, Karra E, et al. Contribution of 32 GWAS-identified common variants to severe obesity in European adults referred for bariatric surgery. *PLoS ONE.* (2013) 8:e70735. doi: 10.1371/journal.pone.0070735

66. Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet Sel Evol.* (2015) 47:29. doi: 10.1186/s12711-014-0074-4

67. Choudhary RK, Capuco AV. Expression of NR5A2, NUP153, HNF4A, USP15 and FNDC3B is consistent with their use as novel biomarkers for bovine mammary stem/progenitor cells. *J Mol Histol.* (2021) 52:289–300. doi: 10.1007/s10735-020-09948-8

68. Pan CJ, Chen SY, Jun HS, Lin SR, Mansfield BC, Chou JY. SLC37A1 and SLC37A2 are phosphate-linked, glucose-6-phosphate antiporters. *PLoS ONE.* (2011) 6:e23157.

69. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics.* (2014) 15:442. doi: 10.1186/1471-2164-15-442

70. Saatchi M, Beever JE. Decker JE,Faulkner DB, Freetly HC, Hansen SL, et al. QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. *BMC Genomics.* (2014) 15:1004. doi: 10.1186/1471-2164-15-1004

71. Tribout T, Croiseau P, Lefebvre R, Barbat A, Boussaha M, Fritz S, et al. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. *Genet Sel Evol.* (2020) 52:55. doi: 10.1186/s12711-020-00575-1

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 December 2022
DOI 10.3389/fgene.2022.1031557

OPEN ACCESS

EDITED BY
Jesús Fernández,
Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain

REVIEWED BY
Maria Jesús Carabaño,
Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain
George R. Wiggans,
Council on Dairy Cattle Breeding, United States

*CORRESPONDENCE
Yachun Wang,
wangyachun@cau.edu.cn

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

RECEIVED 30 August 2022
ACCEPTED 14 November 2022
PUBLISHED 02 December 2022

CITATION
Wang A, Brito LF, Zhang H, Shi R, Zhu L, Liu D, Guo G and Wang Y (2022), Exploring milk loss and variability during environmental perturbations across lactation stages as resilience indicators in Holstein cattle.
*Front. Genet.* 13:1031557.
doi: 10.3389/fgene.2022.1031557

# Exploring milk loss and variability during environmental perturbations across lactation stages as resilience indicators in Holstein cattle

Ao Wang[1], Luiz F. Brito[2], Hailiang Zhang[1], Rui Shi[1], Lei Zhu[1], Dengke Liu[3], Gang Guo[4] and Yachun Wang[1]*

[1]Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China, [2]Department of Animal Sciences, Purdue University, West Lafayette, IN, United States, [3]Hebei Sunlon Modern Agricultural Technology Co., Ltd., Dingzhou, China, [4]Beijing Sunlon Livestock Development Co., Ltd., Beijing, China

Genetic selection for resilience is essential to improve the long-term sustainability of the dairy cattle industry, especially the ability of cows to maintain their level of production when exposed to environmental disturbances. Recording of daily milk yield provides an opportunity to develop resilience indicators based on milk losses and fluctuations in daily milk yield caused by environmental disturbances. In this context, our study aimed to explore milk loss traits and measures of variability in daily milk yield, including log-transformed standard deviation of milk deviations (Lnsd), lag-1 autocorrelation (Ra), and skewness of the deviations (Ske), as indicators of general resilience in dairy cows. The unperturbed dynamics of milk yield as well as milk loss were predicted using an iterative procedure of lactation curve modeling. Milk fluctuations were defined as a period of at least 10 successive days of negative deviations in which milk yield dropped at least once below 90% of the expected values. Genetic parameters of these indicators and their genetic correlation with economically important traits were estimated using single-trait and bivariate animal models and 8,935 lactations (after quality control) from 6,816 Chinese Holstein cows. In general, cows experienced an average of 3.73 environmental disturbances with a milk loss of 267 kg of milk per lactation. Each fluctuation lasted for 19.80 ± 11.46 days. Milk loss traits are heritable with heritability estimates ranging from 0.004 to 0.061. The heritabilities differed between Lnsd (0.135–0.250), Ra (0.008–0.058), and Ske (0.001–0.075), with the highest heritability estimate of 0.250 ± 0.020 for Lnsd when removing the first and last 10 days in milk in a lactation (Lnsd2). Based on moderate to high genetic correlations, lower Lnsd2 is associated with less milk losses, better reproductive performance, and lower disease incidence. These findings indicate that among the variables evaluated, Lnsd2 is the most promising indicator for breeding for improved resilience in Holstein cattle.

KEYWORDS

**milk variability, daily milk yield, cattle resilience, lactation curve, dairy cattle**

## 1 Introduction

Dairy cows are affected by many environmental disturbances throughout their lives (Friggens et al., 2017; Berghof et al., 2018; Silpa et al., 2021), including diseases (Rajala-Schultz et al., 1999a), heat stress (Polsky and von Keyserlingk, 2017; Shi et al., 2021; Wankar et al., 2021), cold stress (Hu et al., 2021a), reproductive events (Macciotta et al., 2011; Guarini et al., 2019), and feed availability and quality (Friggens et al., 2016). These disturbances often result in temporary drop or continuous fluctuations in daily milk yield, which can be considered as milk losses relative to the expected lactation curve (Ben et al., 2021). The pattern of milk losses differs among cows and events and can last for long periods. For instance, mastitis events could affect milk yield for more than 30 days, with a milk loss of 50–300 kg per event (van Soest et al., 2016; Adriaens et al., 2021a). Milk fever can result in lower milk yield for up to 6 weeks with milk losses ranging from 1.1 to 2.9 kg per day (Rajala-Schultz et al., 1999b). Intensive genetic selection for milk production traits has led high-yielding cows to experience negative energy balance (NEB) more often in early lactation, which in turn can result in higher incidences of metabolic disorders (Friggens et al., 2013; Brito et al., 2021). Heat stress also contributes to a reduction in milk yield by affecting endocrine and metabolism processes (Wankar et al., 2021), with reports of milk yield declining by approximately 0.41 kg/d when the temperature and humidity index (THI) exceeds 69 (Bouraoui et al., 2002). However, in the past, production performance and lactation dynamics were mainly analyzed using low frequency test-day records (e.g., weekly or monthly; Adriaens et al., 2021b) due to limitations in large-scale data recording. Disturbances are difficult to be monitored when they are of short duration and in the middle of a test-day interval (Elgersma et al., 2018). With the spread of high frequency milk recording equipment, longitudinal data generated by sensors may contain additional information for deriving novel breeding goals (Peng et al., 2009; Brito et al., 2020, 2021).

To study perturbations in milk production, a theoretically undisturbed lactation curve–the expected lactation curve (ELC), needs to be predicted. The overall objective of predicting an ELC is to eliminate the effect of short-term environmental perturbations on daily milk yield and to reduce the variability, thus enabling the characterization of the lactation potential of each cow in the absence of environmental perturbations (Ben et al., 2021). Identifying environmental perturbations to fit ELC is difficult as information about disturbances is often unavailable (Garcia-Baccino et al., 2021). Therefore, it becomes a mainstream approach to calculate ELC from the actual daily milk yield. Compartment model (Ben et al., 2021), fourth-order polynomial quantile regression model (Poppe et al., 2020), nonparametric trend model (Poppe et al., 2020), and Wood model incorporating iterative procedures (Adriaens et al., 2021a; 2021b) have been used to fit ELC. An important limitation of these approaches is the generalization of the ELC to a single model, thus ignoring differences in lactation trends among cows, which is a topic interest of this current study.

The deviation between the observed and expected daily milk yield can be used for describing the longitudinal dynamics of milk yield and identifying milk losses (Adriaens et al., 2021b). Describing deviations in daily milk yield is needed for evaluating the impact of environmental disturbances in milk yield and for applying effective management decisions. Meanwhile, this provides an opportunity for studying the resilience of lactating cows. Resilience can be defined as the animals' ability to maintain their level of production under environmental disturbances or to recover rapidly to the state pertained before exposure to an environmental disturbance (Colditz and Hine, 2016). Resilience has not been included in any national dairy cattle selection goal to date (Berghof et al., 2018; Poppe et al., 2022b). This is due to the insufficient research on the definition of the best approaches for quantifying resilience, biological validation of resilience indicators, and the selection directions for resilience which are partially encompassed by health, reproduction, and longevity traits in the current selection goals. Genetically selecting for improved resilience could improve herd productivity (Colditz and Hine, 2016; Poppe et al., 2022b), result in better animal welfare (Mulder and Rashidi, 2017), reduce the use of drugs and antibiotics for treating diseases (Konig and May, 2019), and is significantly associated with easier management and lower production cost of herds (Berghof et al., 2018). Many studies have proposed a data-driven approach to derive resilience indicators based on longitudinal data such as daily milk yield (Elgersma et al., 2018; Poppe et al., 2020; Adriaens et al., 2021b; Ben et al., 2021). These methods rely on the assumption that individuals with less fluctuation in longitudinal records are more resilient than those with greater variability. Poppe et al. (2020) used fluctuations in daily milk yield to derive resilience indicators and proposed the log-transformed variance of deviations from lactation curves as the best indicator. Elgersma et al. (2018) defined three traits related to the number of drops in milk yield using the Student *t* test and found that the variance of milk production is the best resilience indicator to predict udder health, ketosis, and longevity. Optimal resilience indicators should have high heritability to enable effective genetic selection for practical applications and ideally favorable genetic correlation with economically important traits. However, the potential of milk loss traits, which directly reflect fluctuations in daily milk yield (e.g., magnitude and duration of milk loss), as suitable resilience indicators has not been previously explored. Furthermore, although resilience indicators based on variability in longitudinal data have been proposed, the calculation of resilience indicators and the genetic relationships with traits already included in selection indexes need to be explored in Chinese Holstein herds.

In this context, the main objectives of this study were 1) to characterize lactation curves and milk yield variability in Holstein cattle; and 2) to investigate the genetic background of milk loss traits and variability traits as resilience indicators and their genetic correlations with economically important traits.

Results of this study will contribute to the identification of appropriate resilience indicators to be used for genetically improving resilience of high-yielding Holstein cattle.

## 2 Materials and methods

### 2.1 Datasets

A total of 11,536,488 daily milk yield records from 22,666 Holstein cows raised in three herds (owned by a single entity) located in Hebei (China) were available for this study. The data was collected from January 2017 to January 2021. The daily milk yield of each cow was extracted from the farm management software. Animals were housed in free-stall systems, fed total mixed rations, and milked three times per day on rotary milking systems. The pedigree of cows with phenotypic records after data editing were traced back as many generations as possible. The final pedigree included 21,574 females and 2,447 males born from 1907 to 2018.

Additional economically important traits were also included in this study. Five reproduction traits were evaluated, including age at first calving in heifers (AFC), age at first insemination in heifers (AFS), interval from first to last insemination in heifers (IFL_H) and cows (IFL_C), and interval from calving to first insemination (ICF), all measured in days. Additional details about the definition of the reproduction traits can be found in Guo et al. (2014) and Liu et al. (2017). Three longevity traits, also measured in days, included the number of days from the first calving to the end of the first (Lon1) and second (Lon2) lactation or culling, and productive life (PL), which refers to the number of days from the first calving to culling or death. The definitions of the longevity traits are described in Zhang et al. (2021). Furthermore, four health traits included udder health (UDDE), reproductive disorders (REPR), metabolic disorders (METB), and digestive disorders (DIGS), as detailed in Wang et al. (2022). The health traits were defined as binary traits with a value of one indicating if a cow had at least one health problem at any time during the corresponding lactation, and 0 otherwise. The number of individuals with reproduction traits, longevity traits, and health traits ranged from 3,871 (IFL_H) to 8,860 (ICF), 883 (PL) to 2,610 (Lon1), and 5,921 (METB and DIGS) to 7,347 (UDDE and REPR), respectively. These traits were recorded until June 2021. The descriptive statistics of these traits used to estimate genetic correlations are presented in Supplemental Table S1.

### 2.2 Data analyses

#### 2.2.1 Data pre-processing

From the initial dataset, only milk yield records measured from days in milk (DIM) 1–305 days, milk yield from 2.5 to 100 kg per day, and non-duplicated records were retained for further analyses. Only cows with age at first calving between 600 and 1,800 days were included in the study. The specific data editing steps, with information on the quality control used, the number of cows, lactations, and records after each editing step, are presented in Supplemental Table S2 (Items 1–9). After the quality control, 22,366 lactations (parity 1 = 7,995; parity 2 = 6,160; parity 3 + = 8,211) were kept for further analyses. A total of 27.61% of lactations had more than 300 milk yield records and 1.85% of the lactations had all 305 milk yield records.

#### 2.2.2 Lactation clustering

Cluster analysis was performed on all lactations in order to group lactations with similar patterns of daily milk yield. The objectives of clustering were 1) to identify and eliminate outliers in each group, 2) to obtain the expected milk yield for missing values for DIM 1–4 days and DIM 305 within the imputation process of missing daily records described in Section 2.2.3, and 3) to account for differences in lactation patterns in the statistical models fitted for resilience indicators.

To minimize clustering divergences caused by differences in the range of milk yield per lactation and emphasize inter-cluster homogeneity (Lee et al., 2020), the phenotypic records were normalized based on the Z-score transformation method. Afterwards, for DIM 31 to 270 within each lactation, the average milk yield for each 10 days was calculated, and 24 average values for each lactation per cow were obtained. Based on these average milk yield records, a principal component analysis (PCA) was performed, and the first five principal components (PC1 to PC5) accounting for 70% of the total variation were considered as attribute points to further measure the similarity across lactations.

The Agglomerative Hierarchical Clustering algorithm (Murtagh and Contreras, 2011) was used to cluster and group lactations, and Euclidean distance was used to measure intra-class distances between two lactations as (Warren Liao, 2005):

$$\boldsymbol{d(A,\,B)} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left(\boldsymbol{m_{A,i}} - \boldsymbol{m_{B,i}}\right)^{\boldsymbol{2}}}$$

where $d(A,\,B)$ is the distance between lactations A and B; $m_{A,i}$ and $m_{B,i}$ are the i$^{th}$ PC in lactation A and B, respectively; and, N is the total number of PCs (equal to 5). To minimize the square sum of intra-class deviations and maximize the square sum of inter-class deviations, the Ward linkage method was used to measure inter-class distance between group pairs (Murtagh and Contreras, 2011).

The silhouette coefficient was adopted for the selection of the number of clusters (Aranganayagi and Thangavel, 2007). Six was the most appropriate number of clusters due to the highest silhouette coefficient, and additional details about the silhouette coefficient of different number of clusters were presented in Supplemental Table S3. The average trends of daily milk yield for each cluster are presented in Figure 1.

Among the six groups, the largest cluster (Figure 1A) included 10,192 lactations (45.57%), while only 173 lactations (0.77%) were included in the smallest cluster (Figure 1F). Descriptive statistics on lactation clustering of the final dataset are detailed in Section 3.1. Within each group, the records deviating three or more SD from the mean were removed for each DIM. A total of 96,601 outlier records (1.50%) were removed as detailed in Supplemental Table S2 (Item 10).

### 2.2.3 Phenotypic data imputation

To obtain complete daily milk yield records from DIM 1 to 305, the missing records were imputed for each lactation. For missing values for DIM 1–4 days and DIM 305, the normalized average milk yield of the corresponding DIM in each cluster was used. Missing milk yield was calculated as the normalized value multiplied by the standard deviation of the non-missing milk yield of the lactation and then added to the mean. After a series of quality control on the record distribution, there was little difference between the average milk yield calculated *via* non-missing values and the true average milk yield. For DIM five to 304, the missing records were sequentially imputed using linear regression interpolation in order of DIM. A total of five records from days n – 4, n – 3, n – 2, n – 1, and n + k was used to fit a first-order linear regression model, where k was the number of days between day n and the next day where daily milk yield was recorded. The regression value for day n was the filled value on that day until all missing values were filled in for each lactation. After imputation, 305 records of daily milk yield for

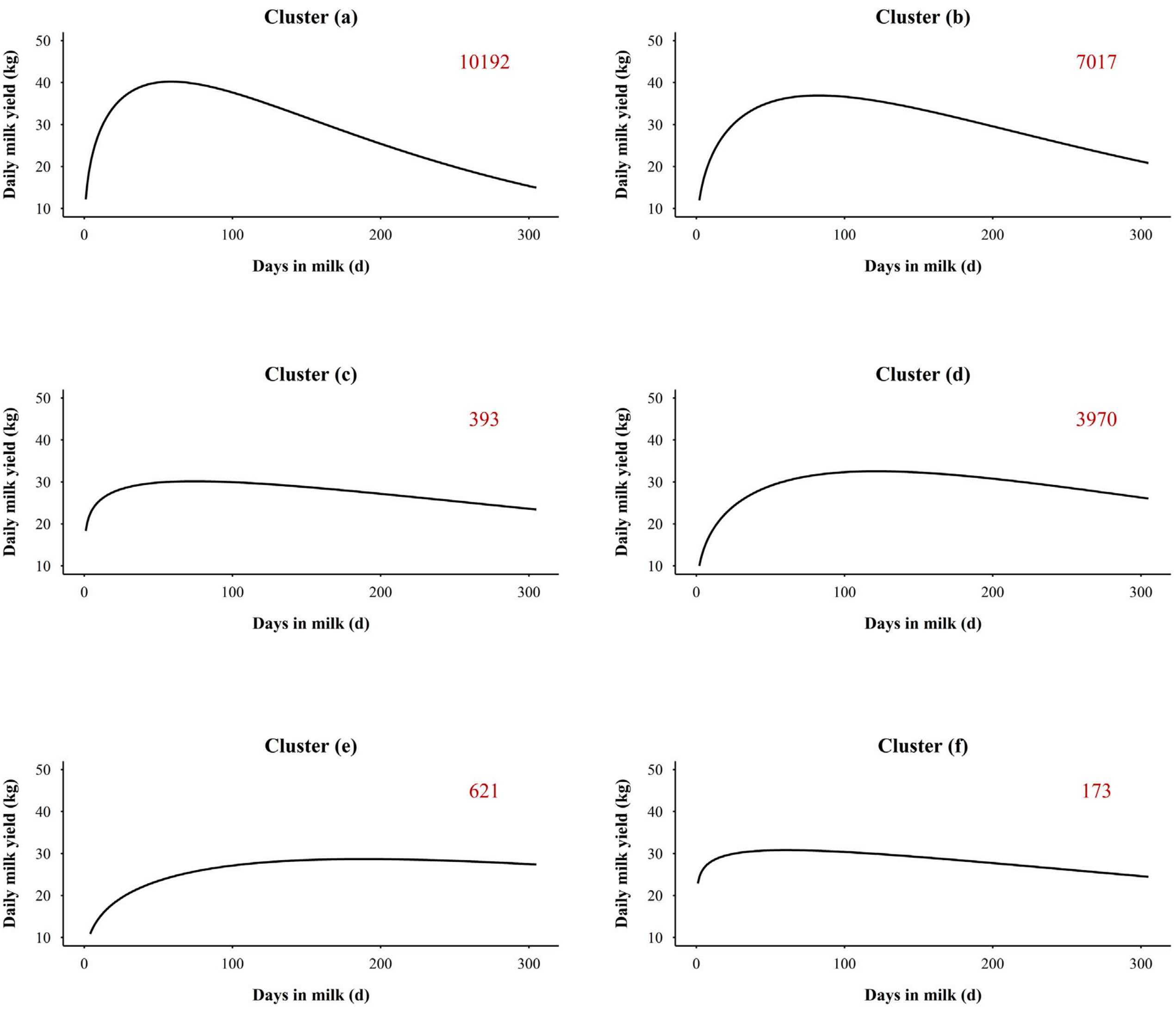


FIGURE 1
Average daily milk yield in six lactation clustering groups. The number in the upper right corner indicates the number of lactations in each cluster. **(A–F)** refer to cluster group (a), (b), (c), (d), (e), and (f), respectively.

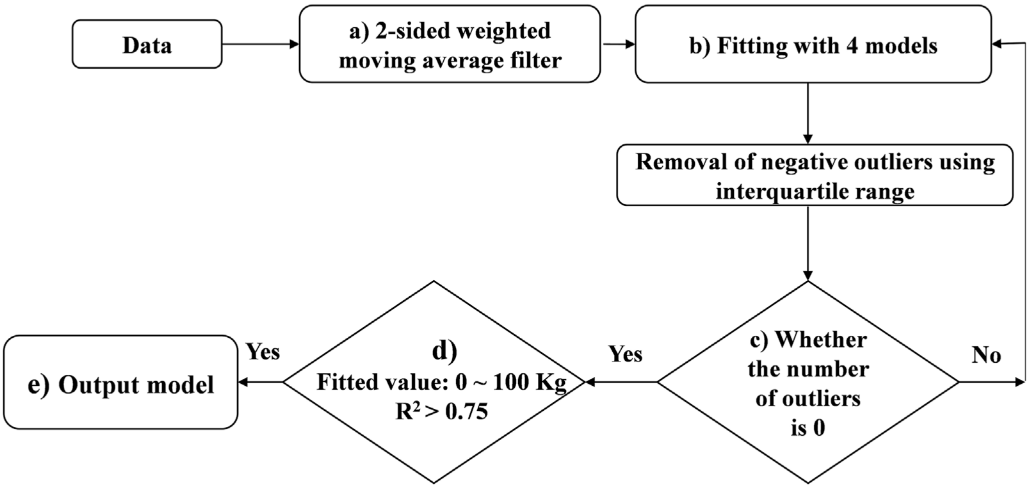

Data
a) 2-sided weighted moving average filter
b) Fitting with 4 models
Removal of negative outliers using interquartile range
c) Whether the number of outliers is 0
No
Yes
d) Fitted value: 0 ~ 100 Kg
R² > 0.75
Yes
e) Output model

FIGURE 2
Illustrative scheme of the process of fitting the expected lactation curve (ELC).

22,366 lactations were obtained as detailed in Supplemental Table S2 (Item 11).

## 2.3 Fitting individual lactation curves

To obtain the expected lactation curve (ELC) of each parity, an iterative procedure was implemented for each lactation with the method presented in Figure 2, and the detailed steps are as follows:

1) A 2-sided weighted moving average filter with a window of 5 days was established in process (a), which means that the expected milk yield on a certain day ($x_t$) is the weighted average of the milk yield in day $x_{t-2}$, $x_{t-1}$, $x_t$, $x_{t+1}$, and $x_{t+2}$. The formula is as follows:

$$x_t = 0.1x_{t-2} + 0.2x_{t-1} + 0.4x_t + 0.2x_{t+1} + 0.1x_{t+2}$$

2) In the first iteration, it was assumed that the expected shape of the optimal lactation curve for each lactation was different. In process (b), four lactation curve models were used to fit each lactation on all data, including the Wood (Wood, 1967), Nelder (Nelder, 1966), Wilmink (Wilmink, 1987), and Ali-Schaeffer (Ali and Schaeffer, 1987) models. The four models can be described as:

$$\boldsymbol{Y_t = at^b e^{-ct}\,(\text{Wood model})}$$

$$\boldsymbol{Y_t^{-1} = a + bt^{-1} + ct\,(\text{Nelder model})}$$

$$\boldsymbol{Y_t = a + bt^{-0.05t} + ct\,(\text{Wilmink model})}$$

$$\boldsymbol{Y_t = a + bt + ct^2 + d\log t + e(\log t)^2\,(\text{Ali} - \text{Schaeffer model})}$$

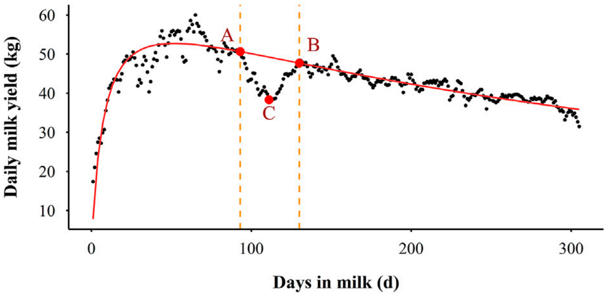


FIGURE 3
An illustrative example of the definition of the milk fluctuation phase. The scatter indicates the actual daily milk yield, the red line represents the expected lactation curve (ELC), the section AB is a fluctuation phase, point A is the start of the fluctuation, point B is the end of the fluctuation, and point C is the highest decrease of the fluctuation.

Where $Y_t$ is the daily milk yield, $t$ is DIM and $a$, $b$, $c$, $d$, and $e$ are the model parameters.

3) Calculate determination coefficient ($R^2$) of the four models and select the model with the highest $R^2$ as the optimal model for that lactation for the subsequent iterative procedure.
4) Calculate the deviations between the actual values and the fitted values for each DIM currently retained (for the first iteration, the number of deviations is 305), as well as the lower quartile (LQ) and the interquartile ranges (IQR) of these deviations.

5) Remove all data with deviation less than LQ-1.5*IQR as outliers to obtain the filtered data resulting from the iteration.
6) Check whether the number of outliers is 0 (as process (c) showed). If not, fit the same lactation curve model on the filtered data from the previous step, and calculate the $R^2$ of the model.
7) Repeat steps (4) to (6) until no outliers are identified. Up to this step, we obtained ELC for each lactation.
8) In process (d), a secondary quality control for ELC was performed. Only ELC with daily milk yield between 0 and 100 kg and $R^2$ (based on the last iteration) > 0.75 were kept in this study.

Furthermore, the lactations with 305 days milk yield deviating three or more SD from the mean and cows with unknown parents were excluded. Finally, 8,935 lactations were obtained for 6,816 cows, as detailed in Supplemental Table S2 (Items 12–14).

## 2.4 Definition of milk loss traits and variability traits as resilience indicators

In this study, the deviations between actual records and the ELC fitted values for each lactation were calculated and expected to contain information about resilience and response to environmental disturbances in Holstein cows. These deviations were expected to be around zero in the absence of perturbations, while during perturbations they would be consistently negative. The number of deviations was 305 for a lactation. A fluctuation was defined as a period of at least 10 successive days of negative deviations for which the milk yield dropped at least once below 90% of the ELC fitted values. An example to illustrate the definition is presented in Figure 3, where the scatters are the daily milk yield in a lactation and the red line indicates the ELC. The section AB is a fluctuation phase. The DIM at points A and B are the beginning and ending of this fluctuation, and the DIM at point C is the highest decrease of this fluctuation. Based on the definitions of deviation and fluctuation, two types of traits were considered as potential resilience indicators in this study: milk loss traits which directly reflect fluctuations in daily milk yield and variability traits obtained by the deviations.

The 305 days milk yield (MY305) and milk loss traits such as the milk loss (ML; in Kg), the number of ML events (NML), the total duration of ML events within a lactation (TDML; in days), the percentage of ML to MY305 (MLP; in %), the duration of each ML period (DML; in days), and milk loss in each ML period (MLF; in Kg) were calculated for each parity. MY305 is calculated by summing up the imputed daily milk yield which included both measured and imputed daily records. ML refers to the sum of the daily milk yield which dropped in all fluctuation phases in a lactation. NML refers to the number of fluctuation events for daily milk yield per lactation (i.e., number of ML). TDML refers to the total duration (in days) of all fluctuation per lactation. MLP refers to the proportion of ML to MY305 per lactation. DML and MLF refer to the duration (in days) and ML in each fluctuation per lactation, respectively. Thus, there may be more than one DML and MLF per lactation.

Through the definitions of deviation, three variability traits were explored within each parity: log-transformed standard deviation of milk deviations (Lnsd), lag-1 autocorrelation of milk deviations (Ra), and skewness of milk deviations (Ske). To identify the effect of lactation stage on resilience, these three variability traits were calculated based on four periods: the entire lactation (Lnsd1, Ra1, and Ske1, from DIM 1–305), lactation period when removing the first and last 10 days (Lnsd2, Ra2, and Ske2, from DIM 11–295), during the lactation peak period (Lnsd3, Ra3, and Ske3, from DIM 60–90), and the period consisting of each DIM when the actual milk yield was below the ELC fitted value (Lnsd4, Ra4, and Ske4).

## 2.5 Genetic analyses

### 2.5.1 Estimation of genetic parameters

The GLM procedure of the SAS software (version 9.4; SAS Institute Inc.) was performed to identify the systematic effects that should be included in the genetic models on milk loss traits and variability traits. Variance and co-variance components were estimated using the Average Information Restricted Maximum Likelihood algorithm implemented in the DMU software (Madsen et al., 2006). Heritability of MY305, milk loss traits (ML, NML, TDML, and MLP), and all variability traits was estimated based on single-trait animal model and heritability of DML and MLF was estimated based on single-trait repeatability animal model.

The single-trait animal model used can be described as:

$$\boldsymbol{y_{ijklmnp}} = \boldsymbol{hys_i} + \boldsymbol{p_j} + \boldsymbol{c_k} + \boldsymbol{m_l} + \boldsymbol{afc_m} + \boldsymbol{a_n} + \boldsymbol{e_{ijklmnp}} \quad (1)$$

where $y_{ijklmnp}$ are the phenotypic records for MY305, milk loss traits (ML, NML, TDML, and MLP), and all variability traits, $hys_i$ is the fixed effect of herd-calving year-calving season (42 levels); $p_j$ is the fixed effect of parity (five levels, including 1, 2, 3, 4, and 5+); $c_k$ is the fixed effect of cluster group (six levels); $m_l$ is the fixed effect of lactation curve model (four levels–the four lactation models described in Section 2.3); $afc_m$ is the fixed effect of age at first calving (four levels, including 22 or less months of age, 23 to 24, 25 to 26, and 27 months and older); $a_n$ is the random additive genetic effect; $e_{ijklmnp}$ is the random residual effect. It was assumed that $a \sim N(0, \mathbf{A}\sigma_a^2)$ and $e \sim N(0, \mathbf{I}\sigma_e^2)$, where $\mathbf{A}$ is the matrix of additive genetic relationships constructed based on pedigree information, $\sigma_a^2$ is the additive genetic variance, $\mathbf{I}$ is an identity matrix, and $\sigma_e^2$ is the residual variance.

The single-trait repeatability animal model can be described as:

$$\boldsymbol{y_{ijklmnpqr}} = \boldsymbol{hys_i} + \boldsymbol{p_j} + \boldsymbol{c_k} + \boldsymbol{m_l} + \boldsymbol{afc_m} + \boldsymbol{DIM_n} + \boldsymbol{a_p} + \boldsymbol{pe_q} + \boldsymbol{e_{ijklmnpqr}} \tag{2}$$

where $y_{ijklmnpqr}$ are the phenotypic records for DML and MLF, $DIM_n$ is the fixed effect of lactation stage at the beginning of the ML (four levels, including 1–44 days, 45–99 days, 100–199 days, and 200–305 days); $pe_q$ is the random permanent environmental effect with $pe \sim N(0, \mathbf{I}\sigma^2_{pe})$. Other fixed and random effects are the same as in the single-trait model.

The genetic correlations between all pairs of resilience indicators were calculated based on bivariate animal models. The bivariate-trait animal model included the same effects as the single-trait model. The assumptions of additive genetic and the residual effects are:

$$\begin{bmatrix} \boldsymbol{a_1} \\ \boldsymbol{a_2} \end{bmatrix} \sim \boldsymbol{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{A} \otimes \begin{pmatrix} \boldsymbol{\sigma^2_{a_1}} & \boldsymbol{\sigma_{a_1a_2}} \\ \boldsymbol{\sigma_{a_1a_2}} & \boldsymbol{\sigma^2_{a_2}} \end{pmatrix}\right]$$

$$\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} \sim \boldsymbol{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{I} \otimes \begin{pmatrix} \boldsymbol{\sigma^2_{e_1}} & \boldsymbol{\sigma_{e_1e_2}} \\ \boldsymbol{\sigma_{e_1e_2}} & \boldsymbol{\sigma^2_{e_2}} \end{pmatrix}\right]$$

where $a_i$ is the additive genetic effects for trait $i$, $\sigma^2_{a_i}$ is the additive genetic variance of trait $i$, $\sigma_{a_ia_j}$ is the additive genetic covariance between trait $i$ and $j$, $e_i$ is the residual effect for trait $i$, $\sigma^2_{e_i}$ is the residual variance of trait $i$, $\sigma_{e_ie_j}$ is the residual covariance between trait $i$ and $j$. The heritability, genetic correlations, and reliability of the estimated breeding value (EBV) for each trait were calculated as described in Su et al. (2007) and Luo et al. (2021).

### 2.5.2 Genetic correlation with milk production, reproduction, longevity, and health traits

Genetic correlations between resilience indicators with economically important traits included milk production, reproduction, longevity, and health traits were calculated based on bivariate animal models. The milk production trait refers to MY305 calculated in this study. For the milk production trait and resilience indicators, the animal models used are the same as model [1]. For the five reproduction traits, the fixed effects included in the models were herd-year of measurement, parity and calving season, and the random effects of animal additive genetic, permanent environment, and residual effects, which are detailed in Guo et al. (2014) and Liu et al. (2017). For the three longevity traits, the fixed effects of age at first calving, herd-year of birth, and birth season and the random effect of additive genetic and residual effects were fitted and are detailed in Zhang et al. (2021). Furthermore, for the four health traits, herd-year of measurement, parity, and calving season were fitted as fixed effects in the model and animal additive genetic, permanent environment, and residual as random effects, as detailed in Wang et al. (2022). These analyses were implemented using the DMU software (Madsen et al., 2006).

## 2.6 Validation

To validate the resilience indicators evaluated in this study and determine whether selection on these indicators can improve the "true" resilience of offspring, thirty-four bulls with at least 40 daughters in first parity with divergent resilience indicators were retained. For each bull, the daughters were divided in prediction and validation datasets based on their birth date with allocation of 80% (older) and 20% (younger) of the animals in the prediction ($n = 2{,}566$) and validation ($n = 641$) datasets, respectively. The EBV of the resilience indicators for each cow in the validation dataset were estimated based on the phenotypes of the prediction dataset and pedigree information, and the model was the same as model 1. In total, the top and bottom 20% resilient animals were selected based on their EBV for each resilience indicator. The differences in EBV for production, reproduction, longevity, and health traits between the top and bottom resilience EBVs were statistically compared based on a Student $t$ test.

# 3 Results

## 3.1 Lactation clustering and lactation curves

The descriptive statistics for the lactation clusters of the final dataset are presented in Supplemental Table S4. The final number of lactations in each cluster group was reduced from the number presented in Figure 1, but the order of numbers of lactations and the trend of daily milk yield within cluster groups did not change. The largest cluster [group (a)] included 3,877 lactations (43.39%), while the smallest cluster [group (f)] contained 16 lactations (0.18%). The differences of lactation curves among the six cluster groups mainly focused on parity, peak day, peak yield, and lactation persistency. The average parity for groups (a), (b), (c), (d), (e), and (f) was 2.67 ± 1.11, 2.12 ± 1.17, 2.03 ± 1.18, 1.37 ± 0.85, 1.36 ± 0.83, and 2.12 ± 1.18, respectively. The highest peak yield was in group (a) (48.27 ± 10.40 kg), with 10.80 kg difference from the lowest group [group (e), 37.47 ± 6.94 kg]. The peak day in group (a), (b), (c), and (f) was at the early lactation period (DIM 1–99), while the peak day in groups (d) and (e) was at the mid lactation (DIM 100–199). The latest peak day was observed for group (e) with 172.14 ± 57.40 days. For the three groups with the highest number of lactations, the groups (a) and (b) presented the highest average parity, normal peak day, and a clear downward phase after peak day, which is more representative of multiparous cows' lactation curve. The group (d) presented a lower average parity and lower peak yield and slower decline in late lactation than groups (a) and (b), which represents the majority of primiparous cows. The groups (c), (e), and (f) exhibited an atypical pattern (5.3%) characterized by higher milk yield in early lactation, a delayed lactation peak, or a slower decline of milk yield in late lactation, while some reversal shaped curves and continuously increasing curve were also included in these groups.

TABLE 1 Descriptive statistics of 305 days milk yield and resilience indicators in Chinese Holstein cattle.

| Trait[1] | N | Mean | SD | Min | Max | Coefficient of variation |
|---|---|---|---|---|---|---|
| MY305, kg | 8,935 | 9,603.12 | 2,354.72 | 2,534.32 | 17,845.34 | 24.52 |
| NML, time | 8,935 | 3.73 | 1.37 | 0 | 9 | 36.73 |
| TDML, d | 8,935 | 73.12 | 26.42 | 0 | 211 | 36.13 |
| ML, kg | 8,935 | 267.00 | 185.04 | 0.00 | 2,170.31 | 69.30 |
| MLP, % | 8,935 | 2.90 | 2.14 | 0.00 | 30.47 | 73.79 |
| DML, d | 31,606 | 19.80 | 11.46 | 10 | 167 | 57.88 |
| MLF, kg | 31,606 | 67.48 | 77.80 | 2.82 | 1,989.74 | 115.29 |
| Lnsd1 | 8,935 | 1.11 | 0.41 | −0.01 | 2.62 | 36.94 |
| Lnsd2 | 8,935 | 0.97 | 0.38 | −0.11 | 2.41 | 39.18 |
| Lnsd3 | 8,935 | 0.89 | 0.44 | −0.47 | 2.65 | 49.44 |
| Lnsd4 | 8,935 | 0.78 | 0.46 | −0.52 | 2.38 | 58.97 |
| Ra1 | 8,935 | 0.83 | 0.08 | 0.36 | 0.98 | 9.64 |
| Ra2 | 8,935 | 0.87 | 0.05 | 0.66 | 0.99 | 5.74 |
| Ra3 | 8,935 | 0.83 | 0.07 | 0.46 | 0.99 | 8.43 |
| Ra4 | 8,935 | 0.77 | 0.10 | 0.37 | 0.98 | 12.99 |
| Ske1 | 8,935 | −1.82 | 1.92 | −10.65 | 6.81 | 105.49 |
| Ske2 | 8,935 | −0.96 | 0.79 | −4.97 | 3.29 | 82.29 |
| Ske3 | 8,935 | −0.68 | 0.76 | −3.24 | 4.18 | 111.76 |
| Ske4 | 8,935 | −1.57 | 0.57 | −5.16 | 0.10 | 36.31 |

[1] N, the number of records or indicators; MY305, 305 days milk yield; NML, number of milk loss events; TDML, total number of days for milk loss per lactation; ML, sum of the milk yield which dropped in all fluctuation phases in a lactation; MLP, the percentage of ML, to MY305; DML, length of each milk loss period in days; MLF, milk loss in each milk loss period; Lnsd, log-transformed standard deviation of milk deviations; Ra, lag-1, autocorrelation of milk deviations; Ske, skewness of milk deviations. These three variability traits were calculated based on records from the entire lactation (Lnsd1, Ra1, and Ske1, from DIM 1–305), lactation period when removing the first and last 10 days (Lnsd2, Ra2, and Ske2, from DIM 11–295), during the lactation peak period (Lnsd3, Ra3, and Ske3, from DIM 60–90), and the period consisting of each DIM, when the actual milk yield was below the expected lactation curve (ELC) fitted value (Lnsd4, Ra4, and Ske4), respectively.

The comparisons of four lactation curve models are presented in Supplemental Table S5. There were 5,137 lactations with the Ali-Schaeffer model as the optimal model in fitting ELC, accounting for 57.49%. While the Nelder model included the lowest number of lactations (731 lactations). After the iterative procedure and quality control, the average amount of data used to predict the ELC was 283.02 ± 14.76, and the average $R^2$ of the ELC was 0.89 ± 0.06. The major difference between the four models was the percentage of the first parity. There were 66.94%, 82.17%, 31.15%, and 37.03% of lactations in which the first parity data were fitted with Wood, Nelder, Wilmink, and Ali-Schaeffer model, respectively.

In this study, cluster group and lactation curve model had a significant effect ($P < 0.05$) on milk loss traits and variability traits. The least squares mean estimates (LSM) of various levels on ML and Lnsd2 and multiple comparisons based on Bonferroni *t* corrected are presented in Supplemental Table S6. The LSM of ML and Lnsd2 in group (c) and (f) were significantly higher than that in other groups ($P < 0.05$), and the ML and Lnsd2 were lowest in group (a). For the lactation curve model, the ELC calculated by Ali-Schaeffer model had the highest ML and Lnsd2, whereas the lowest ones were calculated by Wilmink model.

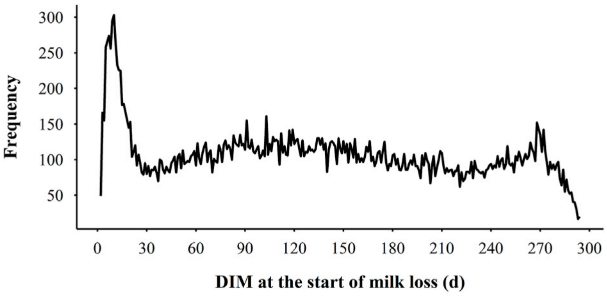


FIGURE 4
The distribution of days in milk (DIM) at the start of milk loss in each lactation.

## 3.2 Descriptive statistics and genetic parameters of resilience indicators

The distributions of MY305, milk loss traits, and variability traits are presented in Supplemental Figure S1. MY305, NML, and TDML were normally distributed and other milk loss traits (ML, MLP, DML, and MLF) showed a right skewed distribution.

TABLE 2 Estimates of additive genetic variance ($\hat{\sigma}_a^2$), permanent environment variance ($\hat{\sigma}_{pe}^2$), residual variance ($\hat{\sigma}_e^2$), and heritability ($\widehat{h^2}$) for 305 days milk yield, milk loss traits, and variability traits.

| Trait[1] | N | $\hat{\sigma}_a^2$ ($\hat{\sigma}_{pe}^2$) | $\hat{\sigma}_e^2$ | $\widehat{h^2}$ |
|---|---|---|---|---|
| MY305, kg | 8,935 | 861,242.143 | 2,704,320.109 | 0.242 ± 0.036 |
| NML, time | 8,935 | 0.043 | 1.740 | 0.024 ± 0.010 |
| TDML, d | 8,935 | 9.155 | 647.551 | 0.014 ± 0.008 |
| ML, kg | 8,935 | 1,758.075 | 26,958.841 | 0.061 ± 0.016 |
| MLP, % | 8,935 | 0.182E-04 | 0.397E-03 | 0.044 ± 0.013 |
| DML[2], d | 31,606 | 0.546 (0.698) | 125.352 | 0.004 ± 0.003 |
| MLF, kg | 31,606 | 29.506 (167.004) | 5,430.781 | 0.005 ± 0.003 |
| Lnsd1 | 8,935 | 0.016 | 0.100 | 0.137 ± 0.020 |
| Lnsd2 | 8,935 | 0.027 | 0.080 | 0.250 ± 0.020 |
| Lnsd3 | 8,935 | 0.020 | 0.131 | 0.135 ± 0.020 |
| Lnsd4 | 8,935 | 0.031 | 0.137 | 0.184 ± 0.020 |
| Ra1 | 8,935 | 0.304E-03 | 0.495E-02 | 0.058 ± 0.015 |
| Ra2 | 8,935 | 0.539E-04 | 0.187E-02 | 0.028 ± 0.011 |
| Ra3 | 8,935 | 0.578E-04 | 0.407E-02 | 0.014 ± 0.008 |
| Ra4 | 8,935 | 0.809E-04 | 0.975E-02 | 0.008 ± 0.007 |
| Ske1 | 8,935 | 0.221 | 2.732 | 0.075 ± 0.016 |
| Ske2 | 8,935 | 0.496E-03 | 0.553 | 0.001 ± 0.005 |
| Ske3 | 8,935 | 0.019 | 0.495 | 0.037 ± 0.012 |
| Ske4 | 8,935 | 0.004 | 0.304 | 0.013 ± 0.008 |

[1] N, the number of records or indicators; MY305, 305 days milk yield; NML, number of milk loss events; TDML, total number of days for milk loss per lactation; ML, sum of the milk yield which dropped in all fluctuation phases in a lactation; MLP, the percentage of ML to MY305; DML, length of each milk loss period in days; MLF, milk loss in each milk loss period; Lnsd, log-transformed standard deviation of milk deviations; Ra, lag-1 autocorrelation of milk deviations; Ske, skewness of milk deviations. These three variability traits were calculated based on records from the entire lactation (Lnsd1, Ra1, and Ske1, from DIM 1–305), lactation period when removing the first and last 10 days (Lnsd2, Ra2, and Ske2, from DIM 11–295), during the lactation peak period (Lnsd3, Ra3, and Ske3, from DIM 60–90), and the period consisting of each DIM when the actual milk yield was below the expected lactation curve (ELC) fitted value (Lnsd4, Ra4, and Ske4), respectively.

The variability traits had different distribution characteristics in the four periods evaluated. All four Lnsd variables were normally distributed and the four Ra variables showed a left skewed distribution. Ske1, Ske2, and Ske4 were left skewed while Ske3 was right skewed.

The descriptive statistics for MY305, milk loss traits, and variability traits are presented in Table 1. MY305 ranged from 2,534.32 kg to 17,845.34 kg, with an average of 9,603.12 ± 2,354.72 kg. In general, cows experienced 3.73 ± 1.37 perturbations per lactation, ranging from 0 to 9. Cows in parity 1, 2, and 3 + experienced 3.70 ± 0.02, 3.76 ± 0.03, and 3.78 ± 0.03 perturbations per lactation, respectively. Only 32 lactations (0.36%) had no perturbations, while 3.54%, 14.08%, 26.98%, 27.53%, and 27.51% lactations had 1, 2, 3, 4, and 5 or more perturbations, respectively. The average TDML was 73.12 ± 26.42 days, with an average ML of 267.00 ± 185.04 kg (2.90% to average MY305). For the cows with the most severe milk loss, the TDML was 221 days, with ML of 2,170.31 kg (30.47% of average MY305). For each perturbation, the average DML was 19.80 ± 11.46 days and MLF was 67.48 ± 77.80 kg on average. The coefficient of variation for DML and MLF was 57.88% and 115.29%, respectively. The highest MLF was 1,989.74 kg, which lasted for 167 days. The distribution of DIM at the beginning of ML is presented in Figure 4. There were larger risks for ML from DIM 5–15, DIM 90–110, and DIM 270 and greater based on the prevalence of variability, while lower risks in mid-late lactation stage (DIM 120–250). The greatest risk of ML was in early lactation, with 12.73% ML events beginning within the first 20 days after calving. The average Lnsd1 was 1.11 ± 0.41, which meant the range of the 95% confidence interval for the deviation of actual milk yield from the expected values was ±5.94 kg. Among the Lnsd variables, the largest and lowest variation was observed for Lnsd4 and Lnsd1 with a coefficient of variation of 58.97% and 36.94%, respectively. Among the four Ra variables, the highest mean value was Ra2 (0.87) which was 0.04–0.10 higher than the other Ra, and its minimum value was 0.66 (0.2–0.3 higher than the other Ra variables). The coefficient of variation for Ra variables was small, with the highest being Ra4 (12.99%) and the lowest being Ra2 (5.74%). The average of four Ske variables were all less than 0. Ske3 had the highest average of −0.68 ± 0.76 and Ske1 had the lowest average of −1.82 ± 1.92. The variation of the four Ske variables was quite different, with the coefficient of variation ranging from 36.31% to 111.76%.

Estimates of variance components and heritability for MY305, milk loss traits, and variability traits are presented in Table 2. The heritability for milk loss traits ranged from 0.004 ± 0.003 (DML) to 0.061 ± 0.016 (ML), all of which had low heritability estimates. All four Lnsd variables had moderate heritability estimates (from 0.135 to 0.250). Lnsd2 had the highest heritability at 0.250 ± 0.021, followed by Lnsd4 at 0.184 ± 0.021. Similar heritability estimates were observed for Lnsd1 and Lnsd3. The heritabilities for Ra and Ske were all below 0.10, ranging from 0.001 ± 0.005 (Ske2) to 0.075 ± 0.016 (Ske1). Ra1 (0.058 ± 0.015) and Ske1 (0.075 ± 0.016) had the highest heritability estimates among Ra and Ske variables.

The genetic correlations among the variability traits are presented in Table 3. The genetic correlations within each trait were high among the four periods. For instance, the genetic correlations among the four Lnsd variables ranged from 0.93 ± 0.02 to 0.99 ± 0.00, and among the four Ra variables ranged from 0.69 ± 0.17 to 0.99 ± 0.12. Within each lactation period, the genetic correlations across the variability traits were not consistent. The genetic correlations between Lnsd and Ra were positive across the different periods, with a minimum of 0.28 ± 0.14 (Lnsd1 and Ra1) and a maximum of 0.77 ± 0.06 (Lnsd2 and Ra2). The genetic correlations between Lnsd and Ske as well as Ra and Ske varied considerably across lactation periods. For instance, positive genetic correlations were observed in the first (between Lnsd1 and Ske1; and, Ra1 and

TABLE 3 Genetic ($r_G$) and phenotypic ($r_P$) correlations among variability traits[1].

| | Lnsd1 | Lnsd2 | Lnsd3 | Lnsd4 | Ra1 | Ra2 | Ra3 | Ra4 | Ske1 | Ske2 | Ske3 | Ske4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lnsd1[2] | | 0.93 (0.02) | 0.96 (0.02) | 0.93 (0.02) | 0.28 (0.14) | 0.43 (0.15) | 0.62 (0.15) | 0.38 (0.25) | 0.09 (0.13) | — | −0.61 (0.12) | 0.45 (0.24) |
| Lnsd2 | 0.79 (0.00) | | 0.99 (0.01) | 0.99 (0.00) | 0.70 (0.07) | 0.77 (0.06) | 0.76 (0.10) | 0.70 (0.14) | 0.51 (0.10) | — | −0.68 (0.09) | 0.74 (0.17) |
| Lnsd3 | 0.57 (0.01) | 0.74 (0.01) | | 0.99 (0.01) | 0.67 (0.10) | 0.15 (0.28) | 0.50 (0.16) | 0.57 (0.19) | 0.35 (0.13) | — | −0.75 (0.10) | 0.85 (0.17) |
| Lnsd4 | 0.76 (0.00) | 0.93 (0.00) | 0.67 (0.01) | | 0.63 (0.09) | 0.70 (0.08) | 0.73 (0.12) | 0.54 (0.16) | 0.56 (0.11) | — | −0.67 (0.10) | 0.74 (0.20) |
| Ra1 | 0.15 (0.01) | 0.49 (0.01) | 0.31 (0.01) | 0.47 (0.01) | | 0.78 (0.10) | 0.69 (0.17) | 0.77 (0.19) | 0.88 (0.06) | — | −0.09 (0.21) | 0.81 (0.18) |
| Ra2 | 0.45 (0.01) | 0.66 (0.01) | 0.43 (0.01) | 0.62 (0.01) | 0.57 (0.01) | | 0.80 (0.16) | 0.99 (0.12) | 0.61 (0.15) | — | 0.15 (0.28) | 0.54 (0.28) |
| Ra3 | 0.24 (0.01) | 0.35 (0.01) | 0.60 (0.01) | 0.32 (0.01) | 0.29 (0.01) | 0.50 (0.01) | | 0.74 (0.32) | 0.32 (0.24) | — | −0.37 (0.29) | 0.50 (0.35) |
| Ra4 | 0.46 (0.01) | 0.60 (0.01) | 0.37 (0.01) | 0.70 (0.01) | 0.48 (0.01) | 0.82 (0.00) | 0.40 (0.01) | | | — | — | 0.53 (0.46) |
| Ske1 | 0.46 (0.01) | 0.09 (0.01) | 0.07 (0.01) | 0.03 (0.01) | 0.48 (0.01) | 0.09 (0.01) | 0.04 (0.01) | 0.05 (0.01) | | — | −0.17 (0.19) | 0.61 (0.19) |
| Ske2 | --[3] | — | — | — | — | — | — | — | — | | — | — |
| Ske3 | 0.15 (0.01) | 0.17 (0.01) | 0.21 (0.01) | 0.30 (0.01) | 0.05 (0.01) | 0.12 (0.01) | 0.18 (0.01) | — | 0.13 (0.01) | — | | −0.78 (0.31) |
| Ske4 | 0.07 (0.01) | 0.05 (0.01) | 0.11 (0.01) | 0.16 (0.01) | 0.05 (0.01) | 0.09 (0.01) | 0.07 (0.01) | 0.23 (0.01) | 0.29 (0.00) | — | 0.14 (0.01) | |

[1]The genetic correlations are presented above the diagonal while the phenotypic correlations are below the diagonal.

[2]Lnsd, log-transformed standard deviation of milk deviations; Ra, lag-1, autocorrelation of milk deviations; Ske, skewness of milk deviations. These three variability traits were calculated based on records from the entire lactation (Lnsd1, Ra1, and Ske1, from DIM 1–305), lactation period when removing the first and last 10 days (Lnsd2, Ra2, and Ske2, from DIM 11–295), during the lactation peak period (Lnsd3, Ra3, and Ske3, from DIM 60–90), and the period consisting of each DIM, when the actual milk yield was below the expected lactation curve (ELC) fitted value (Lnsd4, Ra4, and Ske4), respectively.

[3]-- means that the analyses did not converge.

TABLE 4 Genetic and phenotypic correlations among milk loss traits and genetic correlations between milk loss traits and variability traits.

| s | N[2] | ML | NML | TDML | MLP | Lnsd2 | Ra1 | Ske1 |
|---|---|---|---|---|---|---|---|---|
| ML, kg | 8,935 | | 0.21 (0.21) | 0.29 (0.23) | 0.48 (0.13) | 0.96 (0.01) | 0.62 (0.13) | 0.51 (0.16) |
| NML, time | 8,935 | 0.39 (0.01) | | 0.69 (0.20) | 0.58 (0.18) | 0.45 (0.14) | 0.23 (0.24) | −0.27 (0.21) |
| TDML, d | 8,935 | 0.66 (0.01) | 0.67 (0.01) | | 0.78 (0.13) | 0.58 (0.14) | 0.09 (0.32) | −0.41 (0.30) |
| MLP, % | 8,935 | 0.89 (0.00) | 0.39 (0.01) | 0.67 (0.01) | | 0.54 (0.08) | 0.16 (0.20) | −0.02 (0.18) |

[1]The genetic correlations among milk loss traits are presented above the diagonal while the phenotypic correlations are below the diagonal in the first four columns, while the genetic correlations between milk loss traits and variability traits are presented in the last three columns; ML, sum of the milk yield which dropped in all fluctuation phases in a lactation; NML, number of milk loss events; TDML, total number of days for milk loss per lactation; MLP, the percentage of ML to MY305; Lnsd2, log-transformed standard deviation of milk deviations based on the lactation when removing first and last 10 DIM; Ra1, lag-1 autocorrelation of milk deviations based on the entire lactation; Ske1, skewness of milk deviations based on the entire lactation. Lnsd1, Ra2 and Ske2 are the traits with the highest heritability among the three variability traits, respectively.

[2]N: number of records that were used to calculate the genetic correlations.

Ske1) and fourth periods (between Lnsd4 and Ske4; and, Ra4 and Ske4) and negative correlations in the third period (between Lnsd3 and Ske3; and, Ra3 and Ske3).

The genetic correlations among the milk loss traits and between milk loss traits and variability traits are presented in Table 4. The three traits with the highest heritability among the three variability traits (Lnsd2, Ra1, and Ske1) are presented. Positive genetic correlations were observed between different milk loss traits, ranging from 0.21 ± 0.21 (ML and NML) to 0.78 ± 0.13 (TDML and MLP). Lnsd2 and Ra1 had positive genetic correlations with milk loss traits, ranging from 0.09 ± 0.32 (Ra1 and TDML) to 0.96 ± 0.01 (Lnsd2 and ML), with the exception of Ske1 which had mostly negative genetic correlations. However, only Lnsd2 had statistically significant genetic correlations with all four milk loss traits at the 5% level. There were moderate to high genetic correlations between Lnsd2 and all milk loss traits, ranging from 0.45 ± 0.14 (NML) to 0.96 ± 0.01 (ML).

## 3.3 Genetic correlation with milk production, reproduction, longevity, and health traits

The genetic correlations of resilience indicators with production, reproduction, longevity, and health traits are presented in Table 5. The genetic correlations of DML and MLF with routinely evaluated traits are not presented because the analyses did not converge.

The estimated genetic correlations between milk loss traits (NML, TDML, and MLP) and MY305 were negative and ranged from −0.46 ± 0.14 (NML) to −0.75 ± 0.15 (TDML), except for a positive genetic correlation between ML and MY305 (0.60 ± 0.08). The genetic correlation between variability traits and MY305 were positive and ranged from 0.53 ± 0.09 (Ske1) to 0.80 ± 0.04 (Lnsd2).

The estimated genetic correlations between milk loss traits, variability traits and reproduction, longevity, and health traits

TABLE 5 Genetic correlations between milk loss traits, variability traits and production, reproduction, longevity, and health traits.

| s | N[2] | ML | NML | TDML | MLP | Lnsd2 | Ra1 | Ske1 |
|---|---|---|---|---|---|---|---|---|
| MY305, kg | 8,935 | 0.60 (0.08) | −0.46 (0.14) | −0.75 (0.15) | −0.65 (0.11) | 0.80 (0.04) | 0.56 (0.10) | 0.53 (0.09) |
| AFC, d | 4,222 | 0.25 (0.05) | 0.04 (0.10) | 0.03 (0.07) | 0.64 (0.29) | 0.32 (0.03) | 0.31 (0.07) | −0.17 (0.06) |
| AFS, d | 4,222 | −0.22 (0.39) | 0.28 (0.41) | 0.87 (0.42) | −0.05 (0.37) | 0.05 (0.03) | 0.19 (0.07) | 0.37 (0.07) |
| IFL_H, d | 3,871 | −0.86 (0.66) | 0.34 (0.64) | −0.49 (0.83) | −0.91 (0.42) | 0.12 (0.03) | −0.11 (0.11) | −0.19 (0.06) |
| IFL_C, d | 4,476 | −0.16 (0.36) | 0.25 (0.42) | −0.88 (0.96) | −0.32 (0.35) | 0.59 (0.20) | −0.02 (0.44) | 0.12 (0.32) |
| ICF, d | 8,860 | 0.41 (0.18) | 0.38 (0.21) | 0.29 (0.27) | 0.48 (0.18) | 0.19 (0.11) | 0.14 (0.18) | 0.49 (0.17) |
| Lon1, d | 2,610 | 0.16 (0.17) | −0.16 (0.17) | 0.02 (0.11) | −0.97 (0.14) | 0.24 (0.09) | 0.18 (0.13) | 0.19 (0.15) |
| Lon2, d | 1,350 | 0.01 (0.15) | −0.35 (0.16) | −0.19 (0.13) | −0.96 (0.32) | 0.02 (0.10) | 0.08 (0.15) | 0.28 (0.13) |
| PL, d | 883 | 0.05 (0.25) | −0.08 (0.10) | −0.16 (0.23) | −0.96 (0.74) | 0.49 (0.23) | 0.19 (0.15) | 0.16 (0.17) |
| UDDE | 7,347 | 0.58 (0.18) | 0.18 (0.30) | 0.21 (0.41) | 0.19 (0.24) | 0.87 (0.07) | 0.84 (0.14) | 0.62 (0.18) |
| REPR | 7,347 | 0.66 (0.21) | 0.49 (0.24) | 0.69 (0.31) | 0.70 (0.19) | 0.36 (0.22) | 0.30 (0.25) | 0.52 (0.19) |
| METB | 5,921 | −0.49 (0.27) | −0.55 (0.29) | --[3] | 0.01 (0.33) | −0.47 (0.19) | −0.64 (0.25) | −0.95 (0.23) |
| DIGS | 5,921 | −0.87 (1.77) | −0.88 (1.94) | — | −0.87 (1.96) | 0.05 (0.89) | −0.55 (1.29) | −0.33 (1.34) |

[1] MY305, 305 days milk yield; ML, sum of the milk yield which dropped in all fluctuation phases in a lactation; NML, number of milk loss events; TDML, total number of days for milk loss per lactation; MLP, the percentage of ML to MY305; AFC, age at first calving in heifers; AFS, age at first insemination in heifers; IFL_H, interval from first to last insemination in heifers; IFL_C, interval from first to last insemination in cows; ICF, interval from calving to first insemination; Lon1, the days from the first calving to the end of the first lactation or culling; Lon2, the days from the first calving to the end of the second lactation or culling; PL, productive life referring the days from the first calving to culling or death; UDDE, udder health; REPR, reproductive disorders; METB, metabolic disorders; DIGS: digestive disorders; Lnsd2, log-transformed standard deviation of milk deviations based on the lactation when removing first and last 10 DIM; Ra1, lag-1 autocorrelation of milk deviations based on the entire lactation; Ske1, skewness of milk deviations based on the entire lactation. Lnsd1, Ra2, and Ske2 are the traits with the highest heritability among the three variability traits, respectively.

[2] N: number of records that were used to calculate the genetic correlations.

[3] -- means that the analyses did not converge.

were mostly moderate to high while most of them were not significantly different from zero because of the high standard errors. There were favorable and unfavorable genetic correlations for ML with AFC (0.25 ± 0.05), AFS (−0.22 ± 0.39), IFL_H (−0.86 ± 0.66), IFL_C (−0.16 ± 0.36), and ICF (0.41 ± 0.18), while the genetic correlations between NML and reproduction traits were all positive. For variability traits, the genetic correlations between Lnsd2 and reproduction traits were positive and ranged from 0.05 ± 0.03 (AFS) to 0.59 ± 0.20 (IFL_C), and were all statistically significant at the 5% level. However, for the other two variability traits, the genetic correlations ranged from −0.19 ± 0.06 (Ske1 and IFL_H) to 0.49 ± 0.17 (Ske1 and ICF). There were negative genetic correlations between NML and Lon2, TDML and Lon2, MLP and all longevity traits, ranging from -0.97 ± 0.03 (MLP and Lon1) to −0.19 ± 0.03 (TDML and Lon2), whereas the other estimates between milk loss traits and longevity traits were not significantly different from zero. The genetic correlations between variability traits and longevity traits were positive and ranged from 0.02 ± 0.10 (Lnsd2 and Lon2) to 0.49 ± 0.23 (Lnsd2 and PL). Positive genetic correlations, ranging from 0.18 ± 0.30 (NML and UDDE) to 0.70 ± 0.19 (MLP and REPR) were obtained between milk loss traits and UDDE and REPR. The genetic correlations between milk loss traits and METB and DIGS were mostly unfavorable. Similar correlations were obtained in the genetic correlations between variability traits and health traits. Among all health traits, UDDE had the highest genetic correlation with Lnsd2 (0.87 ± 0.07). Among all genetic correlations with economically important traits, the standard errors were on average higher for the milk loss traits than for the variability traits and Lnsd2 had the lowest standard errors on average. For instance, the standard errors for estimates of genetic correlations between milk loss traits and reproduction traits ranged from 0.05 to 0.96, while the standard errors ranged from 0.03 to 0.44 for the variability traits.

## 3.4 Validation

The comparisons of the milk loss, production, reproduction, longevity, and health traits of the top and bottom 20% EBVs in the validation dataset for Lnsd2 are presented in Table 6. The results for Lnsd1, Lnsd3, and Lnsd4 are presented in Supplemental Tables S7–S9. The top 20% of Lnsd EBVs represent the 20% most resilient cows. The top 20% group was significantly better in MLP and 0.72% lower on average than the bottom 20% group. AFC, IFL_H, and Lon1 were significantly better in the top 20% group than in the bottom 20% group among the ten production, reproduction, longevity, and health traits. For AFC and IFL_H, the top 20% group was 17.83 and 17.08 days less than the bottom 20% group, respectively, and Lon1 was 14.98 days longer. However, for the other traits, although the differences between the two groups were not statistically significant, there was still a trend

TABLE 6 Comparison of top 20% and bottom 20% estimated breeding values (EBVs) in the validation dataset of Lnsd2.

| Trait[1] | N | Top 20% | Bottom 20% | *P*-Value |
|---|---|---|---|---|
| EBV | 128 | −0.02 ± 0.01 | 0.04 ± 0.01 | <0.01** |
| Lnsd2 | 128 | 0.84 ± 0.31 | 0.87 ± 0.34 | 0.26 |
| MY305, kg | 128 | 8,814.02 ± 1,876.79 | 8,529.54 ± 2,044.03 | 0.12 |
| ML, kg | 128 | 195.80 ± 126.37 | 218.72 ± 151.70 | 0.09 |
| NML, time | 128 | 3.53 ± 1.35 | 3.70 ± 1.42 | 0.16 |
| TDML, d | 128 | 68.60 ± 27.08 | 69.72 ± 27.58 | 0.74 |
| MLP, % | 128 | 2.27 ± 1.55 | 2.79 ± 2.32 | 0.02* |
| AFC, d | 128 | 698.08 ± 35.77 | 715.91 ± 86.65 | 0.03* |
| AFS, d | 128 | 415.23 ± 11.51 | 400.60 ± 22.71 | <0.01** |
| IFL_H, d | 119 | 14.92 ± 35.05 | 32.00 ± 54.43 | <0.01** |
| ICF, d | 125 | 66.48 ± 7.18 | 64.60 ± 6.60 | 0.03* |
| Lon1, d | 65 | 380.00 ± 60.21 | 365.02 ± 10.35 | 0.03* |
| Lon2, d | 22 | 664.91 ± 114.05 | 668.77 ± 135.81 | 0.54 |
| UDDE | 105 | 0.28 ± 0.46 | 0.30 ± 0.46 | 0.43 |
| REPR | 105 | 0.19 ± 0.39 | 0.17 ± 0.38 | 0.68 |
| METB | 119 | 0.03 ± 0.18 | 0.04 ± 0.20 | 0.38 |
| DIGS | 112 | 0.02 ± 0.16 | 0.02 ± 0.14 | 0.65 |

[1]EBV, estimated breeding value; Lnsd2, log-transformed standard deviation of milk deviations based on the lactation when removing first and last 10 DIM; MY305, 305 days milk yield; ML, sum of the milk yield which dropped in all fluctuation phases in a lactation; NML, number of milk loss events; TDML, total number of days for milk loss per lactation; MLP, the percentage of ML to MY305; AFC, age at first calving in heifers; AFS, age at first insemination in heifers; IFL_H, interval from first to last insemination in heifers; ICF, interval from calving to first insemination; Lon1, the days from the first calving to the end of the first lactation or culling; Lon2, the days from the first calving to the end of the second lactation or culling; UDDE, udder health; REPR, reproductive disorders; METB, metabolic disorders; DIGS, digestive disorders.

by most traits towards less milk loss, better productive performance, and lower disease incidence in the top 20% group (more resilient animals). For instance, ML was 22.92 kg lower, MY305 was 284.48 kg higher and UDDE was 2% lower in the top 20% group. Nevertheless, AFS, ICF, Lon2, and REPR showed a more favourable trend in the bottom 20% group.

# 4 Discussion

## 4.1 Analyses of longitudinal data

Traits with repeated records over time for the same individual are known as longitudinal traits (Ning et al., 2018; Oliveira et al., 2019), which can be expressed as a series of independent continuous functions (Pletcher and Geyer, 1999; Oliveira et al., 2019). When selecting a longitudinal trait to analyze resilience in cattle, there are several points to be considered. Firstly, the trait should be susceptible to monitorable fluctuations by environmental disturbances. Secondly, the time interval between record points should be less than the duration of the fluctuation (Mehrabbeik et al., 2021), otherwise the short-term fluctuations will not be captured. In the process of recording daily milk yield by automatic monitoring equipment, missing data would inevitably occur due to errors in identifying cows or recording. For lactations missing more than 10 consecutive days, it was assumed that the true fluctuations in that phase could not be known. Afterwards, the quality of raw records is an important factor. The two steps in the quality control which removed the most records were the number of records within a lactation and lactation curve, which caused removal of 19,836 and 9,780 lactations, respectively (the total number of lactations removed was 40,183). In our study, 42.2% lactations did not meet the threshold for the number of records within a lactation. Culling, damage to monitoring equipment, diseases, and a variety of other unknown reasons can result in missing records. In particular, when cows are not milked due to disease, data imputation in milk loss period cannot accurately reflect the disturbance. Matching the two types of data, milk yield and environment disturbance, could be beneficial when possible. To ensure the lactation curve, extreme values and $R^2$ were controlled. This step is important because the empirical lactation model tends to be a quantitative representation of the phenomenon and therefore would be more susceptible to extremes (Macciotta et al., 2011). In addition to the observed phenotypic outliers, some daily milk yield records derived from the imputation analyses were also out of the expected range. The DIM 1–4 days and DIM 305 of data used for the data imputation were based on normalized values for different cluster groups. When the standard deviation of non-missing milk yield is too high, the values converted back to the original scale would likely be negative or too high. As a result, the proportion of extremes in our study was increased. The low $R^2$ for some lactation curves may be due to the atypical shape, as discussed in Section 4.2. Finally, the methodology for analyzing longitudinal data should be precisely tailored to the characteristics of the data. In our study, a weighted moving average filter was established to effectively eliminate the effects of random fluctuations in the raw data (Poppe et al., 2020). This study serves as an example of the analysis of longitudinal data and provides a reference for the future processing of continuous datasets.

## 4.2 Lactation curve and perturbations

Lactation curve is a mathematical model used to describe the trend of daily milk yield in a lactation (Kong et al., 2018; Oliveira et al., 2019). There are individual differences in the shape of this variation in daily milk yield, and even atypical lactation curves, where the shape is completely opposite to the standard curve or shows a linear shape with no peak yield (Lee et al., 2020). These curves account for about 10–20% in a population (Macciotta et al., 2005; Lee et al., 2020). In previous studies, atypical curves have often been ignored or their influence on the overall dataset has been diluted by using average values. Approximately 5.3% of lactations in our population [groups (c), (e), and (f)] exhibited atypical patterns and a greater tendency for atypical curves in low parity cows. Lee et al. (2020) clustered lactation curves using the K-medoids and the proportion of atypical curves was 18%, with average parity of 1.25, which is similar to our results. The reason for identifying fewer atypical curves in our study may be the differences in the type of raw data. Peak yield can easily be missed by using only DHI records to estimate lactation curve (Rekik and Gara, 2004; Macciotta et al., 2005), making the curves unrepresentative of trends of the true daily milk yield. Meanwhile, when there is an abnormal record (too high or too low) in the DHI records, it can have a large influence on the lactation curve. The high milk loss and high Lnsd2 of group (c), (e), and (f) indicated that the occurrence of atypical lactation curves is unfavorable for milk production and resilience breeding. The significant effects of the four lactation curve models on resilience indicators also indicate individual differences in the lactation trend. The identification and application of atypical curves should also be considered in future studies.

In our study, ELC and milk loss were estimated through an iterative procedure. The inclusion of milk loss in the lactation curve is a reasonable modification of the model based on production reality. The assumption is that there is a theoretical production potential for cows that corresponds to their genetic potential, which may not be fully expressed due to various environmental disturbances (Ben et al., 2021). The high variability in ML suggests that fluctuations in daily milk yield may help to identify environmental disturbances and reflect their ability to adapt and resilience to disturbances (Dunne et al., 2018). The maximum TDML was 221 days, which means that 221 days of a lactation had not reached lactation potential, and the maximum DML was 167 days, indicating that the longest period of milk loss in the population was 167 days. This is uncommon and may be related with the low level of milk yield that do not match the trend of milk yield before milk loss occurred. This could also be an issue with the ELC fitted. Although we used four lactation curve models expecting to restore the lactation potential as much as possible, the models still do not fit the data perfectly. Therefore, milk loss can be further addressed by setting thresholds or changing to a more optimal model in future studies. Adriaens et al. (2021b) detected 3.8 perturbations within a lactation, with milk losses ranging from 0 to 29%, using a threshold of five consecutive days of milk losses. Ben et al. (2021) considered each negative deviation as a perturbation and obtained milk losses ranging from 2 to 19%. Milk loss does not occur with the same frequency at all lactation stages. As we set a higher threshold, disturbances of longer duration such as clinical health events (LeBlanc, 2020; Adriaens et al., 2021b) and reproductive events (Strucken et al., 2015) were likely the main reasons for the high probability of milk loss in early and late lactation. The threshold could affect the number of perturbations identified, with more milk loss periods detected when thresholds are reduced. However, it is less directional and may detect decreases in milk yield which last for a short number of days without any cause, which is potentially not what we expect. Therefore, additional studies on milk loss thresholds need to be performed, such as milk yield per shift and specific environmental disturbances.

## 4.3 Resilience in Holstein cattle

In the case of livestock, resilience is defined as "the capacity of the livestock to maintain their level of production under environmental disturbances or to recover rapidly to the state existing before exposure to a disturbance" (Colditz and Hine, 2016; Berghof et al., 2018). Several concepts related to resilience have been discussed in many studies: robustness (De La Torre et al., 2015), tolerance (Bishop, 2012), environmental sensitivity (Ehsaninia et al., 2019, 2020), and plasticity (Debat and David, 2001). Despite the wealth of research in humans (Feder et al., 2019), studies on resilience in livestock are still incipient and there is no clear distinction between these definitions in terms of similarities and differences and their research strategies. It is important to note that we focus on "general" resilience which is a comprehensive breeding goal and not only "specific" resilience (e.g., disease resilience, climatic resilience). When stressors exceed the threshold of the "general" resilience, the homeostasis of the livestock system is disrupted (van Dixhoorn et al., 2018) and performance will be forced to shift from one equilibrium to another (Nazarimehr et al., 2020). In this study, there was a decline in daily milk yield until it reduced to a minimum. Close to the minimum point, the rate of decline in daily milk yield will become slower, a phenomenon known as critical slowing down (Ren and Watts, 2015; Nazarimehr et al., 2020; Mehrabbeik et al., 2021). As a consequence of this phenomenon, the deviation between actual and expected milk yield and its variation will increase, and the autocorrelation between subsequent states will become increasingly tight (Ren and Watts, 2015; Scheffer et al., 2018; van Dixhoorn et al., 2018). Therefore, the standard deviation (Lnsd), autocorrelation (Ra), and skewness (Ske) of the deviations have been proposed as resilience indicators in a complex dynamic system. Lnsd reflects the amplitude of fluctuation in daily milk yield within a lactation. We applied Ln transformation to the standard deviation to make the indicator normally distributed. The smaller the Lnsd, the lower the fluctuation in milk yield, indicating that the cow is less susceptible to environmental disturbance and therefore, more resilient. Ra reflects the length and rate of variation of the milk loss within a lactation. Resilient cows have greater independence between milk yield from successive DIM and therefore, smaller Ra. Ske reflects the balance of positive and negative deviations, with a high Ske indicating low milk loss. In addition, genetic selection for resilience by milk loss traits to reduce milk loss of cows in the general environment seems to be a potential direction which we explored in this study. Several studies have proposed other potential indicators, such as the rate of recovery (Adriaens et al., 2021b), the slope of the reaction norm (Kause and Odegård, 2012), and the cross-correlation between different longitudinal traits (Scheffer et al., 2018).

The best resilience indicators should have high heritability and be genetically correlated with better production, reproduction, longevity, and health traits (Poppe et al., 2020). When high heritability resilience indicator is applied for genetic selection, the accuracy of estimated breeding value as well as genomic selection can be ensured, while genetic antagonism resulting from selection for resilience causing a decrease in milk yield can be avoided as much as possible. Therefore, the selection of appropriate resilience indicators in this study was based on heritability and genetic correlation with economically important traits. Genetic correlations among the four milk loss traits were all positive. Higher ML, more NML, longer TDML, and higher MLP tended to coincide, which showed the overall consistency of milk loss traits. These traits represent the fluctuation of daily milk yield from different perspectives when cows face environmental perturbations. However, the highest heritability estimate was only 0.06 (ML) among milk loss traits which is low. In this context, using milk loss traits for breeding is less efficient. Milk loss traits are favorably genetically associated with several production, reproduction, longevity, and health traits, and in particular the high positive genetic correlations between ML and UDDE and REPR indicate that these health traits might be major causes of fluctuations (decreases) in daily milk yield. There was no clear pattern of genetic correlation between milk loss traits and economically important traits, and the accuracy of the correlation estimates was poor, with standard errors higher than estimates in some cases which were on average higher than the standard errors for the variability trait. This is unfavorable for the genetic selection for resilience through milk loss traits. The high standard errors might be due to the small data size used to estimate genetic correlations, and the complex distribution of phenotypes in different traits, especially for health traits. The low incidence would result in imbalanced binary phenotypes which might also have obstructed the accurate estimation of genetic correlations. A larger data size is required to further determine the relationship between milk loss traits and economically important traits. The genetic correlations between milk loss traits and MY305 were not consistent. The negative genetic correlations between NML, TDML, and MLP and MY305 indicated that fewer milk losses, shorter milk loss duration, and lower milk loss ratios all contributed to higher milk production, as expected. In contrast, the positive genetic correlation between ML and MY305 might be due to scale effects. When high yielding cows experience the same extent of environmental disturbances as low yielding cows, and milk production drops by the same percentage, the absolute value of milk loss is greater in high yielding cows and therefore, ML tends to be greater in high yielding cows. Nevertheless, the absolute amount of ML is important, and it is more necessary to minimize milk loss on high yielding cows to improving herd profitability, rather than focusing on the relative percentage of ML. Therefore, as new traits directly related to milk yield, milk loss traits should be further evaluated, especially using complete datasets with less missing records.

In this study, four periods of variability traits showed different genetic characteristics. The heritability of Lnsd was higher than the heritability of ML, whereas the heritabilities of Ra and Ske were much lower than that of ML. Poppe et al. (2020) obtained heritability estimates of 0.08–0.10 for Ra (higher than this study) and 0.01–0.02 for Ske (lower than this study). The lower heritability for Ra in this study may be due to the establishment of the 2-sided weighted moving average filter which might have removed part of the variability from the deviations. This approach resulted in more similarity between the deviations of successive DIMs, but the natural correlation was broken. The higher heritability for Ske was due to quality control. Ske was too sensitive to extreme milk yield (Poppe et al., 2020). In our study, Ske was more stable and representative due to the strict quality control and fitting procedures. Although these three variability traits referred to different aspect of resilience by definition, the moderate to high genetic correlations between the three highest heritability variability traits (Lnsd2, Ra1, and Ske1) showed that they contain overlapping information on resilience. Lnsd2, which characterizes the amplitude of fluctuation, is also representative of the information about the length of milk loss periods (as presented by Ra) and the negative deviations of milk loss (as presented by Ske). Ra and Ske also provide research value and characterize specific information about resilience. Berghof et al. (2018) pointed that a higher Ra was expected to indicate a slower recovery. However, the results of our study do not provide information on this aspect and individual milk loss require further validations. The reasons for differences in heritability of Lnsd are not the same for various periods. The lactation curves were poorly fitted during the early and late lactation because the raw data were more severely missing in these two periods, particularly when DIM was 1–10 and 296–305. Meanwhile, due to the high sensitivity of the Ali-Schaeffer model to data distribution (Melzer et al., 2017), the curves may take on an abnormally shape when there are episodic extreme values in the records during the early and late lactation. Therefore, calculating variability traits from entire lactation gave poor results. In addition, it may also be inappropriate to use only peak lactation period data due to the lower frequency of milk loss in mid-lactation. For the fourth period, as DIM with positive deviations were not included in the calculation, Lnsd would be based on the negative deviations. However, the number of negative deviations within each lactation is not the same, which results in the calculation of Lnsd not being based on the same scale of data volume and comparability becomes poor. For instance, when only 1 day is in negative deviation, the standard deviation is zero regardless of the amount of milk loss. Therefore, Lnsd2, which has the highest heritability, is the most suitable as a single resilience indicator.

Lower Lnsd2 was correlated with lower milk loss, better reproductive performance, and lower disease incidence at the genetic level with the smaller standard errors than other resilience indicators. The results of validation for Lnsd2 also supported this trend, although the results of the *t*-test were not all statistically significant. These results supported Lnsd2 as a potential resilience indicator. The moderate to high genetic correlations of Lnsd2 with milk loss traits indicate that Lnsd2 can characterize most aspects of milk loss with high genetic correlation of 0.96. Genetic selection for resilience by Lnsd2 is almost completely representative of selection directly by ML and is more efficient. Among the reproduction traits, AFS was less genetically correlated because the age at first insemination tends to be consistent in the herd and phenotypic variation is smaller than other reproduction traits (as presented in Supplemental Table S1). In contrast, all other reproduction traits associated with insemination showed significant genetic correlations with Lnsd2, indicating a strong effect of insemination success on daily milk yield. The genetic correlation between Lnsd2 and UDDE was 0.87. Thus, it is possible that a large part of fluctuations is caused by mastitis. Mastitis-associated milk losses have a large impact on milk yield and herd sustainability. Adriaens et al. (2021a) indicated that milk losses ranged from 38.4 to 215.6 kg within -5–30 days around the first treatment of mastitis. Resilience indicators based on variability in milk yield might reflect resistance to mastitis. However, METB and DIGS were negatively genetically correlated with Lnsd2, in contrast to UDDE and REPR, which was not expected. This might be a statistical artifact. In this study, METB included milk fever, ketosis, and displacement of abomasum which is mainly concentrated in early lactation, and ML and Lnsd2 are lower in early lactation than mid and late lactation. This might have caused the misleading impression that ML and Lnsd2 were less in cows which had METB. Meanwhile, the incidence of UDDE, PRER, METB, and DIGS in the population was 29.2%, 10.7%, 6.5%, and 2.0%, respectively. The imbalance in the raw data for the two binary traits (METB and DIGS) also affected the genetic correlation accuracy and was the main reason for the high SE of the genetic correlation estimates for DIGS. Poppe et al. (2020, 2021a, 2021b, 2021c) used moving average, moving median, Wilmink model, and quantile regression models on raw daily milk yield to explore and validate the variance of deviation, autocorrelation, and skewness of daily milk yield, and the results similarly demonstrated the potential of the variance as resilience indicator. A major difference between our study and theirs was how the lactation curves were fitted. A single longitudinal trait is unlikely to be sensitive to all environmental disturbances. When resilience indicators are defined using other longitudinal traits (e.g., feed intake, activity level), additional resilience mechanisms might be captured. Poppe et al. (2022a) showed that fluctuations on daily step count data are more sensitive to hoof health, fertility, and body condition score. Therefore, the use of multiple high-throughput monitoring data to study resilience in dairy cattle can avoid a heavy reliance on a single trait (milk yield) and be more useful to herds in determining and breeding more resilient cows.

There was a high positive genetic correlation between Lnsd2 and MY305 and longevity traits, indicating that more productive cows tend to be less resilient. The negative correlation between resilience and milk yield may be explained based on the "Resource Allocation Theory" (Rendel, 1963). High-producing cows tend to have fewer resources to resist environmental disturbances due to the high demand for resources for milk production. As a result, high production leads to lower resilience. Moreover, cows with high milk yield have an advantage against active culling in the herd and therefore tend to have a higher productive life (Hu et al., 2021b; Zhang et al., 2021), which might explain the lower longevity of more resilient cows. Therefore, when we improve resilience through genetic selection on resilience indicator, we should also consider milk production, the main breeding goal of dairy farming, to develop a balanced selection index for sustainable production and balanced breeding. Resilience is a comprehensive trait and its economic value is not only related to production, health, and functional traits, but also has additional economic values which are not included in the current breeding goal. For instance, high resilient cows can reduce the cost of disease treatment and human costs for herd. It would be one of the directions of our research to find evidence for Lnsd2 as a breeding target for the next generation of more resilient animals through economic analyses. In summary, the results of the genetic analyses show the high potential and merit of continuous monitoring milk records for deriving novel resilience indicators in dairy cattle breeding. Also, the genetic analyses and phenotypic validation led to the selection of Lnsd2 as the best indicator of resilience in Chinese Holstein cattle.

## 5 Conclusion

The translation of daily milk yield into fluctuations and milk loss based on ELC enables the evaluation of phenotypic and genetic responses of cows to environmental perturbations and the ability of cows to cope with perturbations. Although heritability estimates for milk loss traits are low, there is still variability which reflect variation in daily milk yield as well as the effects of environmental disturbances on cows. Log-transformed standard deviation of milk yield deviations when removing the first and last 10 DIM (Lnsd2) had the highest heritability and was favorably genetically associated with several milk loss, reproduction, longevity, and health traits, while the antagonistic relationship between resilience and milk production indicted the necessity of balanced breeding when improving resilience. In summary, Lnsd2 is recommended as the best resilience indicator among the ones evaluated in this study for genetically improving resilience in Holstein cows. This study also shows the potential of using high frequency automatic monitoring of daily milk yield to characterize and identify the milk yield dynamics during perturbations, which can be used for on-farm monitoring and precision management.

## Data availability statement

The data analyzed in this study is proprietary. Requests to access the datasets used should be directed to YW: wangyachun@cau.edu.cn.

## Ethics statement

Ethical review and approval were not required because all the datasets used were generated as part of routine dairy cattle management and routine genetic evaluations. Thus, no additional animal handling or experiment was performed specifically for this study.

## Author contributions

AW, LB, and YW designed the study. AW, LB, and HZ led the manuscript preparation. AW performed all the data analyses. RS and LZ provided support for the data analyses. DL and GG provided support for the collection of the phenotypic datasets. All authors contributed to the article and approved its final version.

## Funding

This study was funded by National Key R&D Program of China (2021YFD1200903); the Key Research Project of Ningxia Hui Autonomous Region (2022BBF02017), the earmarked fund for CARS-36; and, the Program for Changjiang Scholar and Innovation Research Team in University (IRT_15R62).

## Conflict of interest

DL was employed by Hebei Sunlon Modern Agricultural Technology Co., Ltd. and GG was employed by Beijing Sunlon Livestock Development Co., Ltd.

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1031557/full#supplementary-material

## References

Adriaens, I., Van Den Brulle, I., D'Anvers, L., Statham, J. M. E., Geerinckx, K., De Vliegher, S., et al. (2021b). Milk losses and dynamics during perturbations in dairy cows differ with parity and lactation stage. *J. Dairy Sci.* 104, 405–418. doi:10.3168/jds.2020-19195

Adriaens, I., Van Den Brulle, I., Geerinckx, K., D'Anvers, L., De Vliegher, S., and Aernouts, B. (2021a). Milk losses linked to mastitis treatments at dairy farms with automatic milking systems. *Prev. Vet. Med.* 194, 105420. doi:10.1016/j.prevetmed.2021.105420

Ali, T. E., and Schaeffer, L. R. (1987). Accounting for covariances among test day milk yields in dairy-cows. *Can. J. Anim. Sci.* 67, 637–644. doi:10.4141/cjas87-067

Aranganayagi, S., and Thangavel, K. (2007). "Clustering categorical data using silhouette coefficient as a relocating measure," in *Iccima 2007: International conference on computational intelligence and multimedia applications* (India: Sivakasi), 13. doi:10.1109/ICCIMA.2007.328

Ben, A. A., Puillet, L., Gomes, P., and Martin, O. (2021). Lactation curve model with explicit representation of perturbations as a phenotyping tool for dairy livestock precision farming. *Animal* 15, 100074. doi:10.1016/j.animal.2020.100074

Berghof, T. V. L., Poppe, M., and Mulder, H. A. (2018). Opportunities to improve resilience in animal breeding programs. *Front. Genet.* 9, 692. doi:10.3389/fgene.2018.00692

Bishop, S. C. (2012). A consideration of resistance and tolerance for ruminant nematode infections. *Front. Genet.* 3, 168. doi:10.3389/fgene.2012.00168

Bouraoui, R., Lahmar, M., Majdoub, A., Djemali, M., and Belyea, R. (2002). The relationship of temperature-humidity index with milk production of dairy cows in a Mediterranean climate. *Anim. Res.* 51, 479–491. doi:10.1051/animres:2002036

Brito, L. F., Bedere, N., Douhard, F., Oliveira, H. R., Arnal, M., Peñagaricano, F., et al. (2021). Review: Genetic selection of high-yielding dairy cattle toward sustainable farming systems in a rapidly changing world. *Animal* 15, 100292. doi:10.1016/j.animal.2021.100292

Brito, L. F., Oliveira, H. R., McConn, B. R., Schinckel, A. P., Arrazola, A., Marchant-Forde, J. N., et al. (2020). Large-scale phenotyping of livestock welfare in commercial production systems: A new frontier in animal breeding. *Front. Genet.* 11, 793. doi:10.3389/fgene.2020.00793

Colditz, I. G., and Hine, B. C. (2016). Resilience in farm animals: Biology, management, breeding and implications for animal welfare. *Anim. Prod. Sci.* 56, 1961–1983. doi:10.1071/AN15297

De La Torre, A., Recoules, E., Blanc, F., Ortigues-Marty, I., D Hour, P., and Agabriel, J. (2015). Changes in calculated residual energy in variable nutritional environments: An indirect approach to apprehend suckling beef cows' robustness. *Livest. Sci.* 176, 75–84. doi:10.1016/j.livsci.2015.03.008

Debat, V., and David, P. (2001). Mapping phenotypes: Canalization, plasticity and developmental stability. *Trends Ecol. Evol.* 16, 555–561. doi:10.1016/S0169-5347(01)02266-2

Dunne, F. L., Kelleher, M. M., Walsh, S. W., and Berry, D. P. (2018). Characterization of best linear unbiased estimates generated from national genetic evaluations of reproductive performance, survival, and milk yield in dairy cows. *J. Dairy Sci.* 101, 7625–7637. doi:10.3168/jds.2018-14529

Ehsaninia, J., Hossein-Zadeh, N. G., and Shadparvar, A. A. (2020). Estimation of genetic parameters for micro-environmental sensitivities of production traits in Holstein cows using two-step method. *Anim. Prod. Sci.* 60, 752–757. doi:10.1071/AN18687

Ehsaninia, J., Hossein-Zadeh, N. G., and Shadparvar, A. A. (2019). Estimation of genetic variation for macro- and micro-environmental sensitivities of milk yield and composition in Holstein cows using double hierarchical generalized linear models. *J. Dairy Res.* 86, 145–153. doi:10.1017/S0022029919000293

Elgersma, G. G., de Jong, G., van der Linde, R., and Mulder, H. A. (2018). Fluctuations in milk yield are heritable and can be used as a resilience indicator to breed healthy cows. *J. Dairy Sci.* 101, 1240–1250. doi:10.3168/jds.2017-13270

Feder, A., Fred-Torres, S., Southwick, S. M., and Charney, D. S. (2019). The biology of human resilience: Opportunities for enhancing resilience across the life span. *Biol. Psychiatry* 86, 443–453. doi:10.1016/j.biopsych.2019.07.012

Friggens, N. C., Blanc, F., Berry, D. P., and Puillet, L. (2017). Review: Deciphering animal robustness. A synthesis to facilitate its use in livestock breeding and management. *Animal* 11, 2237–2251. doi:10.1017/S175173111700088X

Friggens, N. C., Brun-Lafleur, L., Faverdin, P., Sauvant, D., and Martin, O. (2013). Advances in predicting nutrient partitioning in the dairy cow: Recognizing the central role of genotype and its expression through time. *Animal* 7, 89–101. doi:10.1017/S1751731111001820

Friggens, N. C., Duvaux-Ponter, C., Etienne, M. P., Mary-Huard, T., and Schmidely, P. (2016). Characterizing individual differences in animal responses to a nutritional challenge: Toward improved robustness measures. *J. Dairy Sci.* 99, 2704–2718. doi:10.3168/jds.2015-10162

Garcia-Baccino, C. A., Marie-Etancelin, C., Tortereau, F., Marcon, D., Weisbecker, J. L., and Legarra, A. (2021). Detection of unrecorded environmental challenges in high-frequency recorded traits, and genetic determinism of resilience to challenge, with an application on feed intake in lambs. *Genet. Sel. Evol.* 53 (1), 4. doi:10.1186/s12711-020-00595-x

Guarini, A. R., Lourenco, D. A. L., Brito, L. F., Sargolzaei, M., Baes, C. F., Miglior, F., et al. (2019). Genetics and genomics of reproductive disorders in Canadian Holstein cattle. *J. Dairy Sci.* 102 (2), 1341–1353. doi:10.3168/jds.2018-15038

Guo, G., Guo, X., Wang, Y., Zhang, X., Zhang, S., Li, X., et al. (2014). Estimation of genetic parameters of fertility traits in Chinese Holstein cattle. *Can. J. Anim. Sci.* 94, 281–285. doi:10.4141/CJAS2013-113

Hu, H., Mu, T., Ma, Y., Wang, X., and Ma, Y. (2021b). Analysis of longevity traits in Holstein cattle: A Review.The genetic analysis of tolerance to infections: A review. *Front. Genet. Genet.* 123, 695543695543. doi:10.3389/fgene.2021.695543

Hu, L., Brito, L. F., Abbas, Z., Sammad, A., Kang, L., Wang, D., et al. (2021a). Investigating the short-term effects of cold stress on metabolite responses and metabolic pathways in inner-Mongolia Sanhe cattle. *Animals.* 11 (9), 2493. doi:10.3390/ani11092493

Kong, L., Li, J., Li, R., Zhao, X., Ma, Y., Sun, S., et al. (2018). Estimation of 305-day milk yield from test-day records of Chinese Holstein cattle. *J. Appl. Animal Res.* 46, 791–797. doi:10.1080/09712119.2017.1403918

Konig, S., and May, K. (2019). Invited review: Phenotyping strategies and quantitative-genetic background of resistance, tolerance and resilience associated traits in dairy cattle. *Animal* 13, 897–908. doi:10.1017/S1751731118003208

LeBlanc, S. J. (2020). Review: Relationships between metabolism and neutrophil function in dairy cows in the peripartum period. *Animal* 14, s44–s54. doi:10.1017/S1751731119003227

Lee, M., Lee, S., Park, J., and Seo, S. (2020). Clustering and characterization of the lactation curves of dairy cows using K-medoids clustering algorithm. *Animals.* 10, 1348. doi:10.3390/ani10081348

Liu, A., Lund, M. S., Wang, Y., Guo, G., Dong, G., Madsen, P., et al. (2017). Variance components and correlations of female fertility traits in Chinese Holstein population. *J. Anim. Sci. Biotechnol.* 8, 56. doi:10.1186/s40104-017-0189-x

Luo, H., Brito, L. F., Li, X., Su, G., Dou, J., Xu, W., et al. (2021). Genetic parameters for rectal temperature, respiration rate, and drooling score in Holstein cattle and their relationships with various fertility, production, body conformation, and health traits. *J. Dairy Sci.* 104, 4390–4403. doi:10.3168/jds.2020-19192

Macciotta, N. P. P., Dimauro, C., Rassu, S. P. G., Steri, R., and Pulina, G. (2011). The mathematical description of lactation curves in dairy cattle. *Italian J. Animal Sci.* 10, e51. doi:10.4081/ijas.2011.e51

Macciotta, N. P. P., Vicario, D., and Cappio-Borlino, A. (2005). Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. *J. Dairy Sci.* 88, 1178–1191. doi:10.3168/jds.S0022-0302(05)72784-3

Madsen, P., Sorensen, P., Su, G., Damgaard, L., Thomsen, H., and Labouriau, R. (2006). "Dmu - a package for analyzing multivariate mixed models," in *Proceedings of the 8th world congress on genetics applied to livestock production* (Belo Horizonte, Brazil: Instituto Prociencia).

Mehrabbeik, M., Ramamoorthy, R., Rajagopal, K., Nazarimehr, F., Jafari, S., and Hussain, I. (2021). Critical slowing down indicators in synchronous period-doubling for salamander flicker vision. *Eur. Phys. J. Spec. Top.* 230, 3291–3298. doi:10.1140/epjs/s11734-021-00113-0

Melzer, N., Trissl, S., and Nurnberg, G. (2017). Short communication: Estimating lactation curves for highly inhomogeneous milk yield data of an F2 population (Charolais × German Holstein). *J. Dairy Sci.* 100, 9136–9142. doi:10.3168/jds.2017-12772

Mulder, H. A., and Rashidi, H. (2017). Selection on resilience improves disease resistance and tolerance to infections. *J. Anim. Sci.* 95, 3346–3358. doi:10.2527/jas.2017.1479

Murtagh, F., and Contreras, P. (2011). Algorithms for hierarchical clustering: An overview. *WIREs Data Min. Knowl. Discov.* 2, 86–97. doi:10.1002/widm.53

Nazarimehr, F., Jafari, S., Perc, M., and Sprott, J. C. (2020). Critical slowing down indicators. *EPL* 132, 18001. doi:10.1209/0295-5075/132/18001

Nelder, J. A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics* 22, 128. doi:10.2307/2528220

Ning, C., Wang, D., Zheng, X., Zhang, Q., Zhang, S., Mrode, R., et al. (2018). Eigen decomposition expedites longitudinal genome-wide association studies for milk production traits in Chinese Holstein. *Genet. Sel. Evol.* 50, 12. doi:10.1186/s12711-018-0383-0

Oliveira, H. R., Brito, L. F., Lourenco, D. A. L., Silva, F. F., Jamrozik, J., Schaeffer, L. R., et al. (2019). Invited review: Advances and applications of random regression models: From quantitative genetics to genomics. *J. Dairy Sci.* 102 (9), 7664–7683. doi:10.3168/jds.2019-16265

Peng, C. K., Costa, M., and Goldberger, A. L. (2009). Adaptive data analysis of complex fluctuations in physiologic time series. *Adv. Adapt. Data Anal.* 1, 61–70. doi:10.1142/S1793536909000035

Pletcher, S. D., and Geyer, C. J. (1999). The genetic analysis of age-dependent traits: Modeling the character process. *Genetics* 153, 825–835. doi:10.1093/genetics/153.2.825

Polsky, L., and von Keyserlingk, M. A. G. (2017). Invited review: Effects of heat stress on dairy cattle welfare. *J. Dairy Sci.* 100, 8645–8657. doi:10.3168/jds.2017-12651

Poppe, M., Bonekamp, G., van Pelt, M. L., and Mulder, H. A. (2021a). Genetic analysis of resilience indicators based on milk yield records in different lactations and at different lactation stages. *J. Dairy Sci.* 104, 1967–1981. doi:10.3168/jds.2020-19245

Poppe, M., Mulder, H. A., Kamphuis, C., and Veerkamp, R. F. (2021c). Between-herd variation in resilience and relations to herd performance. *J. Dairy Sci.* 104, 616–627. doi:10.3168/jds.2020-18525

Poppe, M., Mulder, H. A., van Pelt, M. L., Mullaart, E., Hogeveen, H., and Veerkamp, R. F. (2022a). Development of resilience indicator traits based on daily step count data for dairy cattle breeding. *Genet. Sel. Evol.* 54, 21. doi:10.1186/s12711-022-00713-x

Poppe, M., Mulder, H. A., and Veerkamp, R. F. (2021b). Validation of resilience indicators by estimating genetic correlations among daughter groups and with yield responses to a heat wave and disturbances at herd level. *J. Dairy Sci.* 104, 8094–8106. doi:10.3168/jds.2020-19817

Poppe, M., Veerkamp, R. F., Mulder, H. A., and Hogeveen, H. (2022b). Observational study on associations between resilience indicators based on daily milk yield in first lactation and lifetime profitability. *J. Dairy Sci.* 105, 8158–8176. doi:10.3168/jds.2021-21532

Poppe, M., Veerkamp, R. F., van Pelt, M. L., and Mulder, H. A. (2020). Exploration of variance, autocorrelation, and skewness of deviations from lactation curves as resilience indicators for breeding. *J. Dairy Sci.* 103, 1667–1684. doi:10.3168/jds.2019-17290

Rajala-Schultz, P. J., Gröhn, Y. T., and McCulloch, C. E. (1999b). Effects of milk fever, ketosis, and lameness on milk yield in dairy cows. *J. Dairy Sci.* 82, 288–294. doi:10.3168/jds.S0022-0302(99)75235-5

Rajala-Schultz, P. J., Gröhn, Y. T., McCulloch, C. E., and Guard, C. L. (1999a). Effects of clinical mastitis on milk yield in dairy cows. *J. Dairy Sci.* 82, 1213–1220. doi:10.3168/jds.S0022-0302(99)75344-0

Rekik, B., and Gara, A. B. (2004). Factors affecting the occurrence of atypical lactations for Holstein-Friesian cows. *Livest. Prod. Sci.* 87, 245–250. doi:10.1016/j.livprodsci.2003.09.023

Ren, H., and Watts, D. (2015). Early warning signals for critical transitions in power systems. *Electr. Power Syst. Res.* 124, 173–180. doi:10.1016/j.epsr.2015.03.009

Rendel, J. M. (1963). Correlation between the number of scutellar and abdominal bristles in Drosophila melanogaster. *Genetics* 48, 391–408. doi:10.1093/genetics/48.3.391

Scheffer, M., Bolhuis, J. E., Borsboom, D., Buchman, T. G., Gijzel, S. M. W., Goulson, D., et al. (2018). Quantifying resilience of humans and other animals. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11883–11890. doi:10.1073/pnas.1810630115

Shi, R., Brito, L. F., Liu, A., Luo, H., Chen, Z., Liu, L., et al. (2021). Genotype-by-environment interaction in Holstein heifer fertility traits using single-step genomic reaction norm models. *BMC Genomics* 22 (1), 193–220. doi:10.1186/s12864-021-07496-3

Silpa, M. V., König, S., Sejian, V., Malik, P. K., Nair, M. R. R., Fonseca, V. F. C., et al. (2021). Climate-resilient dairy cattle production: Applications of genomic tools and statistical models. *Front. Vet. Sci.* 8, 625189. doi:10.3389/fvets.2021.625189

Strucken, E. M., Laurenson, Y. C., and Brockmann, G. A. (2015). Go with the flow-biology and genetics of the lactation cycle. *Front. Genet.* 6, 118. doi:10.3389/fgene.2015.00118

Su, G., Lund, M. S., and Sorensen, D. (2007). Selection for litter size at day five to improve litter size at weaning and piglet survival rate. *J. Anim. Sci.* 85, 1385–1392. doi:10.2527/jas.2006-631

van Dixhoorn, I. D. E., de Mol, R. M., van der Werf, J. T. N., van Mourik, S., and van Reenen, C. G. (2018). Indicators of resilience during the transition period in dairy cows: A case study. *J. Dairy Sci.* 101, 10271–10282. doi:10.3168/jds.2018-14779

van Soest, F. J. S., Santman-Berends, I. M. G. A., Lam, T. J. G. M., and Hogeveen, H. (2016). Failure and preventive costs of mastitis on Dutch dairy farms. *J. Dairy Sci.* 99, 8365–8374. doi:10.3168/jds.2015-10561

Wang, K., Zhang, H., Dong, Y., Chen, S., Guo, G., Liu, L., et al. (2022). Definition and genetic parameters estimation for health traits by using on-farm management data in dairy cattle. *Sci. Agric. Sin.* 55, 1227–1240. doi:10.3864/j.issn.0578-1752.2022.06.014

Wankar, A. K., Rindhe, S. N., and Doijad, N. S. (2021). Heat stress in dairy animals and current milk production trends, economics, and future perspectives: The global scenario. *Trop. Anim. Health Prod.* 53, 70. doi:10.1007/s11250-020-02541-x

Warren Liao, T. (2005). Clustering of time series data—A survey. *Pattern Recognit.* 38, 1857–1874. doi:10.1016/j.patcog.2005.01.025

Wilmink, J. B. M. (1987). Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livest. Prod. Sci.* 16, 335–348. doi:10.1016/0301-6226(87)90003-0

Wood, P. D. P. (1967). Algebraic model of the lactation curve in cattle. *Nature* 216, 164–165. doi:10.1038/216164a0

Zhang, H., Liu, A., Wang, Y., Luo, H., Yan, X., Guo, X., et al. (2021). Genetic parameters and genome-wide association studies of eight longevity traits representing either full or partial lifespan in Chinese holsteins. *Front. Genet.* 12, 634986. doi:10.3389/fgene.2021.634986

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 06 December 2022
DOI 10.3389/fgene.2022.1072853

# Transcriptomic changes underlying glucocorticoid-induced suppression of milk production by dairy cows

Anna Sadovnikova[1,2], Sergio C. Garcia[3], Josephine F. Trott[1], Alice T. Mathews[1], Monica T. Britton[4], Blythe P. Durbin-Johnson[4] and Russell C. Hovey[1]*

[1]Department of Animal Science, University of California, Davis, Davis, CA, United States, [2]School of Medicine, University of California, Davis, Sacramento, CA, United States, [3]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia, [4]UC Davis Bioinformatics Core, University of California, Davis, Davis, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Asha Marie Miles,
United States Department of Agriculture (USDA), United States

REVIEWED BY
Theresa M. Casey,
Purdue University, United States
Chi Tang,
Tarim University, China

*CORRESPONDENCE
Russell C. Hovey,
rchovey@ucdavis.edu

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

RECEIVED 18 October 2022
ACCEPTED 23 November 2022
PUBLISHED 06 December 2022

CITATION
Sadovnikova A, Garcia SC, Trott JF, Mathews AT, Britton MT, Durbin-Johnson BP and Hovey RC (2022), Transcriptomic changes underlying glucocorticoid-induced suppression of milk production by dairy cows.
*Front. Genet.* 13:1072853.
doi: 10.3389/fgene.2022.1072853

Milk production by dairy cows is sensitive to increased levels of stress hormones such as glucocorticoids (GC) that also regulate the transcription of several genes required for milk synthesis. Whereas previous studies identified that an exogenous GC such as dexamethasone (DEX) transiently suppresses milk yield in several species without any pronounced effect on milk protein or fat percentage, the mechanism underlying this effect has not been established. In this study we sought to establish changes within the mammary glands of non-pregnant dairy cows in their second lactation (*n* = 3–4; 648–838 kg) following a single dose of exogenous DEX. Changes in the udder were monitored by serial biopsy of alternating quarters, concurrent with quarter-level monitoring of milk yield and composition. Dexamethasone increased serum glucose levels from 12–36 h ($p < 0.05$), reduced milk yield from 12–48 h ($p < 0.05$), increased % milk protein content at 24 h post-DEX, and transiently decreased both milk lactose and α-lactalbumin content, while not altering the level of milk fat. After 72 h, all aspects of milk production had returned to pre-treatment levels. Transcriptomic changes in the mammary glands in response to DEX were identified by RNA sequencing followed by differential gene expression analysis. Coincident with the milk yield and composition changes was the differential expression of 519 and 320 genes at 12 and 24 h after DEX (adjusted $p < 0.05$), respectively, with the return of all gene expression to baseline levels by 72 h. Among the transcriptomic changes in response to DEX, there was notable downregulation of elements in the lactose synthesis pathway, specifically *AQP3, GALE* and *LALBA* (α-lactalbumin) at 12 h, and sustained downregulation of *LALBA* at 24 h. One gene in the pathway, *UGP2*, was upregulated at 12–24 h post-DEX. This work supports the hypothesis that there is a direct relationship between the response to DEX and the concurrent suppression of milk yield due to the reduced synthesis of α-lactalbumin and lactose by the mammary epithelium. The ability of glucocorticoids to modulate the homeorrhetic requirements for glucose during stressful states concurrent with immune activation bears significance for dairy animals as well as a broad range of lactating mammals.

KEYWORDS

**dexamethasone, lactose, alpha-lactalbumin, lactation, inflammation**

# Introduction

Stress can suppress milk production by dairy animals (Romero et al., 2015; Hong et al., 2019) in association with a range of negative outcomes, including depressed feed intake and increased susceptibility to mastitis and metritis (Menta et al., 2022). The overarching stress response is mediated, in large part, by the endocrine environment including reduced responsiveness to oxytocin (Bruckmaier and Wellnitz, 2008), immunosuppression (Waller, 2000), glucose sparing, and gluconeogenesis (Sapolsky et al., 2000). Many of these changes are coordinated by increased circulating glucocorticoids (GC) that are elevated in response to stressors, including change of environment, heat stress, transport and disease (Johnson and Vanjonack, 1976).

During lactation the extreme demand for glucose by the mammary glands is part of a homeorhetic/homeostatic balance, that is, coordinated through mechanisms including elevated GC. While GC are essential for the transcription of milk protein genes by mammary epithelial cells (Casey and Plaut, 2007), the extent to which the lactating mammary glands respond to elevated GC remains unclear. Several studies have demonstrated that an acute, high dose of exogenous GC, including a synthetic GC such as dexamethasone (DEX), leads to the abrupt and transient suppression of milk production (Hartmann and Kronfeld, 1973; Shamay et al., 2000b; Babwah et al., 2013), which is more pronounced in cows than goats (Shamay et al., 2000a). Coincident with this DEX-induced suppression of milk yield was a reduction in the extraction of glucose from the circulation by the mammary glands, as determined from arterio-venous difference (Hartmann and Kronfeld, 1973). Further to these findings, Shamay et al. (2000b) identified that the reduction in milk production following DEX was associated with a specific reduction in the proportion of lactose in milk, whereas the level of protein and fat in milk was unchanged. While GC have also been implicated in the regulation of tight junction integrity (Stelwagen et al., 1998), exogenous DEX did not affect the ratio of Na/K in the milk (Shamay et al., 2000b), suggesting that the effect of DEX was not due to altered integrity of these intercellular junctions.

The pronounced and transient effects of a GC such as DEX on the synthesis and composition of milk raise questions about the mechanism(s) underlying this response, including whether it occurs through a systemic mode of action, or through local effects on the mammary glands. To this end, we sought to establish the temporal transcriptomic response within the udder of high-producing dairy cows following an acute exposure to DEX. Our data establish that a primary target of acute DEX exposure is the lactose synthesis pathway, including through the marked down-regulation of α-lactalbumin (LALBA) gene transcription.

# Materials and methods

## Animals and study design

All animal experimentation was approved by the UC Davis Institutional Animal Care and Use Committee. Four non-pregnant Holstein cows were enroled in the study (average 738.2 kg, range 648–838 kg) in their second lactation (average 55 DIM, range 40–64 DIM). None had a prior history of clinical mastitis. Cows were housed in separate pens and were bedded on rice hulls with *ad libitum* access to water and feed. Cows were fitted with rumination collars (SCR Engineers Limited, Israel).

The study period included an 8 days acclimation prior to the single administration of DEX on day 9. Four days prior to DEX, each cow was fitted with an indwelling jugular catheter that was flushed daily with saline and locked with heparinized saline (250 IU per ml). On day 9, each cow was administered a single injection of DEX (40 mg, IM, VetOne, Boise, Idaho) between 19:00 and 21:30, immediately after the first biopsy and the subsequent milking. Blood was collected into vacutainers containing potassium oxalate and sodium fluoride every 12 h out to 5 days post-DEX and was processed by centrifugation at x 2,000 g for 10 min to yield serum that was stored at -80°C.

## Feed intake, composition, and rumination

The lactating cow ration consisted of (w/w, as fed) rolled corn (40.4%), alfalfa hay (32.3%), chopped wheat hay (9.3%), cottonseed (7.7%), almond hulls (7.7%), mineral mix (1.2%), EnerGII supplement (Virtus, 1%), Strata (Virtus, 0.3%), and salt (0.2%). Each cow was offered 40 kg (as fed) of total mixed ration daily, which was delivered as 10 kg portions at 06:00, 12:00, 18:00 and 0:00. Refusals were collected and weighed daily at 18:00 for 5 days prior to, and 4 days following, administration of DEX. Proximate analysis of the ration was performed by a commercial laboratory (DairyOne, Ithaca, NY; Supplementary Table S1).

## Milk collection procedure and milk yield and composition analysis

Cows were milked twice daily, at 12 h intervals (06:00–08:00 and 18:00–20:00) using a portable milking machine that allowed for separate collection of milk from each quarter (QTR). The left rear QTR was designated as QTR1, the left front was QTR2, the left right was QTR3, and the right rear was QTR4. During the experimental period, fore- and hindmilk were collected, weighed, and sampled separately prior to, and following, administration of oxytocin (30U, IV, VetOne, Boise, Idaho), respectively. The fore- and hindmilk from each QTR was then combined and sampled in duplicate. When specified, hindmilk samples were from QTR4. When a biopsy was performed, milking and sampling of all QTR was performed immediately thereafter. After biopsy, some samples contained contaminating blood and were not analyzed for composition, namely QTR1 (0 h post DEX), QTR2 (12 h post DEX), QTR3 (24 h post DEX), and QTR4 (72 h post DEX). Duplicate milk samples were chilled on ice and supplemented with bronopol preservative (Microtabs II, Nelson- Jameson) then stored at 4°C or −20°C. Refrigerated samples were analyzed for lactose, fat, casein, total protein, solids, and somatic cell count (SCC) by a commercial laboratory (DairyOne, Ithaca, NY). Minerals (Na, K, Mg, Ca, Cl, and P) were analyzed in frozen milk samples (DairyOne, Ithaca, NY).

## Milk α-lactalbumin

The content of LALBA in milk sampled at −24, −12, 0, 12, 24, 36, 48, 60, 73, and 84 h relative to DEX was determined using a bovine LALBA ELISA (Bethyl Laboratories, Montgomery, TX United States) per the manufacturer's instructions. Concentrations were established from a standard curve generated with the provided bovine LALBA, where the resultant absorbance was measured at 280 nm using a Synergy HT microplate spectrophotometer (BioTek, Winooski, VT). All samples were assayed in triplicate.

## Serum glucose

Glucose levels were quantified using a glucose colorimetric assay kit (Cayman Chemical, Ann Arbor, MI) according to the manufacturer's instructions. The standard curve was prepared with the provided glucose that was serially-diluted. Absorbance was measured at 520 nm using a Synergy HT microplate spectrophotometer (BioTek, Winooski, VT), where all samples were assayed in triplicate.

## Mammary biopsy

One or two cores of tissue were collected using a needle biopsy tool (16 ga. Magnum, Bard, Covington, GA) that was inserted through a small incision in the skin following local anesthesia (0.125% bupivacaine, SC). Sequential biopsy across the experimental period was performed on alternating udder QTR at either time 0 (QTR1), 12 h (QTR2), 24 h (QTR3), or 72 h (QTR4) post-DEX to capture the anticipated full range of the milk yield response (Shamay et al., 2000b). Tissue cores were flash frozen in liquid nitrogen and stored at −80°C. Cows received prophylactic ampicillin (Polyflex, Boehringer Ingelheim, IM) for 3 days spanning the biopsy period.

## RNA isolation, cDNA library preparation and sequencing

Total RNA was isolated from biopsy cores (~10–50 mg tissue) from 4 cows at 0, 12, and 24 h, and from 3 cows at 72 h, using TRIzol (Invitrogen, ThermoFisher, Waltham, MA) according to the manufacturer's instructions. The integrity of the total RNA and its yield were confirmed by formaldehyde gel electrophoresis with staining (SybrSafe, Invitrogen) and UV visualization, and spectrophotometry (Nanodrop, ThermoScientific), respectively. Total RNA (5 µg) was treated with DNaseI (Zymo Research, Irvine, CA) and analyzed for quality (Experion RNA StdSens, BioRad, Hercules, CA), where all samples had an RNA integrity value greater than 8.3.

Gene expression profiling was performed using 3′Tag-RNA-Seq. Barcoded sequencing libraries were prepared using the QuantSeq FWD kit (Lexogen, Vienna, Austria) for multiplexed sequencing according to the manufacturer's recommendations, using 700 ng input RNA and 13 cycles of PCR for final library amplification. Fragment size distribution of the libraries was verified *via* microcapillary gel electrophoresis on a Bioanalyzer 2100 (Agilent, Santa Clara, CA). The library masses were quantified on a Qubit fluorometer (LifeTechnologies, Carlsbad, CA), and pooled in equimolar ratios. The final pool was treated with Exonuclease VII followed by bead clean-up to remove free primer. The pool was quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems, loaction). Fifteen libraries were sequenced per lane on a HiSeq 4000 sequencer (Illumina, San Diego, CA) with single-end 90 bp reads generating an average of 6 million reads per sample.

## Bioinformatic analyses

Raw reads were processed with HTStream (https://ibest.github.io/HTStream/) to remove adapter and low-quality sequences. On average, 0.2% of reads were removed. The trimmed reads were aligned to the *Bos taurus* UMD3.1 genome with Ensembl gene annotation release 93 using the aligner STAR v. 2.6.0c (Dobin et al., 2013) to generate raw counts per gene. On average, over 97% of the reads aligned to the *B. taurus* genome, and 76% of the trimmed reads uniquely aligned to a *B. taurus* gene. The RNA-seq data was submitted to GEO under the accession number GSE217369.

Prior to analysis, genes having an expression level across all samples of less than 4 counts per million reads were filtered out, leaving 10,241 genes. Differential expression analysis was conducted using the limma-voom Bioconductor pipeline (limma version 3.38.3, edgeR version 3.24.3, R version 3.5.1). The model used within limma was a single-factor ANOVA model for comparisons between timepoints, and a linear regression model for correlations between continuous milk characteristics and gene expression. In all limma analyses, standard errors and estimates of log fold changes were adjusted for within-cow correlations. Gene ontology (GO) enrichment analyses were conducted by Kolmogorov-Smirnov testing as implemented in the Bioconductor package topGO (version 2.32.0.). Gene enrichment analyses were also conducted with ShinyGO 0.76.3 (http://bioinformatics.sdstate.edu/go/), using a background list containing 9,745 of our 10,241 expressed genes that were annotated with a gene symbol. Linear mixed effects models were used to evaluate the correlation between module eigengenes and the phenotype variables of total milk yield, total lactose %, total casein %, total protein %, total solids %, or total fat %.

Genes that were differentially-expressed at 12 and 24 h relative to time 0 (adjusted $p$ <0 .05) were filtered by up- or down-log-fold change, then uploaded to the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8) using the *B. taurus* background list (Huang et al., 2009b; a). After selecting GOTERM_MF_DIRECT, the functional annotation chart was used where a threshold of two genes, an EASE score of 1, fold-enrichment, and false discovery rate (FDR) were selected. Enrichment terms with FDR>0.05 were removed. Genes that were differentially-expressed at 12 and 24 h post-DEX were also aligned with gene lists that were generated for the lactose synthesis pathway, or for the GO terms “tight junctions”, “inflammation”, “response to corticosteroids”, and “regulation of blood vessel diameter” (Ashburner et al., 2000; Lemay et al., 2013; Sadovnikova et al., 2021a; Gene Ontology Consortium, 2021). Upstream regulators of genes that were differentially-expressed at 12 and/or 24 h (adjusted $p$ <0 .05) were predicted using Enrichr (https://maayanlab.cloud/Enrichr/) and its Drug Signatures Database.

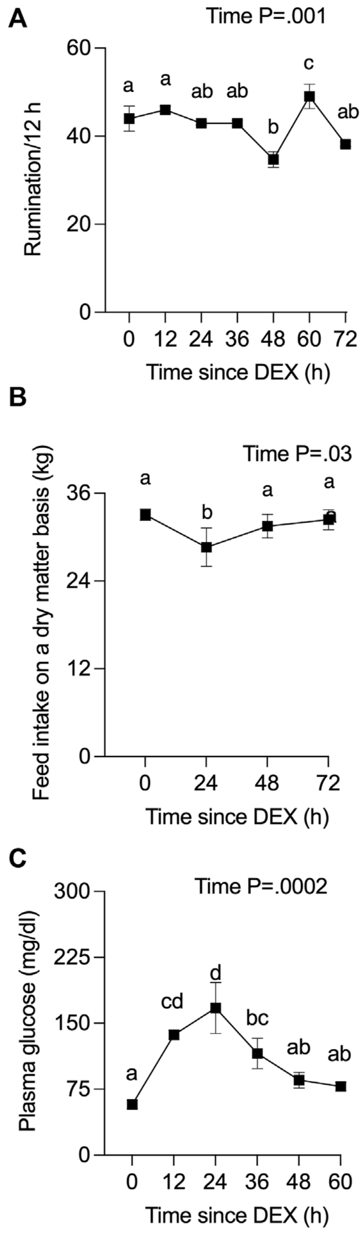


FIGURE 1
Effect of a single administration of dexamethasone (DEX) on **(A)** rumination, **(B)** daily feed intake, and **(C)** plasma glucose. Time zero represents the average of 12 data points from −24 to 0 h (h) for rumination, the average of −24 and 0 h for feed intake, and the average of −24, −12, and 0 h for plasma glucose. Data are means ± SEM ($n$ = 3 cows). a, b, c, d Means with different superscripts are different ($p$ <0 .05). The $p$-value for the main effect of time is indicated on each panel when $p$ <0.05.

## Statistical analyses

All data were analyzed as a mixed-effects model with repeated measures using Prism 9 (GraphPad Software, San Diego, CA). Data were checked for normality and homogeneity of variance and transformed where necessary. Cow was the experimental unit, with time (post-DEX) and milk type (fore-vs. hindmilk) being fixed effects. A post-hoc

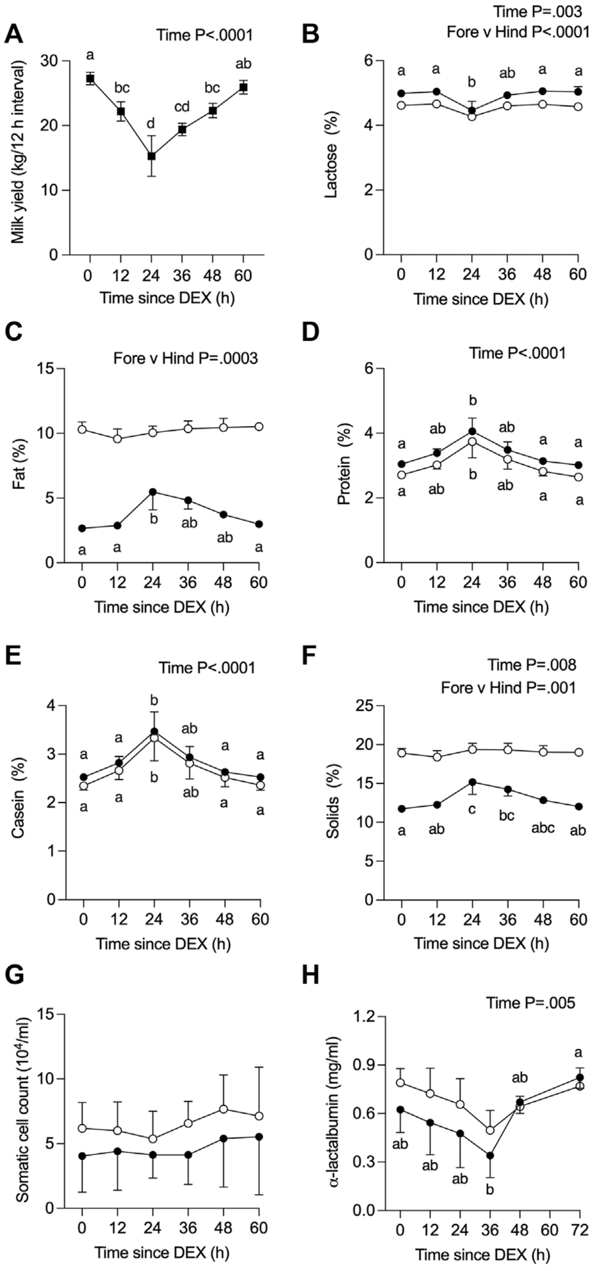


FIGURE 2
Effect of dexamethasone (DEX) on **(A)** total milk yield, or on components (%) in foremilk (black circles) and hindmilk (open circles) for **(B)** lactose, **(C)** fat, **(D)** protein, **(E)** casein, and **(F)** total solids, as well as **(G)** somatic cell count, and **(H)** α-lactalbumin (LALBA) concentration. Data are means ± SEM (*n* = 3 cows). Time zero represents an average of −36, −24, −12, and 0 h (h) relative to DEX for all components except for LALBA, where the baseline value is the average of −24, −12, and 0 h a, b, c, d Means with different superscripts are different (*p* <0.05). The *p*-value for a main effect of time or milk type is indicated on each panel when *p* <0.05.

Tukey test was performed for all data except the RNA-seq data where a post-hoc Dunnet test was performed. Significance was declared at *p* <0.05.

# Results

## Effect of DEX on rumination, feed intake, and plasma glucose levels

Rumination frequency was captured for cows from 42 h before, to 100 h after, DEX. One cow ceased ruminating by 24 h post-DEX, developed hematochezia starting at 36 h post-DEX, and was removed from the study at that time. All data for that cow are presented in Supplemental Data given that many physiological parameters up to 24 h post-DEX were notably similar to the responses recorded for other cows. The rate of rumination for the three remaining cows varied slightly over the experimental period following DEX (Figure 1A), where there was a small transient reduction in dry matter-adjusted feed intake (Figure 1B) during the 24 h after DEX that then returned to baseline. Plasma glucose levels (Figure 1C) were increased more than 2-fold by 12 h post-DEX (*p* <0 .0001), reaching a peak of 167 mg/dl at 24 h, before they returned to euglycemic values by 48 h after DEX.

## Effect of DEX on milk yield and composition

The average milk yield per 12 h interval decreased over time (*p* <0 .0001) from 27.3 kg (pre-DEX) to 15.3 kg at 24 h post-DEX, remained low (19.0 kg) at 36 h post-DEX, then returned to baseline values by 60 h post-DEX (Figure 2A). There was a parallel, transient decline in the calculated energy-corrected milk yield (Sjaunja et al., 1991), from 31.3 to 22.6 kg per 12 h interval by 24 h post-DEX (*p* < 0.01). For the one cow that was removed from the study, no foremilk could be collected by machine milking at 24 h post-DEX and required oxytocin for ejection, such that the yield and composition data reported for that cow at 24 h reflect the entire volume collected as hindmilk following oxytocin (Supplemental Data Sheet S1, Figure 2).

The composition of all fore- and hindmilk samples collected from QTR4, from time 0 to 60/72 h post-DEX, is depicted in Figure 2 for the *n* = 3 cows that completed the study. Values for the additional cow excluded from the analysis are presented as Supplemental data (Supplemental Data Sheet S1, Figure 2). The concentration of lactose in milk changed over time (*p* = 0.003), was higher in fore-*versus* hind milk (*p* <0 .0001), and in foremilk was lower at 24 h after DEX compared to 0 h (*p* <0 .05) before returning to baseline levels (Figure 2B). There was no effect of time on the fat content of milk (*p* >0.05, Figure 2C), albeit its concentration in foremilk was increased at 24 h post-DEX compared to 0 h, before returning to baseline (*p* <0.05). As expected, there was a higher fat content in hind milk (*p* = 0.0003).

The concentration of total protein (Figure 2D) and casein (Figure 2E) in milk changed over time (*p* <0.0001) and was increased at 24 h post-DEX (*p* < 0.05), without differences

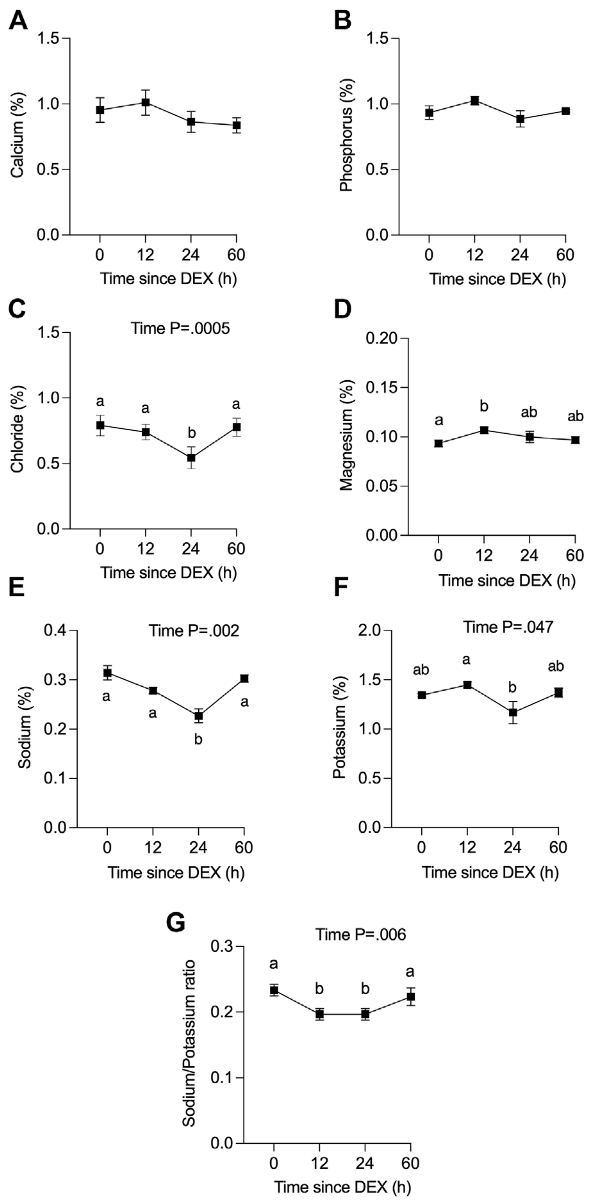


FIGURE 3
Effect of dexamethasone (DEX) on the concentration (%) of minerals in combined fore- and hindmilk. **(A)** calcium, **(B)** phosphorus, **(C)** chloride, **(D)** magnesium, **(E)** sodium, **(F)** potassium and **(G)** sodium:potassium ratio. Data are means ± SEM (n = 3 cows). Time 0 h (h) is for a single datapoint per cow (not an average of previous values). a, b Means with different superscripts are different ($p$ <0.05). The $p$-value for a main effect of time is indicated on each panel when $p$ <0.05.

TABLE 1 Pairwise comparisons for differential gene expression in response to DEX. *Adjusted $p$ <0.05, $n$ = 3 cows.

| Comparison (hours) | Number of genes* |
|---|---|
| 0 v 12 | 519 |
| 0 v 24 | 320 |
| 0 v 72 | 0 |
| 12 v 24 | 99 |
| 12 v 72 | 519 |
| 24 v 72 | 516 |

between fore- and hindmilk. The concentration of LALBA in milk changed over time ($p$ = 0.005), where in foremilk its concentration was lower at 36 h compared to 72 h after DEX (Figure 2H; $p$ = 0.007). Total solids changed over time ($p$ = 0.008) reflecting increased levels in foremilk at 24 and 36 h post-DEX (Figure 2F), whereas solids in hindmilk were unchanged. There was no change in the SCC of milk over the experimental period in either fore- or hindmilk (Figure 2G).

We also determined the level of electrolytes in QTR4 foremilk collected at 0, 12, 24, and 60 h relative to DEX (Figure 3). There was no change in the concentration of Ca or P in response to DEX (Figures 3A,B). The concentration of Cl (Figure 3C, $p$ < 0.005) and Na (Figure 3E, $p$ < 0.05) was decreased at 24 h in response to DEX compared to all other time points, while Mg was higher at 12 h compared to 0 h (Figure 3D, $p$ < 0.05). There was a small reduction in milk K at 24 h compared to 12 h (Figure 3F, $p$ < 0.05). The Na/K ratio decreased (Figure 3G) from 0.23 to a nadir of 0.20 at 12 and 24 h, then returned to baseline at 60 h ($p$ < 0.05). Data for the omitted cow are presented in Supplementary Figure S3.

## Effect of DEX on the mammary transcriptome

Relative to baseline gene expression at time 0, the expression of 519 and 320 genes was altered at 12 and 24 h after DEX, respectively (Table 1). Figure 4 shows the GO biological processes and KEGG pathways enriched in the lactating mammary gland at 12 h post-DEX (FDR<0.05), while Figure 5 shows the GO biological processes and KEGG pathways enriched at 24 h post-DEX (FDR<0.05). Of note, by 72 h post-DEX, no genes differed in their expression relative to that at time 0 (Table 1), highlighting that the mammary gland transcriptome was completely restored by 72 h after DEX.

Regression analysis across the entire study period identified seven genes (RDH12, TUBA1B, AZGP1, CEP57L1, SESN1, EPHX2, and TMEM35B) having an expression profile that associated with the change in milk yield (adjusted $p$ <0.05). By contrast, no genes had an expression profile across time that associated with the change in milk fat or lactose content. After adjusting for gene expression changes attributable to altered milk yield, the expression of one gene (ENSBTAG00000047609) retained a negative association with the change in milk fat content over time (adjusted $p$ <0.05).

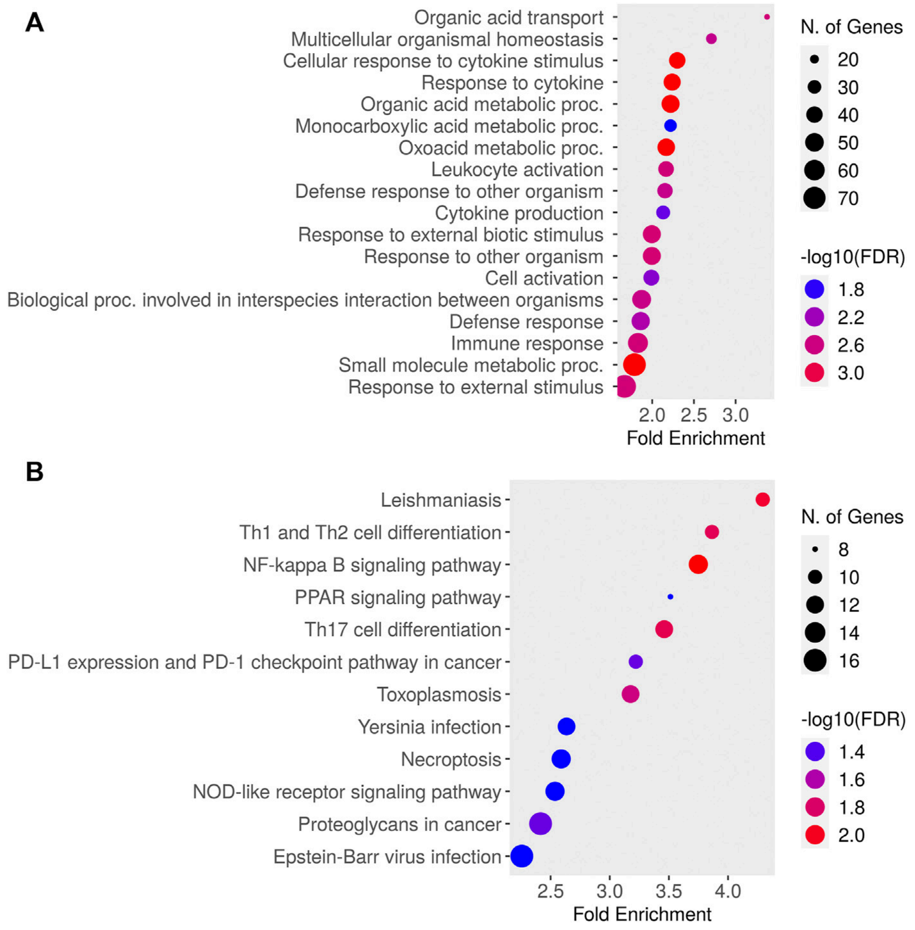


FIGURE 4
Dotplot of **(A)** Gene Ontology (GO) Biological Processes and **(B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched for genes differentially expressed in the lactating udder at 12 h post dexamethasone.

We next defined changes in the expression of individual genes specifically at 12 and 24 h after DEX. After 12 h, 519 genes were differentially expressed compared to time 0. The top ten most significant biological process ontologies (Table 2) were "immune system process", "actin cytoskeleton reorganization", "positive regulation of fat cell differentiation", "negative regulation of protein kinase activity", "heart contraction", "cGMP-mediated signaling", "female gonad development", "cellular response to cAMP", "response to bacterium", and "response to drug". By 24 h after DEX, the expression of 320 genes had changed, where the top ten biological process ontologies were "translation", "cytoplasmic translation", "formation of cytoplasmic translation initiation complex", "intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator", and "response to oxidative stress". Of the 204 genes having upregulated expression at 24 h after DEX, 136 were functionally annotated in the DAVID database. Of note among these, the expression of twenty unique genes was significantly upregulated (FDR<0.05) and belonged to three biological process ontologies: translation (RPL34, RPS27, RPS13, RPS2, EEF2, RPL13A, ENSBTAG00000047136, RPL5, MRPL10, RPL23A, RPL23, RPL24, RPL13, SLC25A3, and RPL30), translational initiation (EIF2S3, EIF3E, EIF3D, EIF3H, EIF3F, and EIF1), and formation of translation preinitiation complex (EIF2S3, EIF3E, EIF3D, EIF3H, and EIF3F).

## Effect of DEX on expression of candidate genes/pathways within the mammary gland

Given the established global effects of GC on gene expression (Ratman et al., 2013), we further examined the effect of DEX on candidate GC targets in the mammary glands including the local inflammasome (Rhen and Cidlowski, 2005), blood flow

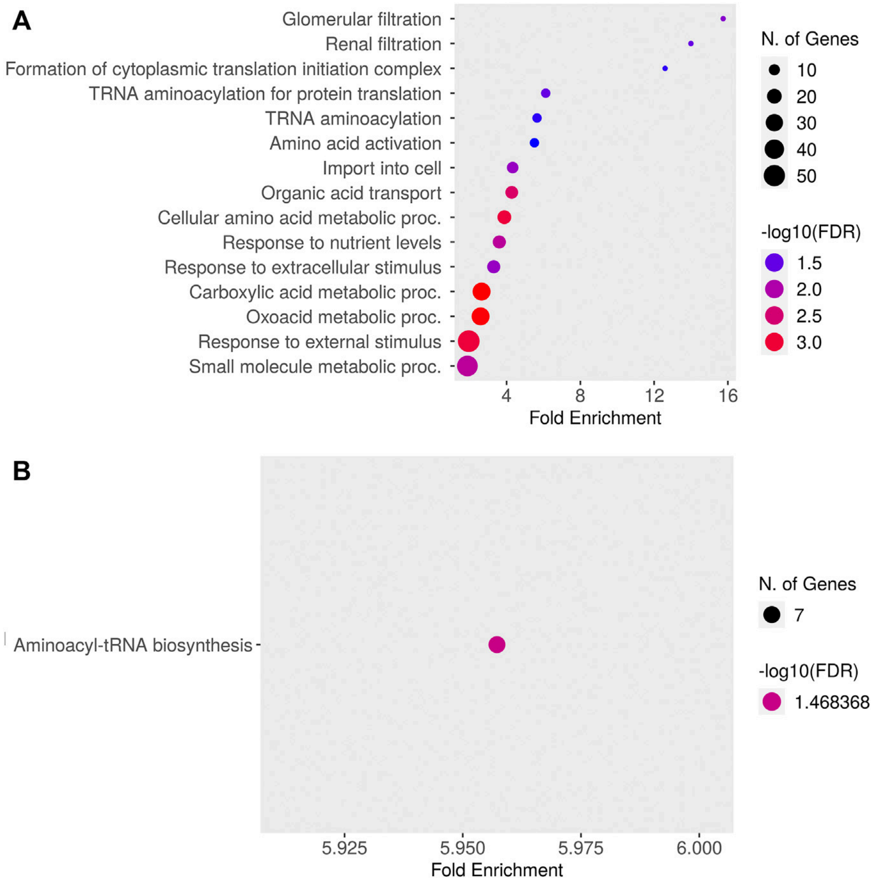


FIGURE 5
Dotplot of **(A)** Gene Ontology (GO) Biological Processes and **(B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched for genes differentially expressed in the lactating udder at 24 h post dexamethasone.

**TABLE 2** The top ten most significant biological process ontologies for genes that were upregulated or downregulated in response to DEX.

| Name | Count | Up | Down |
|---|---|---|---|
| Response to bacterium | 8 | CAV1, COLEC12, IRAK1, TICAM2, LPO | CFD, MPEG1, TLR4 |
| immune system process | 36 | TSPAN6, B4GALT1, CAV1, HSP90AB1, CD46, IMPDH2, CNOT7, COLEC12, STAT3, IRAK1, DDIT4, TICAM2, PHB, MPP1, LPO, GCNT1, FST, PTX3, SNX10 | VAV1, TLR3, BLA-DQB, TMEM106A, ALOX15, PSMB9, ENPP3, CD320, LGALS9, CASP4, CFD, LGALS1, SOD1, AQP3, MSN, PSMB10, TLR4 |
| actin cytoskeleton reorganization | 1 | — | PDLIM4 |
| positive regulation of fat cell differentiation | 1 | — | MEDAG |
| negative regulation of protein kinase activity | 7 | CAV1, HMGCR, DNAJA1, GSKIP | FABP4, TRIB2, WARS1 |
| heart contraction | 3 | SNTA1, CAV1 | SOD1 |
| cGMP-mediated signaling | 1 | PDE2A | — |
| female gonad development | 2 | FST | SOD1 |
| cellular response to cAMP | 1 | FDX1 | — |
| response to drug | 4 | — | FBP1, SOD1, PDE2A, SLC1A3 |

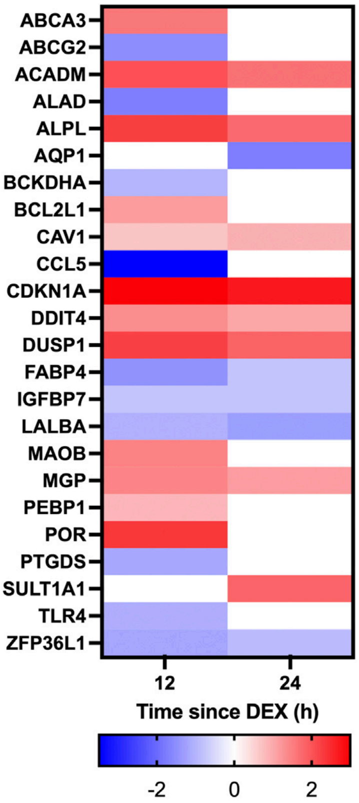


FIGURE 6
Heatmap depicting changes in differential gene expression in response to dexamethasone (DEX) for genes categorized under the gene ontology term "Response to corticosteroid". Only genes with a significant (adjusted $p$-value <0.05) change in expression for either 12 v 0 h (h), or 24 v 0 h, and a log fold change between −2 and 2 in response to DEX are shown. Gene expression data is for $n$ = 3 cows.

(Kerachian et al., 2009; Ozmen et al., 2017), and the integrity of tight junctions between the mammary epithelium (Stelwagen et al., 1998; Stelwagen et al., 1999). Within the category "response to corticosteroid" there was 10 and 5 genes having expression that was downregulated at 12 and 24 h, respectively, while 12 and 8 genes had upregulated expression at 12 and 24 h, respectively (Figure 6). Within the category "inflammation" there was 29 and 9 genes downregulated at 12 and 24 h after DEX, respectively, and 12 and 11 genes having expression that was upregulated at 12 and 24 h (Figure 7). Data for the omitted cow are presented in Supplementary Figure S4. As noted previously, there were no signs of mastitis for any of the cows during the experimental period.

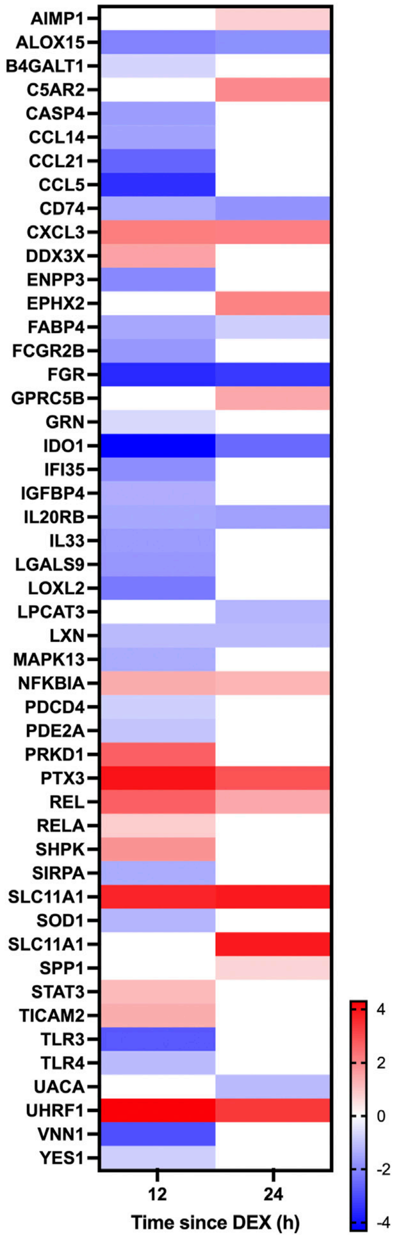


FIGURE 7
Heatmap depicting changes in differential gene expression in response to dexamethasone (DEX) for genes categorized under the gene ontology term "Inflammation." Only genes having a significant (adjusted $p$-value < .05) change in expression for either 12 v 0 h (h) or 24 v 0 h, and a log fold change between −4 and 4 in response to DEX are shown. Gene expression data is for $n$ = 3 cows.

Several genes categorized under "blood vessel diameter maintenance" were differentially regulated in response to DEX, including three that were downregulated (ADD3, FGG, and SOD1) and 5 that were upregulated (CAV1, CBS, HMGCR, KCNMB4, KCNMB4, and SNTA1) at 12 h post-DEX, of which three genes (CAV1, KCNMB4, and SNTA1) remained upregulated at 24 h. Among genes defined by the GO term "tight junction", two were downregulated (CLDN15, ESAM) and 3 were upregulated (USP53, C1QTNF5, and YBX3) at 12 h post-DEX, while only two genes (DLG3 and YBX3) were

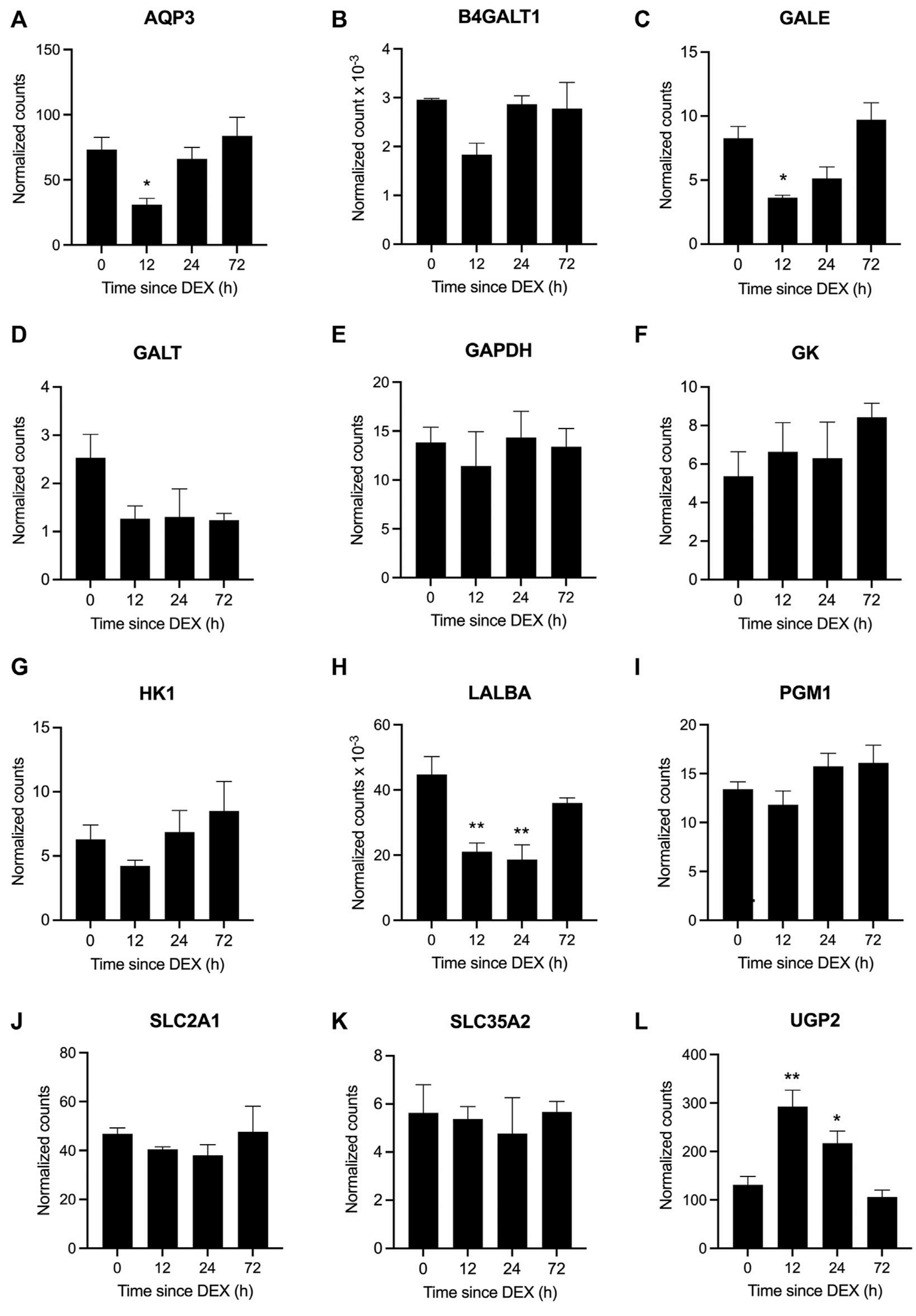


FIGURE 8
Differential gene expression in response to dexamethasone (DEX) for genes in the lactose synthesis pathway. Data represent mean ± SEM (n = 3 cows) for the genes **(A)** aquaporin 3, AQP3, **(B)** beta-1,4-galactosyltransferase 1, B4GALT1, **(C)** UDP-galactose-4-epimerase, GALE, **(D)** galactose-1-phosphate uridylyltransferase, GALT, **(E)** glyceraldehyde-3-phosphate dehydrogenase, GAPDH, **(F)** glycerol kinase, GK, **(G)** hexokinase 1, HK, **(H)** lactalbumin alpha, LALBA, **(I)** phosphoglucomutase 1, PGM1, **(J)** solute carrier family 2 member 1, SLC2A1, **(K)** solute carrier family 35 member A2, SLC35A2, and **(L)** UDP-glucose pyrophosphorylase 2, UGP2. **P* <0.05 and ***P* <0.01 versus time = 0 hours (h).

upregulated by 24 h post-DEX, where the expression of YBX3 was upregulated at both time points.

## Effect of DEX on genes in the lactose synthesis pathway

A cohort of genes involved in the lactose synthesis pathway was among those that were differentially expressed at 12 and 24 h post-DEX (Figure 8). Specifically, *AQP3, GALE,* and *LALBA* were all downregulated at 12 h (adjusted $p$ <0.05; Figures 8A,C,H), while the expression of *UGP2* was upregulated at 12 h (adjusted $p$ <0 .05). The expression of *LALBA* continued to be suppressed at 24 h post-DEX, while *UGP2* expression remained elevated (adjusted $p$ <0.05; Figure 8H,L). The expression of other genes associated with the lactose synthesis pathway including *B4GALT1, GAPDH, GALT, GK, HK1, PGM1, SLC2A1,* and *SLC35A2* was unaffected by DEX, where some genes showed numerical changes in expression that may have been statistically underpowered due to omission of the fourth cow (Figure 8B,D–G,I–K, and Supplemental Data Sheet S1, Supplementary Figure S5). Of note, the expression of another major milk protein, *CSN2* was unchanged in response to DEX (data not shown).

## Predicted upstream regulators

Not surprisingly, several GC were among the top upstream regulators identified using the Drug Signatures Database in Enrichr, where flumetasone, diflorasone and fluorometholone were the top 3 predicted upstream regulators for the lists of differentially-expressed genes at 12 and 24 h post-DEX. Indolo [3,2-b]carbazole, mastinib (AB1010), and etynodiol HL60 were the top three predicted upstream regulators for genes having downregulated expression at 12 h, while valrubicin, fenbendazole, and ofloxacin were the predicted upstream regulators for genes with downregulated expression at 24 h.

# Discussion

Here we investigated the effect of elevated systemic GC, as occurs during various states of stress, on milk synthesis and gene expression by the mammary glands of dairy cows. Consistent with earlier reports (Hartmann and Kronfeld, 1973; Shamay et al., 2000b), a single, high dose of DEX administered to multiparous cows transiently suppressed the yield and lactose content of milk, and increased its protein content, without negatively affecting its concentration of fat. Lactose synthesis returned to baseline levels by 36 h post-DEX, after a nadir at 24 h, alongside the transitional normalization of milk yield, composition, and the expression of *LALBA* and *B4GALT1*, concomitant with the restoration of euglycemia. Our transcriptomic analysis highlighted that DEX downregulated several aspects of lactose synthesis and function in the mammary epithelium, spanning from precursor uptake and hexose metabolism through to lactose synthesis, in a time-dependent manner. This suppression of lactose synthesis is in keeping with its recognized and critical role as the major osmole in milk (Sadovnikova et al., 2021a). These changes were also accompanied by a marked reduction in water secretion, where water transporters in the mammary glands localize to the capillary endothelium (AQP1) or mammary epithelium (AQP3), although limited data exist for how their expression is regulated in the mammary gland (Mobasheri and Barrett-Jolley, 2014).

Central to the reduction in lactose output was the downregulation of gene expression for several components of the lactose synthase complex (LSC) within 12 h post-DEX, including for the essential modifier protein, LALBA, which remained suppressed out to 24 h. At the same time, expression of *B4GALT1*, the enzymatic component of the LSC, was suppressed at 12 h, but was then restored by 24 h. By contrast, gene expression for another major milk protein, *CSN2*, was unchanged following DEX exposure. This acute and specific suppression of LSC activity in response to elevated GC aligns with the differential regulation of milk protein gene expression across a range of GC concentrations in rodent mammary tissue *ex vivo/in vitro* (Sadovnikova et al., 2021b). Specifically, the expression of LALBA *ex vivo/in vitro* is stimulated by low, relatively-physiological concentrations of GC, whereas high concentrations are suppressive (Ono and Oka, 1980; Nagamatsu and Oka, 1983). By contrast, the positive effect of GC on gene expression for milk proteins such as CSN2 and whey acidic protein was distinctly sigmoidal and monotonic across a range of GC concentrations (Ono and Oka, 1980; Nagamatsu and Oka, 1983). We should highlight that these reductions in *LALBA* and *B4GALT1* expression and overall lactose synthesis coincided with evidence for other coordinated changes implicated in precursor transport and hexose metabolism (Sadovnikova et al., 2021a), despite not always reaching statistical significance due to the challenged sample size within this study. For example, there were indications for reduced abundance of *GLUT1* (*SLC2A1*) at both 12 and 24 h in keeping with its established role during glucose uptake (Zhao, 2014). Likewise, the downregulated expression of *AQP3* not only aligns with its role as a water transporter but also for transporting glycerol (Hara-Chikuma and Verkman, 2006), which may have led to a reduction in its availability as an alternative precursor for galactose synthesis (Sadovnikova et al., 2021a). Similarly, a non-significant decline in *HK1* expression at 12 h post-DEX coincided with the anticipated reduction in glucose uptake and demand, where hexokinase activity controls 80% of glucose being metabolized for lactose synthesis (Xiao and Cant, 2005). At the same time, transient downregulation of *GALE* at 12 (and non-significantly at 24 h) would have reduced the accumulation of UDP-galactose being provided for lactose synthesis, thereby accumulating glucose-1-phosphate. The increased and sustained expression of *UGP2* at 12 and 24 h post-DEX would have then potentially rerouted this excess glucose-1-phosphate to UDP-glucose for alternative metabolism, such as toward glycogen synthesis. Indeed, Emerman et al. (1980) proposed that glycogen accumulation was an important shunt during lactogenesis in the absence of maximal lactose synthesis, and that stored glycogen could be recycled for lactose synthesis during its subsequent activation around parturition. Combined, these data point to a mechanism whereby elevated GC activates the rapid and targeted downregulation of lactose synthesis in the mammary glands across multiple steps, beginning at the level of gene expression, thereby affording a reversible and rapid glucose-sparing benefit to the female as would be physiologically-warranted during an acute, stressful event. This mechanism of glucose diversion away from the mammary glands to realize the transient suppression of milk yield and lactose output, alongside increased/stable fat and protein content, was similarly evident during insulin-induced hypoglycemia (Rook et al., 1965). Indeed, diverting glucose away from the mammary glands results in the rapid cessation of milk synthesis, as occurred in perfused udders (Hardwick et al., 1963). These examples substantiate how the glucose-sparing effect of acute DEX, as documented by others (Kronfeld and Hartmann, 1973; Shamay et al., 2000b), underlies its therapeutic benefit when administered to ketotic dairy cows (Gordon et al., 2013).

The complete reversal of milk production, composition and the mammary transcriptome after DEX highlights the plasticity of the mammary epithelium and lactose synthesis in response to elevated GC. These changes mirror those recorded during several examples of the transient reversal of milk production loss following exposure to a range of stressors including elevated temperature (Collier et al., 2017; Becker et al., 2020), conversion to once-daily milking (Littlejohn et al., 2010), and after the systemic response to mastitis (Shuster et al., 1991). On this last front, our data contribute toward an understanding of the negative systemic regulation of milk synthesis during mastitis. As reviewed by Shangraw and Mcfadden. (2022), an informative model for addressing this question has been the local challenge of one mammary quarter with lipopolysaccharide (LPS) to induce transient hypogalactia in adjacent glands, alongside altered milk composition, reduced lactose output, hyperglycemia and increased circulating GC. In reviewing the mechanisms involved as well as their own data, Shangraw and Mcfadden. (2022) suggested that either inflammatory cytokines derived from the LPS-treated gland, or circulating GC that are elevated in response to either intramammary or intravenous LPS exposure, are the likely mediators underlying this hypogalactia. A comparison of the differentially-expressed gene sets among our data with those of Shangraw et al. (2021) revealed several notable similarities. Of the 14 genes that were differentially expressed (upregulated) in the untreated glands at 3 h after adjacent intramammary LPS (Shangraw et al., 2021), 5 were also upregulated at 12 h post-DEX (*ARRDC2, RGS1, CDKN1A, NFKBIA, and PTX3*; adj $p$ <0 .05), where all these genes have been described as sharing anti-inflammatory properties. Shangraw et al. (2021) also identified that *AQP1* expression was downregulated during LPS-induced hypogalactia, where we recorded that *AQP1* was downregulated at 12 h (albeit adj $p$ = 0.11), and moreso at 24 h (adj $p$ = 0.014), in keeping with a likelihood that its expression changed after the reduction in lactose synthesis. Along similar lines, Littlejohn et al. (2010) found that transcriptomic changes within the udder following its transition to once-daily milking, alongside a reduction in milk yield, LALBA and lactose synthesis, mirrored several of those we recorded following DEX. Of note, in both our study and that of Littlejohn et al. (2010) there was increased expression of *RELA* and downregulation of the toll-like receptors (*TLR2* during once-daily milking, and *TLR3* and *TLR4* after DEX). The fact that these types of immune-associated changes occurred in the mammary glands across all 3 studies [(Shangraw et al., 2021) and our present data], absent any pathogenic response, points to GC-induced activation of local mediators as being a likely mechanism at play during various stress responses. In turn, the systemic increase in GC during states such as mastitis would serve to acutely suppress glucose uptake by all quarters, thereby prioritizing the availability of glucose for the immune system over that for milk production. Certainly the glucose requirements of the immune system are significant, where that of a lactating cow consumes >1 kg during an acute LPS challenge (Kvidera et al., 2017). While there is wide acceptance that the glucose demands in support of a normal lactation are directed by homeorhetic adaptation (Bauman and Currie, 1980), these transcriptomic profiles support the notion that the glucose requirements of an activated immune system trump those of the mammary glands in order to maintain homeostasis (Bradford and Swartz, 2020). We posit that systemic GC levels serve as the central mediator of this balance.

## Conclusion

Our data show that DEX administered to lactating dairy cows leads to the temporary and specific suppression of milk yield and lactose synthesis due to reduced expression of *LALBA* and other lactose synthesis intermediates within the mammary glands. We conclude this response allows the homeorrhetic repartitioning of glucose toward immune activation during physiological stress responses. This work is an important step towards understanding how stress and exogenous GC contribute to transient hypogalactia.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE217369.

## Ethics statement

The animal study was reviewed and approved by University of California Davis Institutional Animal Care and Use Committee.

## Author contributions

AS, SG, and RH contributed to conception and design of the study. AS, SG, JT, AM, and RH conducted the experimentation and sample analysis. AS drafted and RH edited the manuscript. AS, MB, BD-J, and JT performed the statistical analysis. All authors contributed to manuscript revision, read and approved the submitted version.

## Funding

This work was funded in part by a UC Davis-University of Sydney Priority Partnerships Collaboration Award to RCH and SG. AS was supported by the National Institutes of Health (F30 HD101295, UL1 TR001860, and linked award TL1 TR001861), a University of California, Davis Henry A. Jastro Graduate Research Award, and the Daniel T. O'Connor, M.D. Memorial Research Grant.

## Acknowledgments

The authors wish to express their appreciation for animal care and support provided by UC Davis dairy facility managers Doug Gisi and Maria Patino, and research assistance provided by Carmen Banks, Sarah Gabriel, Nicole Hage, Thorben Krueger, Sarah Legge, Nicole Marter, Michael Mui, and Isabella Pendergast.

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1072853/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, 25–29. doi:10.1038/75556

Babwah, T. J., Nunes, P., and Maharaj, R. G. (2013). An unexpected temporary suppression of lactation after a local corticosteroid injection for tenosynovitis. *Eur. J. Gen. Pract.* 19, 248–250. doi:10.3109/13814788.2013.805198

Bauman, D. E., and Currie, W. B. (1980). Partitioning of nutrients during pregnancy and lactation: A review of mechanisms involving homeostasis and homeorhesis. *J. Dairy Sci.* 63, 1514–1529. doi:10.3168/jds.s0022-0302(80)83111-0

Becker, C. A., Collier, R. J., and Stone, A. E. (2020). Invited review: Physiological and behavioral effects of heat stress in dairy cows. *J. Dairy Sci.* 103, 6751–6770. doi:10.3168/jds.2019-17929

Bradford, B. J., and Swartz, T. H. (2020). Review: Following the smoke signals: Inflammatory signaling in metabolic homeostasis and homeorhesis in dairy cattle. *Animal* 14, s144–s154. doi:10.1017/S1751731119003203

Bruckmaier, R. M., and Wellnitz, O. (2008). Induction of milk ejection and milk removal in different production systems. *J. Anim. Sci.* 86, 15–20. doi:10.2527/jas.2007-0335

Casey, T. M., and Plaut, K. (2007). The role of glucocorticoids in secretory activation and milk secretion, a historical perspective. *J. Mammary Gland. Biol. Neoplasia* 12, 293–304. doi:10.1007/s10911-007-9055-3

Collier, R. J., Renquist, B. J., and Xiao, Y. (2017). A 100-year review: Stress physiology including heat stress. *J. Dairy Sci.* 100, 10367–10380. doi:10.3168/jds.2017-13676

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

Emerman, J. T., Bartley, J. C., and Bissell, M. J. (1980). Interrelationship of glycogen metabolism and lactose synthesis in mammary epithelial cells of mice. *Biochem. J.* 192, 695–702. doi:10.1042/bj1920695

Gene Ontology Consortium (2021). The gene ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* 49, D325–D334. doi:10.1093/nar/gkaa1113

Gordon, J. L., Leblanc, S. J., and Duffield, T. F. (2013). Ketosis treatment in lactating dairy cattle. *Vet. Clin. North Am. Food Anim. Pract.* 29, 433–445. doi:10.1016/j.cvfa.2013.03.001

Hara-Chikuma, M., and Verkman, A. S. (2006). Physiological roles of glycerol-transporting aquaporins: The aquaglyceroporins. *Cell. Mol. Life Sci.* 63, 1386–1392. doi:10.1007/s00018-006-6028-4

Hardwick, D. C., Linzell, J. L., and Mepham, T. B. (1963). The metabolism of acetate and glucose by the isolated perfused udder. 2. The contribution of acetate and glucose to carbon dioxide and milk constituents. *Biochem. J.* 88, 213–220. doi:10.1042/bj0880213

Hartmann, P. E., and Kronfeld, D. S. (1973). Mammary blood flow and glucose uptake in lactating cows given dexamethasone. *J. Dairy Sci.* 56, 896–902. doi:10.3168/jds.S0022-0302(73)85274-9

Hong, H., Lee, E., Lee, I. H., and Lee, S.-R. (2019). Effects of transport stress on physiological responses and milk production in lactating dairy cows. *Asian-Australas. J. Anim. Sci.* 32, 442–451. doi:10.5713/ajas.18.0108

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Johnson, H. D., and Vanjonack, W. J. (1976). Effects of environmental and other stressors on blood hormone patterns in lactating animals. *J. Dairy Sci.* 59, 1603–1617. doi:10.3168/jds.S0022-0302(76)84413-X

Kerachian, M. A., Séguin, C., and Harvey, E. J. (2009). Glucocorticoids in osteonecrosis of the femoral head: A new understanding of the mechanisms of action. *J. Steroid Biochem. Mol. Biol.* 114, 121–128. doi:10.1016/j.jsbmb.2009.02.007

Kronfeld, D. S., and Hartmann, P. E. (1973). Glucose redistribution in lactating cows given dexamethasone. *J. Dairy Sci.* 56, 903–908. doi:10.3168/jds.S0022-0302(73)85275-0

Kvidera, S. K., Horst, E. A., Abuajamieh, M., Mayorga, E. J., Fernandez, M. V., and Baumgard, L. H. (2017). Glucose requirements of an activated immune system in lactating Holstein cows. *J. Dairy Sci.* 100, 2360–2374. doi:10.3168/jds.2016-12001

Lemay, D. G., Ballard, O. A., Hughes, M. A., Morrow, A. L., Horseman, N. D., and Nommsen-Rivers, L. A. (2013). RNA sequencing of the human milk fat layer transcriptome reveals distinct gene expression profiles at three stages of lactation. *PLoS ONE* 8, e67531. doi:10.1371/journal.pone.0067531

Littlejohn, M. D., Walker, C. G., Ward, H. E., Lehnert, K. B., Snell, R. G., Verkerk, G. A., et al. (2010). Effects of reduced frequency of milk removal on gene expression in the bovine mammary gland. *Physiol. Genomics* 41, 21–32. doi:10.1152/physiolgenomics.00108.2009

Menta, P. R., Machado, V. S., Pineiro, J. M., Thatcher, W. W., Santos, J. E. P., and Vieira-Neto, A. (2022). Heat stress during the transition period is associated with impaired production, reproduction, and survival in dairy cows. *J. Dairy Sci.* 105, 4474–4489. doi:10.3168/jds.2021-21185

Mobasheri, A., and Barrett-Jolley, R. (2014). Aquaporin water channels in the mammary gland: From physiology to pathophysiology and neoplasia. *J. Mammary Gland. Biol. Neoplasia* 19, 91–102. doi:10.1007/s10911-013-9312-6

Nagamatsu, Y., and Oka, T. (1983). The differential actions of cortisol on the synthesis and turnover of alpha-lactalbumin and casein and on accumulation of their mRNA in mouse mammary gland in organ culture. *Biochem. J.* 212, 507–515. doi:10.1042/bj2120507

Ono, M., and Oka, T. (1980). The differential actions of cortisol on the accumulation of alpha-lactalbumin and casein in midpregnant mouse mammary gland in culture. *Cell* 19, 473–480. doi:10.1016/0092-8674(80)90522-x

Ozmen, A., Unek, G., and Korgun, E. T. (2017). Effect of glucocorticoids on mechanisms of placental angiogenesis. *Placenta* 52, 41–48. doi:10.1016/j.placenta.2017.02.015

Ratman, D., Vanden Berghe, W., Dejager, L., Libert, C., Tavernier, J., Beck, I. M., et al. (2013). How glucocorticoid receptors modulate the activity of other transcription factors: A scope beyond tethering. *Mol. Cell. Endocrinol.* 380, 41–54. doi:10.1016/j.mce.2012.12.014

Rhen, T., and Cidlowski, J. A. (2005). Antiinflammatory action of glucocorticoids-new mechanisms for old drugs. *N. Engl. J. Med.* 353, 1711–1723. doi:10.1056/NEJMra050541

Romero, G., Restrepo, I., Muelas, R., Bueso-Ródenas, J., Roca, A., and Díaz, J. R. (2015). Within-day variation and effect of acute stress on plasma and milk cortisol in lactating goats. *J. Dairy Sci.* 98, 832–839. doi:10.3168/jds.2014-8052

Rook, J. A., Storry, J. E., and Wheelock, J. V. (1965). Plasma glucose and acetate and milk secretion in the ruminant. *J. Dairy Sci.* 48, 745–747. doi:10.3168/jds.s0022-0302(65)88336-9

Sadovnikova, A., Garcia, S. C., and Hovey, R. C. (2021a). A comparative review of the cell biology, biochemistry, and genetics of lactose synthesis. *J. Mammary Gland. Biol. Neoplasia* 26, 181–196. doi:10.1007/s10911-021-09490-7

Sadovnikova, A., Garcia, S. C., and Hovey, R. C. (2021b). A comparative review of the extrinsic and intrinsic factors regulating lactose synthesis. *J. Mammary Gland. Biol. Neoplasia* 26, 197–215. doi:10.1007/s10911-021-09491-6

Sapolsky, R. M., Romero, L. M., and Munck, A. U. (2000). How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. *Endocr. Rev.* 21, 55–89. doi:10.1210/edrv.21.1.0389

Shamay, A., Mabjeesh, S. J., Shapiro, F., and Silanikove, N. (2000a). Adrenocorticotrophic hormone and dexamethasone failed to affect milk yield in dairy goats: Comparative aspects. *Small Rumin. Res.* 38, 255–259. doi:10.1016/s0921-4488(00)00152-8

Shamay, A., Shapiro, F., Barash, H., Bruckental, I., and Silanikove, N. (2000b). Effect of dexamethasone on milk yield and composition in dairy cows. *Ann. Zootech.* 49, 343–352. doi:10.1051/animres:2000125

Shangraw, E. M., and Mcfadden, T. B. (2022). Graduate student literature review: Systemic mediators of inflammation during mastitis and the search for mechanisms underlying impaired lactation. *J. Dairy Sci.* 105, 2718–2727. doi:10.3168/jds.2021-20776

Shangraw, E. M., Rodrigues, R. O., Choudhary, R. K., Zhao, F. Q., and Mcfadden, T. B. (2021). Hypogalactia in mammary quarters adjacent to lipopolysaccharide-infused quarters is associated with transcriptional changes in immune genes. *J. Dairy Sci.* 104, 9276–9286. doi:10.3168/jds.2020-20048

Shuster, D. E., Harmon, R. J., Jackson, J. A., and Hemken, R. W. (1991). Endotoxin mastitis in cows milked four times daily. *J. Dairy Sci.* 74, 1527–1538. doi:10.3168/jds.S0022-0302(91)78313-6

Sjaunja, L., Baevre, L., Junkkarinen, L., Pedersen, J., and Setala, J. (1991). "A Nordic proposal for an energy corrected milk (ECM) formula," in *Performance recording of animals: State of the art, 1990 : Proceedings of the 27th biennial session of the international committee for animal record*. Editors P. Gaillon and Y. Chabert (Wageningen, Netherlands: Centre for Agricultural Publishing and Documentation Pudoc), 156–157.

Stelwagen, K., Mcfadden, H. A., and Demmer, J. (1999). Prolactin, alone or in combination with glucocorticoids, enhances tight junction formation and expression of the tight junction protein occludin in mammary cells. *Mol. Cell. Endocrinol.* 156, 55–61. doi:10.1016/s0303-7207(99)00145-8

Stelwagen, K., Van Espen, D. C., Verkerk, G. A., Mcfadden, H. A., and Farr, V. C. (1998). Elevated plasma cortisol reduces permeability of mammary tight junctions in the lactating bovine mammary epithelium. *J. Endocrinol.* 159, 173–178. doi:10.1677/joe.0.1590173

Waller, K. P. (2000). Mammary gland immunology around parturition. Influence of stress, nutrition and genetics. *Adv. Exp. Med. Biol.* 480, 231–245. doi:10.1007/0-306-46832-8_29

Xiao, C. T., and Cant, J. P. (2005). Relationship between glucose transport and metabolism in isolated bovine mammary epithelial cells. *J. Dairy Sci.* 88, 2794–2805. doi:10.3168/jds.S0022-0302(05)72959-3

Zhao, F. Q. (2014). Biology of glucose transport in the mammary gland. *J. Mammary Gland. Biol. Neoplasia* 19, 3–17. doi:10.1007/s10911-013-9310-8

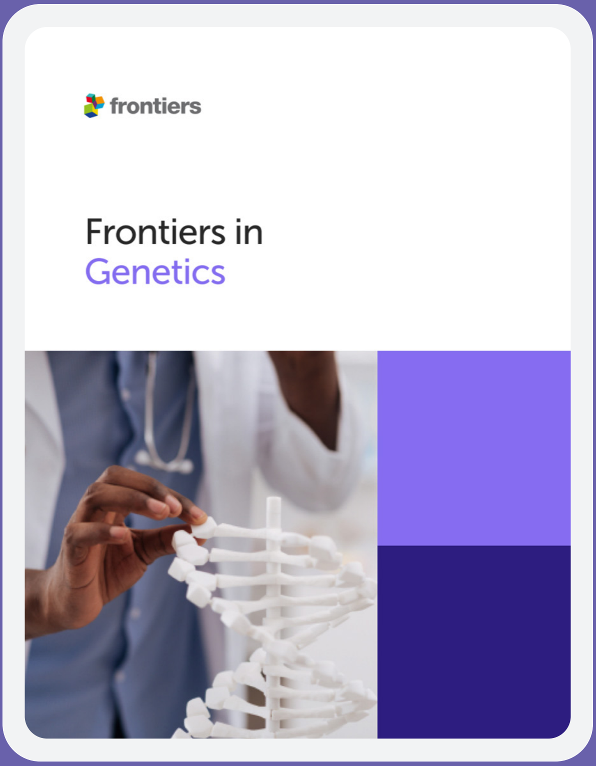
# Frontiers in Genetics

**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

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact