frontiers | Research Topics

Application of genomics in livestock populations under selection or conservation

Edited by
Anupama Mukherjee and Zexi Cai

Coordinated by
Sabyasachi Mukherjee

Published in
Frontiers in Genetics
Frontiers in Veterinary Science

ISSN 1664-8714
ISBN 978-2-8325-4420-4
DOI 10.3389/978-2-8325-4420-4

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

Application of genomics in livestock populations under selection or conservation

Topic editors
Anupama Mukherjee — Indian Council of Agricultural Research (ICAR), India
Zexi Cai — Aarhus University, Denmark

Topic Coordinator
Sabyasachi Mukherjee — Indian Council of Agricultural Research (ICAR), India

Citation
Mukherjee, A., Cai, Z., Mukherjee, S., eds. (2024). *Application of genomics in livestock populations under selection or conservation*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4420-4

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 24 January 2024
DOI 10.3389/fgene.2024.1363839

OPEN ACCESS

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

*CORRESPONDENCE
Sabyasachi Mukherjee,
sabayasachimukherje@gmail.com

†These authors have contributed equally to this work

RECEIVED 31 December 2023
ACCEPTED 17 January 2024
PUBLISHED 24 January 2024

CITATION
Mukherjee A, Cai Z and Mukherjee S (2024), Editorial: Application of genomics in livestock populations under selection or conservation. *Front. Genet.* 15:1363839. doi: 10.3389/fgene.2024.1363839

Editorial: Application of genomics in livestock populations under selection or conservation

Anupama Mukherjee[1†], Zexi Cai[2†] and Sabyasachi Mukherjee[1*†]

[1]ICAR-National Dairy Research Institute, Karnal, Haryana, India, [2]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark

KEYWORDS

livestock genomics, selection, conservation, WGS, GWAS, genotyping

Editorial on the Research Topic
Application of genomics in livestock populations under selection or conservation

Genomics is one of the newest branches of biology that has progressed tremendously during the last decades. Genomics deals with the molecular structures, functions, evolution, and mapping of the genomes of any species and has significantly generated new information that has improved our understanding of the complex biology and genetic mechanisms of animal production systems. The advancement of genomics is linked with a number of key developments, which include the rapid expansion of next-generation sequencing and chip-based genotyping assays. Large-scale genomics data are now utilized more and more due to the dwindling cost of such sequencing and genotyping techniques. Livestock breeding programs, including selection and conservation efforts, have attained huge success due to affordable genomic prediction, particularly in dairy cattle. It is expected that there will be further reduction in the cost of these high-throughput genomic data generation platforms and more development of precise estimation methodologies. Multi-disciplinary involvement is going to further benefit the genomics community with the advancement of robust and reliable tools in the field of bioinformatics and their use in livestock breeding.

Keeping these developments in the area of livestock genomics in mind, the present Research topic of the Frontiers in Genetics titled "Application of Genomics in Livestock Populations under Selection or Conservation" was aptly selected with several major themes that highlighted the usage of genomics for conservation, current methods of genomics, application of whole-genome- and genome-wide-based techniques, and use of different bioinformatics tools and pipelines for the processing of genomic data. The resulting efforts contributed to the publication of a total 19 research papers in the current volume, comprising major focal points in the area of genomics of livestock and other species with the concerns of the present day. However, the ocean of genomics is too vast, and even this wide-array of published articles could hardly justify an ounce of that vastness!

Nonetheless, genuine efforts were made to include articles in this volume on those central themes of genomics that comprise the major skills and techniques employed in various animal populations for selection and conservation issues. These include genome-wide association studies (GWAS), differential gene expression utilizing transcriptome data, and analysis of selection signatures through whole-genome sequencing and high-density genotyping datasets, which are utilized for discovering genes and genomic variants that control significant traits of importance in livestock species.

Terefe et al. in African cattle, Naji et al. in Austrian and Chinese cattle, Karam et al. in sobaity seabream, Heidaritabar et al. in crossbred and purebred pigs, Li et al. in yak populations, Chen et al. in buffaloes, and Liu et al. in Chinese Qaidam cattle utilized whole-genome-sequenced datasets for the analysis of variant calling, selection signatures, and genomic copy number variations to identify genes related to major economic traits of importance, including production, growth, immunity, and adaptability, in these livestock species. Terefe et al. provided possible examples of convergent selection between cattle and humans through the identification of unique selection signatures in African cattle living in the Ethiopian highlands.

Liu et al. generated skin transcriptome data from cashmere goats and identified key pathways and six hub genes (*PDGFRA, WNT5A, PPP2R1A, BMPR2, BMPR1A*, and *SMAD1*) that regulate cashmere development in these goats that are mediated by exogenous melatonin. This study is expected to provide a foundation for understanding the mechanism of melatonin-regulated cashmere growth. Liu et al. identified LTBP2 gene polymorphisms and their association with the thoracolumbar vertebrae number, body size, and carcass traits in Dezhou donkeys, which will be useful as a molecular marker to improve the production performance in this donkey population.

Genomic selection (GS) is another potential breeding tool that can reduce the generation interval, improve the accuracy of selection, and bring genetic improvement and, therefore, has been successfully employed in many farm animals for more than a decade now (Hayes et al., 2009; Gorjanc et al., 2015; de Koning, 2016; Meuwissen et al., 2016; Wiggans et al., 2017; Yang et al., 2020). Ning et al. studied various marker densities and designed several statistical models to increase the accuracy of genomic selection for wool traits in Angora rabbits. They are the first to estimate genomic heritability in Angora rabbits and showed that their work will be able to provide key strategies to optimize GS using 50k marker density in rabbits for early selection of various wool traits. Wiggans and Carrillo reviewed the progress of GS in the United States and found that the dairy genomic selection program has doubled the rate of annual genetic gain since 2010, with a rapid increase in the number of genotype evaluations for over 50 traits. The use of genomic information has enabled us to determine the value of animals at a much earlier age and has contributed to a dramatic increase in the rate of genetic improvement.

Culver and Labow (2002) mentioned that genomics is a multi-disciplinary field of biology that focuses on the structure, function, evolution, mapping, and editing of genomes. Unlike genetics, which refers to the study of individual genes having a role in inheritance and variation, genomics targets the combined characterization of all of an organism's genes, their combined inheritance, and variation on the organism (WHO, 2020). Genomics also includes studies of various within-the-genome phenomena such as epistasis (the effect of one gene on another), pleiotropy (one gene affecting more than one trait), heterosis or hybrid vigor, and the different interplay of loci and alleles within the genome (Pevsner, 2009). Since the domain of genomics is quite broad, it is unlikely that the coverage of the present Research Topic will be able to encompass all. Nonetheless, the 19 articles did cover a great number of aspects of genomics and their application in various animal populations for selection and conservation. Given that these articles are interesting and timely, we agree that there is still scope for improvement in incorporating genomics Research Topic for the selection of complex traits, i.e., the selection for traits with low heritability and disease resistance in farm animals and other species. High-throughput phenotyping or phenomics and precision farming are such tools that are being applied along with genomics in most advanced countries for livestock breeding programs (Pedrosa et al., 2023). We sincerely hope that a future Research Topic on genomics may well embrace many other emerging areas within genomics, i.e., phenomics, epigenomics, and metagenomics, utilized not only for genetic improvement programs but also for the sustainability of livestock production systems.

Author contributions

AM: Writing–original draft, Writing–review and editing, Conceptualization. ZC: Writing–original draft, Writing–review and editing, Conceptualization. SM: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors acknowledge the Directors, ICAR-NDRI, Karnal, India, and the Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark, for all the facilities to carry out this work.

Conflict of interest

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

Publisher's note

References

Culver, K. W., and Labow, M. A. (2002). "Genomics," in *Genetics*. Editor R. Robinson (New York, NY: Macmillan Science Library, Macmillan Reference USA).

de Koning, D. J. (2016). Meuwissen et al. on Genomic Selection. *Genetics* 203 (1), 5–7. doi:10.1534/genetics.116.189795

Gorjanc, G., Cleveland, M. A., Houston, M. D., and Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Sel. Evol.* 47, 12. doi:10.1186/s12711-015-0102-z

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92 (2), 433–443. doi:10.3168/jds.2008-1646

Meuwissen, T. H., Hayes, B., and Goddard, T. (2016). Genomic selection: a paradigm shift in animal breeding. *Anim. Front.* 6, 6–14. doi:10.2527/af.2016-0002

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

Pevsner, J. (2009). *Bioinformatics and functional genomics*. 2nd Edn. Hoboken, NJ: Wiley-Blackwell.

WHO (2020). *WHO definitions of genetics and genomics*. World Health Organization, 20. Archived from the original on December.

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

Yang, C. J., Sharma, R., Gorjanc, G., Hearne, S., Powell, W., and Mackay, I. (2020). Origin specific genomic selection: a simple process to optimize the favorable contribution of parents to progeny. *G3 Genes| Genomes| Genet.* 10 (7), 2445–2455. doi:10.1534/g3.120.401132

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fgene.2022.968712

OPEN ACCESS

EDITED BY
Guosheng Su,
Aarhus University, Denmark

REVIEWED BY
Tianfei Liu,
Guangdong Academy of Agricultural Sciences, China
Xinsheng Wu,
Yangzhou University, China

*CORRESPONDENCE
Dan Wang,
wangd_18@163.com
Xinzhong Fan,
xzfan@sdau.edu.cn

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

RECEIVED 14 June 2022
ACCEPTED 17 August 2022
PUBLISHED 02 September 2022

CITATION
Ning C, Xie K, Huang J, Di Y, Wang Y, Yang A, Hu J, Zhang Q, Wang D and Fan X (2022), Marker density and statistical model designs to increase accuracy of genomic selection for wool traits in Angora rabbits.
Front. Genet. 13:968712.
doi: 10.3389/fgene.2022.968712

Marker density and statistical model designs to increase accuracy of genomic selection for wool traits in Angora rabbits

Chao Ning, Kerui Xie, Juanjuan Huang, Yan Di, Yanyan Wang, Aiguo Yang, Jiaqing Hu, Qin Zhang, Dan Wang* and Xinzhong Fan*

College of Animal Science and Veterinary Medicine, Shandong Agricultural University, Tai'an, China

The Angora rabbit, a well-known breed for fiber production, has been undergoing traditional breeding programs relying mainly on phenotypes. Genomic selection (GS) uses genomic information and promises to accelerate genetic gain. Practically, to implement GS in Angora rabbit breeding, it is necessary to evaluate different marker densities and GS models to develop suitable strategies for an optimized breeding pipeline. Considering a lack in microarray, low-coverage sequencing combined with genotype imputation was used to boost the number of SNPs across the rabbit genome. Here, in a population of 629 Angora rabbits, a total of 18,577,154 high-quality SNPs were imputed (imputation accuracy above 98%) based on low-coverage sequencing of 3.84X genomic coverage, and wool traits and body weight were measured at 70, 140 and 210 days of age. From the original markers, 0.5K, 1K, 3K, 5K, 10K, 50K, 100K, 500K, 1M and 2M were randomly selected and evaluated, resulting in 50K markers as the baseline for the heritability estimation and genomic prediction. Comparing to the GS performance of single-trait models, the prediction accuracy of nearly all traits could be improved by multi-trait models, which might because multiple-trait models used information from genetically correlated traits. Furthermore, we observed high significant negative correlation between the increased prediction accuracy from single-trait to multiple-trait models and estimated heritability. The results indicated that low-heritability traits could borrow more information from correlated traits and hence achieve higher prediction accuracy. The research first reported heritability estimation in rabbits by using genome-wide markers, and provided 50K as an optimal marker density for further microarray design, genetic evaluation and genomic selection in Angora rabbits. We expect that the work could provide strategies for GS in early selection, and optimize breeding programs in rabbits.

KEYWORDS

angora rabbit, wool, genomic selection, marker density, model

1 Introduction

The Angora rabbit is a well-known breed for fiber production that provides wool usually chosen for the production of luxury textile materials. Genetic improvement of wool production and quality is essential for achieving sustained increase in fiber production. Genomic selection (GS) is a potential breeding tool, and has been successfully employed in many farm animals, such as pigs and dairy cattle (Meuwissen et al., 2013; Gorjanc et al., 2015; Wiggans et al., 2017; Yang et al., 2020). GS can reduce the interval of generation, improve the accuracy and intensity of selection, and contribute to genetic improvement (He et al., 2019). A number of simulation and empirical studies on GS has realized impacts on improvement in the animal production (Solberg et al., 2008; Wiggans et al., 2017; Karimi et al., 2019; Yang et al., 2020), and GS has been effectively used in animal breeding programs for more than a decade (Hayes et al., 2009; Jannink et al., 2010). The exploitation of genome-assisted approaches could greatly benefit breeding efforts in Angora rabbits, though rabbits breeding is slower to adopt this technology. In rabbits, a high-density commercial SNP microarray (Affymetrix Axiom OrcunSNP Array, around 200k SNPs) was not available until 2015, and a lack in inexpensive chips and high genotyping cost by genome sequencing in rabbits delay genomic selection application; Additional issues such as the small economic value of paternal rabbits and the short generation interval are still limiting genomic selection as an evaluating method (Mancin et al., 2021).

Various factors appear to affect prediction accuracy in genomic selection (Covarrubias-Pazaran et al., 2018; Krishnappa et al., 2021). Marker density is a force driving the prediction accuracies of GS, and has been so far one of the most studied factors. It is suggested that high density markers can improve the prediction accuracy (Hickey et al., 2014; Al-Khudhair et al., 2021), and the consensus is that a higher number of markers usually yield higher accuracy reaching a plateau (Wang et al., 2017; Krishnappa et al., 2021). In the presence of genome resequencing, genome-wide SNPs are available for rabbits, but what density of markers is optimal for GS in Angora rabbits, *i.e.*, the density reaching a plateau, remains obscure, since the efficient SNP number could reduce the dimensionality of the GS model.

Various studies related to GS have been mostly confined to single trait in the recent past, although they performed not very well in cases of pleiotropy, missing data and low heritability (Boison et al., 2017; Budhlakoti et al., 2019). Gradually, studies were carried out to explore the possibility of methods for GS based on multiple traits that enabled to provide accurate genomic prediction by exploiting the information of correlated structure among response (Budhlakoti et al., 2019). In addition, breeders in animal breeding usually record one trait at multiple times throughout the lifetime of animals that are often strongly genetically correlated. The optimal estimation procedure is to combine information from multiple records of different traits or different times to obtain genomic estimated breeding values (GEBV) using the multi-trait models (Okeke et al., 2017; Covarrubias-Pazaran et al., 2018). In the breeding of Angora rabbits, we have very little idea about the performance of GS, so single-trait and multi-trait models should be explored.

In this study, we used the genomic resources of Angora rabbits in hand to test the usefulness of genomic selection. In order to maximize genomic prediction accuracy, we focused on estimating the optimal marker density undergoing a renaissance thanks to genome resequencing, and comparing the GS performance between single-trait and multi-trait models for genomic best linear unbiased prediction (GBLUP). The research would provide strategies for GS in early selection of wool production, and optimize breeding programs in Angora rabbits.

2 Materials and methods

2.1 Animal phenotypes and genotypes

A total of 629 Agora rabbits (298 males and 331 females) used for this study were from same batch. All rabbits were housed under the same conditions on a farm, including diet, water and temperature. In production practice, the rabbits are artificially inseminated with mixed semen, so there is not a definite pedigree information for the studied population. The wool is harvested around every 70 days from 70 days of age, and after the third shearing, the rabbits are selected for breeding. The associated wool traits including length of fine wool (LFW), diameter of fine wool (DFW), coefficient of variation of diameter of fine wool (CVDFW), length of coarse wool (LCW), rate of coarse wool (RCW) and weight of sheared wool (WSW) were measured at 70, 140 and 210 days of age. In addition, body weight (BW) was measured at the same days of age. The descriptive summary was provided for the traits in Supplementary Table S1.

Ear samples were collected from the individuals. Genomic DNA was isolated using the Qiagen MinElute Kit. Genomic DNA from each sample was used to construct a paired-end library (PE150) with an insert size of ~350 bp. All libraries were sequenced on the DNBSEQ-T7 platform. By low-coverage whole genome sequencing (LCS), an average of 3.84X genomic coverage was sequenced, with the read depth varying from 1.51X to 8.03X. Read quality was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters and low-quality bases were removed using Trimmomatic (Bolger et al., 2014). Sample reads were mapped to the rabbit reference sequence GCF_000003625.3 (*Oryctolagus cuniculus*) using BWA-mem (Li and Durbin, 2009). SNPs were called using BaseVar (Liu et al., 2018) and imputed genotype dosages at missing sites using STITCH (Davies et al., 2016). The SNPs were filtered for an imputation info

score >0.4 using Bcftools (Li, 2011), and then with 'MAF >0.05, genotype missing rate <0.1 and a Hardy-Weinberg equilibrium (HWE) *p*-value > 1E-6' using PLINK (Chang et al., 2015). The sites which were missing in 10% of the individuals after STITCH imputation were then imputed by Beagle v5.1 Beagle v5.1 (Browning et al., 2018). A total of 18,577,154 high-quality SNPs (imputation accuracy above 98%) were used after stringent quality control. The manipulation of phenotypes and genotypes is detailed in the previous study (Wang et al., 2022).

2.2 Models

In our studies, univariate linear mixed models (uvLMM) were used to analyze the traits measured at three time points, respectively. The univariate linear mixed models are formulated as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{1}$$

Here, **y** is the phenotypic vectors of a specific time point; **b** is the vector of fixed effects including population mean, batch and sex; **u** is the vector of random genetic effects; **e** is the vector of random residuals. **X** and **Z** are the corresponding design matrixes. The distributions of random effects are

$$\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_u^2\right),\ \mathbf{e} \sim N\left(\mathbf{0},\ \mathbf{I}\sigma_e^2\right) \tag{2}$$

Where, σ_u^2 is the genetic variance; σ_e^2 is the residual variance; **G** is the genomic relationship matrix built with method of VanRaden (Vanraden, 2008); **I** an identity matrix.

To test the performance of GS using multivariate linear mixed models (mvLMM), we regarded the records from three time points of one trait as different traits and used mvLMM to analyze the data. The multivariate linear mixed models are formulated as

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & & \\ & \mathbf{X}_2 & \\ & & \mathbf{X}_3\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \mathbf{b}_3\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1 & & \\ & \mathbf{Z}_2 & \\ & & \mathbf{Z}_3\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \mathbf{u}_3\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{e}_3\end{bmatrix} \tag{3}$$

All symbols have the same meaning with the single-trait models, and subscripts ($i = 1, 2, 3$) indicate the ith time point. The distributions of random effects are

$$\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \mathbf{u}_3\end{bmatrix} \sim N\left(\mathbf{0},\ \sum_u \otimes \mathbf{G}\right),\ \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{e}_3\end{bmatrix} \sim N\left(\mathbf{0},\ \sum_e \otimes \mathbf{I}\right) \tag{4}$$

Where, $\sum_u$ and $\sum_e$ are a 3 × 3 covariance matrix for the genetic effects and residual errors.

2.3 Marker densities

To evaluate the influence of marker density on the heritability estimation and genomic prediction, we randomly selected 0.5K, 1K, 3K, 5K, 10K, 50K, 100K, 500K, 1M and 2M from the original 18.6M markers. Then, we used these randomly selected markers to build the genomic relationship matrix, and estimate heritabilities and genomic breeding values with univariate linear mixed models. For each marker density, we repeat this process for 30 times to obtain stable results.

2.4 Cross-validation

We used 10-fold cross-validation (CV) to assess the accuracy of the genomic prediction. The 629 individuals were randomly shuffled and split into 10 groups. One of them was used as a validation population in turn, and the remaining nine groups used as a training population. The accuracy of genomic prediction was assessed by the correlation between corrected phenotypic values (y_c) and GEBVs in the validation population ($r_{y_c,GEBV}$). Here, the corrected phenotypic values were calculated with general linear model, which removed sex and batch effects from the original phenotypic values. For the three-trait models analysis, we also compared different leave-out strategies: 1) leave out the observations of all the three time points; 2) leave out the observations of the last two time points; 3) leave out the observations of the last time point. The aim of the three leave-out strategies was to explore whether and how the accuracy of the genomic prediction would be improved with early measured traits. In the study, we used two-sample *t*-test to determine whether the prediction accuracies from two experiments (varied marker densities or models) were significantly different from each other.

2.5 Implements

The genomic relationship matrix was built with GMAT (https://github.com/chaoning/GMAT), and uvLMM and mvLMM were implemented with DMU package (https://dmu.ghpc.au.dk/dmu/).

3 Results

3.1 50K markers are the baseline to estimate heritability for angora rabbits

In order to produce different marker densities, we randomly selected 0.5K, 1K, 3K, 5K, 10K, 50K, 100K, 500K, 1M and 2M from the original sequencing markers and repeat 30 times for each marker density to reduce the sampling error. We calculated the Pearson correlation coefficients between all genomic relationship matrixes built from randomly selected

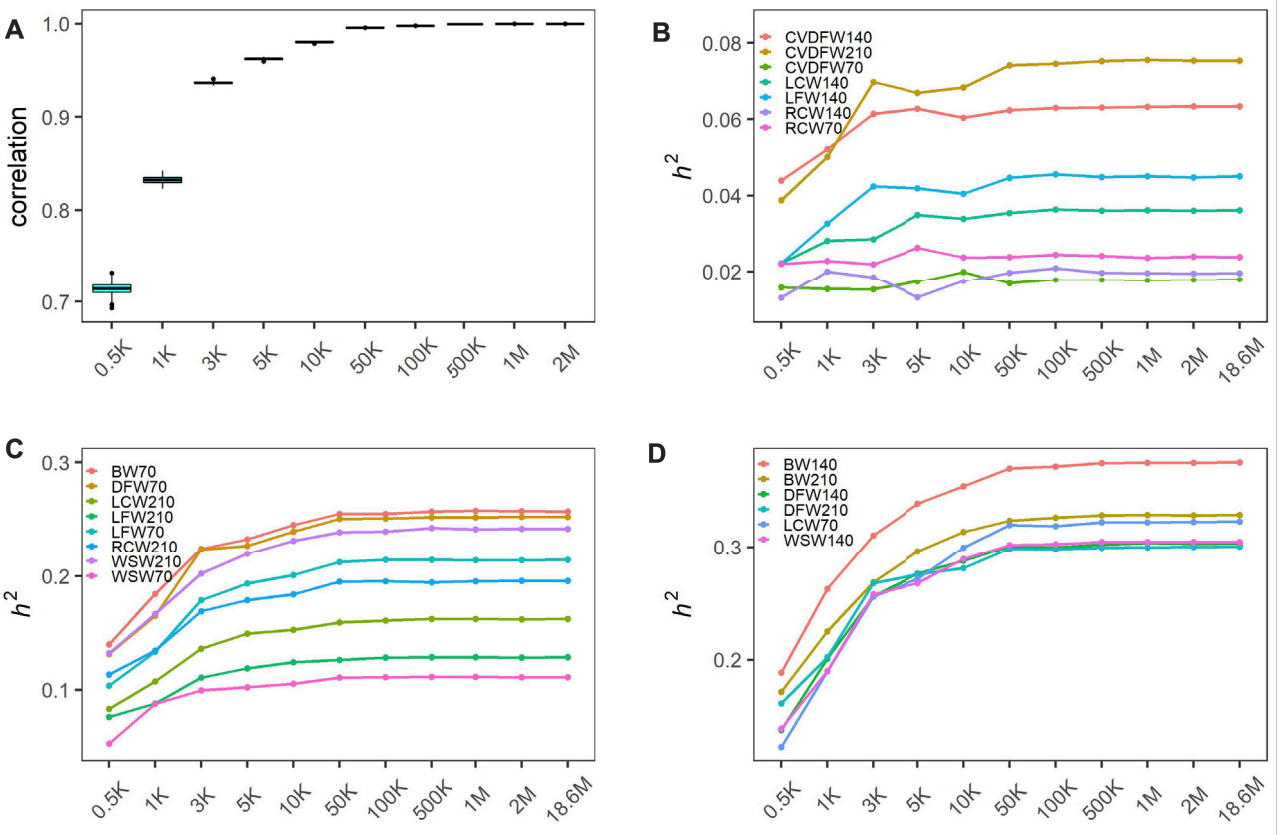


FIGURE 1
Heritability estimation of wool traits and body weight with varied marker densities. **(A)** Pearson correlation coefficients between all genomic relationship matrixes built from randomly selected markers; The changing estimated heritability for low heritability traits **(B)**, medium heritability traits **(C)** and high heritability traits **(D)**. Traits were categorized into three categories according to the heritability estimated with the original sequencing markers: low heritability (<0.1), medium heritability (0.1–0.3) and high heritability (>0.3).

TABLE 1 Heritability estimated by the genome markers of 18.6M in the single-trait model.

Trait	Heritability	Trait	Heritability
BW70	0.257±0.05367	LFW70	0.214±0.0512
BW140	0.375±0.05271	LFW140	0.045±0.0385
BW210	0.328±0.05272	LFW210	0.129±0.04422
CVDFW70	0.018±0.03479	RCW70	0.024±0.03705
CVDFW140	0.063±0.03697	RCW140	0.02±0.03683
CVDFW210	0.075±0.04199	RCW210	0.196±0.04849
DFW70	0.252±0.04953	WSW70	0.111±0.04758
DFW140	0.303±0.05209	WSW140	0.305±0.05328
DFW210	0.301±0.05111	WSW210	0.242±0.0502
LCW70	0.323±0.05669		
LCW140	0.036±0.03552		
LCW210	0.162±0.04733		

markers for each maker density. We found that the Pearson correlation coefficients increased rapidly and the dispersion degree decreased with the increase of marker density from 0.5K to 50K, and the values tended to be steady with the minimum value exceeding 0.99 (Figure 1A). We categorized traits into three categories according to the heritability estimated with the original sequencing markers: low heritability (<0.1), medium heritability (0.1–0.3) and high heritability (>0.3), and showed the average estimated heritability of 30 random selections for each marker density in Figures 1B–D. We observed that estimated heritabilities increased rapidly with the marker density increasing from 0.5K to 50K, and then maintained steady when the marker density continued to increase for all levels of heritability. The heritabilities estimated by the genome markers of 18.6M were listed in Table 1.

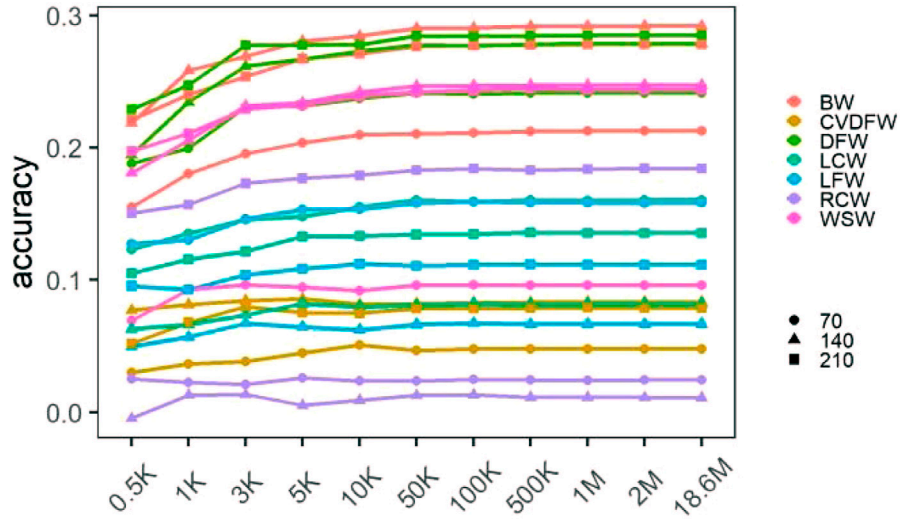


FIGURE 2
Mean prediction accuracies of cross-validation under different marker densities.

3.2 50K markers can achieve ideal prediction accuracy for angora rabbits

We calculated the prediction accuracy for each trait by averaging the cross-validation results of 30 random selections (Supplementary Table S2), and showed the changing prediction accuracy with the increase of marker density in Figure 2. For all traits, the mean accuracies were lower than 0.3 regardless of marker density. Similar to the change tendency of estimated heritability, we found that the prediction accuracy increased rapidly with the increase of marker density from 0.5K to 50K, and it improved very little when the marker density continued to increase. In addition, the significance of the differences between the prediction accuracies under different marker densities was listed in Supplementary Table S3. There was no significant difference between the accuracies under the marker density of 50K and 100K for all the traits, while when comparing 50K–10K, the difference between the accuracies was significant for several traits such as BW140, BW210, DFW210, LCW70 and WSW140.

3.3 Multiple-trait models can improve the prediction accuracy in genomic selection

We applied the multiple-trait models to analyze the records from three time points of one trait. Compared to the single-trait models, the prediction accuracy of nearly all traits could be improved, except that it was slightly decreased for BW140 which was decreased from 0.292 from 0.288 (Figure 3, Supplementary Table S4 and S5). The Pearson correlation coefficient between the increased prediction accuracy from single-trait to multiple-trait models and estimated heritability was -0.584 (p = 0.0055), which indicated that the prediction accuracy of traits with lower heritability can be improved more with multiple-trait models. For example, CVDFW belonging to low heritability traits, its estimated heritabilities at three time points were 0.018, 0.063 and 0.075, respectively, but their prediction accuracy could be improved by 71.35%, 85.71% and 68.81%, respectively.

In the cross-validation experiments, if we kept the observations of early time points in the validation groups, the prediction accuracy could be further improved by multiple-trait models; and the more observations of early time points kept, the higher prediction accuracy could reach for the majority of traits (Figure 3, Supplementary Table S4 and S5).

4 Discussion

Genomic selection promises to accelerate genetic gain in animal breeding programs (Meuwissen et al., 2013; Gorjanc et al., 2015; Yang et al., 2020). Practically, to implement GS in Angora rabbit breeding, it is necessary to evaluate different marker densities and GS models to develop suitable strategies for an optimized breeding pipeline.

Low-coverage sequencing combined with genotype imputation boosts the number of SNPs across the genome. It plays out advantages in obtaining genotyping information since both DNA library and sequencing cost decreased (Nicod et al., 2016; Meier et al., 2021) especially when lacking in microarray (Davies et al., 2016; Davies et al., 2021). In this study, we evaluated different marker densities for heritability estimation

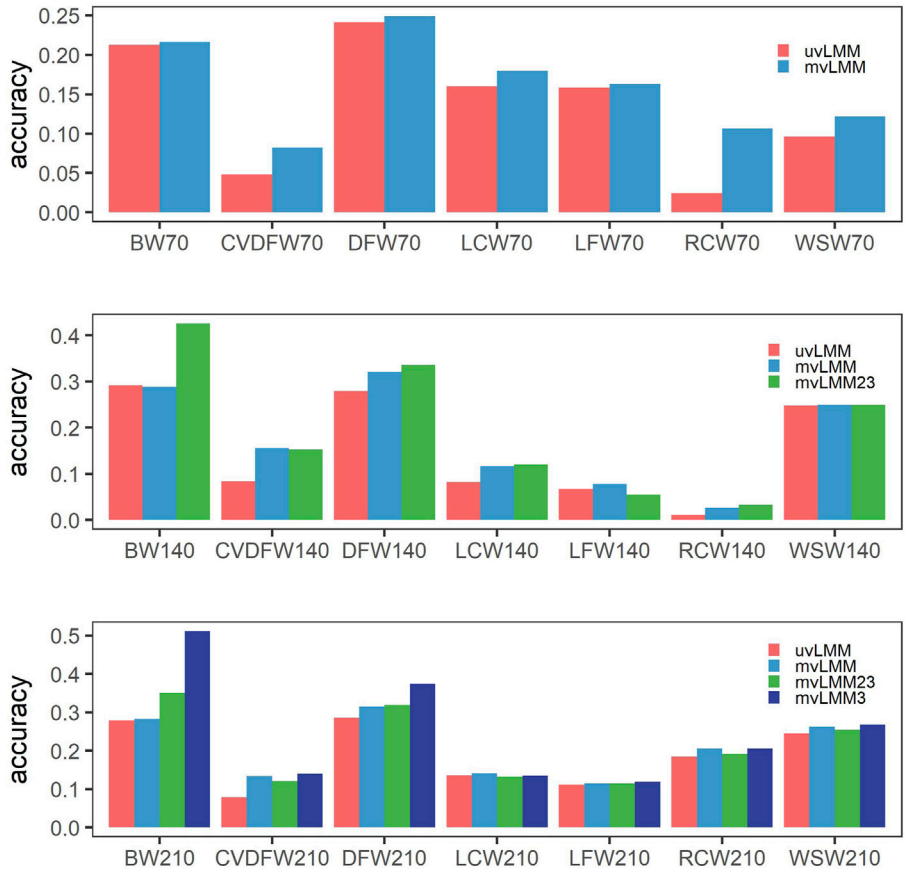


FIGURE 3
Mean prediction accuracies of cross-validation by different models. uvLMM: univariate linear mixed models which analyzed records of different time points, respectively; mvLMM: multivariate linear mixed models which considered the correlations of different time points and leave out the observations of all the three time points in the cross-validation experiments; mvLMM23: leave out the observations of the second and third time points in the cross-validation experiments; mvLMM3: leave out the observations of the third time points in the cross-validation experiments.

and genomic selection, and provided 50K as an optimal marker density for further microarray design, genetic evaluation and genomic selection in rabbits, since the efficient SNP number could reduce the dimensionality of the calculation model.

The heritability of various traits in rabbits was traditionally estimated by using pedigree information (Dige et al., 2012; El Nagar et al., 2020; Montes-Vergara et al., 2021). To our knowledge, this study was the first report for heritability estimation in rabbits by using genome-wide markers. What's more, for Angora rabbits, little information is available on heritabilities of production performance and economic traits. The previous estimation using pedigree information included the heritability of wool production, coarse wool rate and body weight of Wan-strain Angora rabbits at 11-month-old (0.33, 0.21 and 0.43, respectively) (Zhao. et al., 2016), and the heritability of weaning weight, wool yield of first, second and third clips of Angora rabbits (0.24, 0.22, 0.20 and 0.21, respectively) (Niranjan et al., 2011). By exploring the influence of marker density on heritability estimation, we estimated stable heritabilities for wool and body weight traits of Angora rabbits at the marker density of 50K.

It becomes clear that the increase in the marker density by panels and even genome sequencing could not result in ceaselessly increase in the accuracy of genomic selection (Chang et al., 2019). In this study, the marker density showed major effects on the improvement of prediction accuracy below 50K, which showed the accuracy predicted by GS increased as the marker density increased for all traits in the rabbit population. However, above a threshold of 50K, the marker density showed minor effects. 50K is a density of genome markers in common usage for livestock genetics and breeding (Nandolo et al., 2019; He et al., 2020; Bhuiyan et al., 2021; Liu et al., 2021; Singh et al., 2021). Similar phenomenon was found in other species though the baseline of marker density was different (Spindel et al., 2015; Wang et al., 2017). The threshold where the plateau takes place might be associated with the extent of linkage disequilibrium

(LD) between genome markers and QTLs. At a long extent of LD, the number of independent segments in the genome is expected to be small, which means fewer markers are needed to mark all segments (Wientjes et al., 2013; Wang et al., 2017). In the present study, the average pairwise LD r2 values decreased to 0.16 at 500 kb and to 0.11 at 1 Mb (Wang et al., 2022), and the population was considered to have a relatively slow decay of LD similar to other livestock population, hence 50K, a small number of markers, was sufficient to produce the accurate prediction.

A large number of genomic selection studies have focused on single-trait analyses (Boison et al., 2017; Budhlakoti et al., 2019). However, many traits are genetically correlated, such as the Angora wool traits among different shearing times. It has been shown that a multiple-trait genomic model had higher prediction accuracy than a single-trait genomic model, and the use of multiple-trait models is one of the ideas to increase the predictive ability of GS (Guo et al., 2014; Covarrubias-Pazaran et al., 2018). In this study, the majority of traits reached higher accuracy predicted by multiple-trait models than by single-trait models, because multiple-trait models used information from genetically correlated traits (Jia and Jannink, 2012). Furthermore, we observed high significant negative correlation between the increased prediction accuracy from single-trait to multiple-trait models and estimated heritability. The results indicated that low-heritability traits can borrow more information from correlated traits and hence achieve higher prediction accuracy. Especially, the prediction accuracy of BW140 with the highest heritability among the analyzed traits, was slightly decreased. Since many wool traits belong to medium and low heritability, this characteristic of multiple-trait could be very important in Angora rabbits breeding (Jia and Jannink, 2012).

5 Conclusion

Genomic selection was applied to Angora rabbits based on low-coverage sequencing combined with genotype imputation. A total of 18,577,154 high-quality SNPs were obtained with imputation accuracy above 98%. From the original markers, 0.5K, 1K, 3K, 5K, 10K, 50K, 100K, 500K, 1M and 2M were randomly selected and evaluated, resulting in 50K markers as the baseline for the heritability estimation and genomic prediction. Comparing to the GS performance of single-trait models, the prediction accuracy of nearly all traits could be improved by multi-trait models. Furthermore, we observed high significant negative correlation between the increased prediction accuracy from single-trait to multiple-trait models and estimated heritability. The results indicated that low-heritability traits could borrow more information from correlated traits and hence achieve higher prediction accuracy. The research first reported heritability estimation in rabbits by using genome-wide markers, and provided 50K as an optimal marker density for further microarray design, genetic evaluation and genomic selection in Angora rabbits. We expect that the work could provide strategies for early selection, and optimize breeding programs in rabbits.

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/, PRJNA810279.

Ethics statement

The animal study was reviewed and approved by Animal Care and Use Committee of Shandong Agricultural University.

Author contributions

XF, CN, and DW conceived the idea. YW, and AY collected the data. CN, and DW performed the theoretical study. KX, JH, YD, and CN analyzed the data. DW, CN, and XF wrote the manuscript with input from all authors. XF, QZ, and JH supervised the research. All authors contributed to the discussions of the results.

Funding

This work was supported by the Agricultural Improved Seed Project of Shandong Province (2021LZGC002), Shandong Province Special Economic Animal Innovation Team (SDAIT-21–02), National Natural Science Foundation of China (32102526 and 32002172), China Postdoctoral Science Foundation (2020M682217), Shandong Provincial Postdoctoral Program for Innovative Talent and Shandong Provincial Natural Science Foundation (ZR2020QC176 and ZR2020QC175).

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.2022.968712/full#supplementary-material

References

Al-Khudhair, A., Vanraden, P. M., Null, D. J., and Li, B. (2021). Marker selection and genomic prediction of economically important traits using imputed high-density genotypes for 5 breeds of dairy cattle. *J. Dairy Sci.* 104, 4478–4485. doi:10.3168/jds.2020-19260

Bhuiyan, M. S. A., Lee, S. H., Hossain, S. M. J., Deb, G. K., Afroz, M. F., Lee, S. H., et al. (2021). Unraveling the genetic diversity and population structure of Bangladeshi indigenous cattle populations using 50K SNP markers. *Animals.* 11, 2381. doi:10.3390/ani11082381

Boison, S. A., Utsunomiya, A. T. H., Santos, D. J. A., Neves, H. H. R., Carvalheiro, R., Meszaros, G., et al. (2017). Accuracy of genomic predictions in Gyr (*Bos indicus*) dairy cattle. *J. Dairy Sci.* 100, 5479–5490. doi:10.3168/jds.2016-11811

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

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

Budhlakoti, N., Mishra, D. C., Rai, A., Lal, S. B., Chaturvedi, K. K., and Kumar, R. R. (2019). A comparative study of single-trait and multi-trait genomic selection. *J. Comput. Biol.* 26, 1100–1112. doi:10.1089/cmb.2019.0032

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8

Chang, L. Y., Toghiani, S., Aggrey, S. E., and Rekaya, R. (2019). Increasing accuracy of genomic selection in presence of high density marker panels through the prioritization of relevant polymorphisms. *BMC Genet.* 20, 21. doi:10.1186/s12863-019-0720-5

Covarrubias-Pazaran, G., Schlautman, B., Diaz-Garcia, L., Grygleski, E., Polashock, J., Johnson-Cicalese, J., et al. (2018). Multivariate GBLUP improves accuracy of genomic selection for yield and fruit weight in biparental populations of vaccinium macrocarpon ait. *Front. Plant Sci.* 9, 1310. doi:10.3389/fpls.2018.01310

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

Davies, R. W., Kucka, M., Su, D. W., Shi, S. N., Flanagan, M., Cunniff, C. M., et al. (2021). Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* 53, 1104-+1111. doi:10.1038/s41588-021-00877-0

Dige, M. S., Kumar, A., Kumar, P., Dubey, P. P., and Bhushan, B. (2012). Estimation of variance components and genetic parameters for growth traits in New Zealand white rabbit (*Oryctolagus cuniculus*). *J. Appl. Animal Res.* 40, 167–172. doi:10.1080/09712119.2011.645037

El Nagar, A. G., Sanchez, J. P., Ragab, M., Minguez, C., and Baselga, M. (2020). Genetic variability of functional longevity in five rabbit lines. *Animal* 14, 1111–1119. doi:10.1017/S1751731119003434

Gorjanc, G., Cleveland, M. A., Houston, R. D., and Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Sel. Evol.* 47, 12. doi:10.1186/s12711-015-0102-z

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15, 30. doi:10.1186/1471-2156-15-30

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92, 433–443. doi:10.3168/jds.2008-1646

He, J., Lopes, F. B, and Wu, X. L. (2019). Methods and applications of animal genomic mating. *Yi Chuan* 41, 486–493. doi:10.16288/j.yczz.19-053

He, J., Wu, X. L., Zeng, Q. H., Li, H., Ma, H. M., Jiang, J., et al. (2020). Genomic mating as sustainable breeding for Chinese indigenous Ningxiang pigs. *Plos One* 15, e0236629. doi:10.1371/journal.pone.0236629

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi:10.2135/cropsci2013.03.0195

Jannink, J. L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genomics* 9, 166–177. doi:10.1093/bfgp/elq001

Jia, Y., and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192, 1513–1522. doi:10.1534/genetics.112.144246

Karimi, K., Sargolzaei, M., Plastow, G. S., Wang, Z., and Miar, Y. (2019). Opportunities for genomic selection in American mink: A simulation study. *PLoS One* 14, e0213873. doi:10.1371/journal.pone.0213873

Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Mamrutha, H. M., Kumar, S., et al. (2021). Integrated genomic selection for rapid improvement of crops. *Genomics* 113, 1070–1086. doi:10.1016/j.ygeno.2021.02.007

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. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Liu, B., Shen, L., Guo, Z., Gan, M., Chen, Y., Yang, R., et al. (2021). Single nucleotide polymorphism-based analysis of the genetic structure of Liangshan pig population. *Anim. Biosci.* 34, 1105–1115. doi:10.5713/ajas.19.0884

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., et al. (2018). Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell.* 175, 347–359. doi:10.1016/j.cell.2018.08.016

Mancin, E., Sosa-Madrid, B. S., Blasco, A., and Ibanez-Escriche, N. (2021). Genotype imputation to improve the cost-efficiency of genomic selection in rabbits. *Anim. (Basel)* 11, 803. doi:10.3390/ani11030803

Meier, J. I., Salazar, P. A., Kucka, M., Davies, R. W., Dreau, A., Aldas, I., et al. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2015005118. doi:10.1073/pnas.2015005118

Meuwissen, T., Hayes, B., and Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci.* 1, 221–237. doi:10.1146/annurev-animal-031412-103705

Montes-Vergara, D. E., Hernndez-Herrera, D. Y., and Hurtado-Lugo, N. A. (2021). Genetic parameters of growth traits and carcass weight of New Zealand white rabbits in a tropical dry forest area. *J. Adv. Vet. Anim. Res.* 8, 471–478. doi:10.5455/javar.2021.h536

Nandolo, W., Meszaros, G., Banda, L. J., Gondwe, T. N., Lamuno, D., Mulindwa, H. A., et al. (2019). Timing and extent of inbreeding in african goats. *Front. Genet.* 10, 537. doi:10.3389/fgene.2019.00537

Nicod, J., Davies, R. W., Cai, N., Hassett, C., Goodstadt, L., Cosgrove, C., et al. (2016). Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nat. Genet.* 48, 912–918. doi:10.1038/ng.3595

Niranjan, S. K., Sharma, S. R., and Gowane, G. R. (2011). Estimation of genetic parameters for wool traits in Angora rabbit. *Asian-Australas. J. Anim. Sci.* 24, 1335–1340. doi:10.5713/ajas.2011.10456

Okeke, U. G., Akdemir, D., Rabbi, I., Kulakow, P., and Jannink, J. L. (2017). Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genet. Sel. Evol.* 49, 88. doi:10.1186/s12711-017-0361-y

Singh, A., Kumar, A., Mehrotra, A., Karthikeyan, A., Pandey, A. K., Mishra, B. P., et al. (2021). Estimation of linkage disequilibrium levels and allele frequency distribution in crossbred Vrindavani cattle using 50K SNP data. *Plos One* 16, e0259572. doi:10.1371/journal.pone.0259572

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86, 2447–2454. doi:10.2527/jas.2007-0010

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice (oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11, e1004982. doi:10.1371/journal.pgen.1004982

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

Wang, D., Xie, K. R., Wang, Y. Y., Hu, J. Q., Li, W. Q., Zhang, Q., et al. (2022). Cost-effectively dissecting the genetic architecture of complex wool traits in rabbits by low-coverage sequencing. *BIORXIV*, 483689.doi:10.1101/2022.03.09.483689

Wang, Q., Yu, Y., Yuan, J., Zhang, X., Huang, H., Li, F., et al. (2017). Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. *BMC Genet.* 18, 45. doi:10.1186/s12863-017-0507-5

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi:10.1534/genetics.112.146290

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

Yang, A. Q., Chen, B., Ran, M. L., Yang, G. M., and Zeng, C. (2020). The application of genomic selection in pig cross breeding. *Yi Chuan* 42, 145–152. doi:10.16288/j.yczz.19-253

Zhao, H. L., Huang, D. W., Chen, S., Cheng, G. L., Yang, Y. X., Zhao, X. W., et al. (2016). "Wan strain Angora rabbit - a novel breed in China," in World Rabbit Science Association Proceedings 11th World Rabbit Congress, Qingdao, China, June, 15-18, 2016.()

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 09 September 2022
DOI 10.3389/fgene.2022.994466

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Bayode O. Makanjuola,
Michigan State University, United States
Satish Kumar Illa,
Sri Venkateswara Veterinary University, India

*CORRESPONDENCE
George R. Wiggans,
George.Wiggans@UScdcb.com

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

RECEIVED 14 July 2022
ACCEPTED 18 August 2022
PUBLISHED 09 September 2022

CITATION
Wiggans GR and Carrillo JA (2022), Genomic selection in United States dairy cattle.
Front. Genet. 13:994466.
doi: 10.3389/fgene.2022.994466

Genomic selection in United States dairy cattle

George R. Wiggans* and José A. Carrillo

Council on Dairy Cattle Breeding, Bowie, MD, United States

The genomic selection program for dairy cattle in the United States has doubled the rate of genetic gain. Since 2010, the average annual increase in net merit has been $85 compared to $40 during the previous 5 years. The number of genotypes has been rapidly increasing both domestically and internationally and reached over 6.5 million in 2022 with 1,134,593 submitted in 2021. Evaluations are calculated for over 50 traits. Feed efficiency (residual feed intake), heifer and cow livability, age at first calving, six health traits, and gestation length have been added in recent years to represent the economic value of selection candidates more accurately; work is underway to develop evaluations for hoof health. Evaluations of animals with newly submitted genotypes are calculated weekly. In April 2019, evaluations were extended to crossbreds; to support that effort, evaluations are initially calculated on an all-breed base and then blended by an estimated breed composition. For animals that are less than 90% of one breed, the evaluation is calculated by weighting contributions of each of the five major dairy breeds evaluated (Ayrshire, Brown Swiss, Guernsey, Holstein, and Jersey) by the breed proportion. Nearly 200,000 animals received blended evaluations in July 2022. Pedigree is augmented by using haplotype matching to discover maternal grandsires and great-grandsires. Haplotype analysis is also used to discover undesirable recessive conditions. In many cases, the causative variant has been identified, and results from a gene test or inclusion on a genotyping chip improves the accuracy of those determinations for the current 27 conditions reported. Recently discovered recessive conditions include neuropathy with splayed forelimbs in Jerseys, early embryonic death in Holsteins, and curly calves in Ayrshires. Techniques have been developed to support rapid searches for parent-progeny relationships and identical genotypes among all likely genotypes, which substantially reduces processing time. Work continues on using sequence data to discover additional informative single nucleotide polymorphisms and to incorporate those previously discovered. Adoption of genotyping by sequencing is expected to improve flexibility of marker selection. The success of the Council on Dairy Cattle Breeding in conducting the genetic evaluation program is the result of close cooperation with industry and research groups, including the United States Department of Agriculture, breed associations, genotyping laboratories, and artificial-insemination organizations.

KEYWORDS

genomic selection, dairy cattle, genetic gain, genetic-economic index, breed composition, ancestor discovery, haplotype, recessive discovery

1 Introduction

Genomic selection has revolutionized dairy cattle breeding by doubling the rate of genetic gain primarily through halving the generation interval. In the United States, the Council on Dairy Cattle Breeding (CDCB) conducts a genetic evaluation program that includes genotypes from all over the world. The number of genotypes in the collection has been rapidly increasing and reached 6.6 million in August 2022 with 1,134,593 submitted in 2021 (Council on Dairy Cattle Breeding, 2022b). Continuous refinement of the program involves incorporating research results to improve accuracy, exploiting technological advances, and adapting to changes in the industry. Changes include extending evaluations to crossbreds, increasing the frequency of evaluation, revising the set of genotype markers used, adding evaluations for health, reproduction and feed efficiency traits, updating genetic indexes to improve ranking based on economic value, detecting additional deleterious genetic factors, augmenting pedigree by discovering ancestors, and providing breed composition information. This article is an update and expansion of Wiggans (2017).

2 Characteristics of genomic evaluation system in the United States

The first official genomic evaluations were released in January 2009 for Holsteins and Jerseys. Figure 1 shows the growth in number of genotypes submitted (excluding withdrawn) by year. In 2008 and 2009, more genotypes for bulls were received than for cows, however, in later years, the number of bull genotypes received has been nearly constant while the number of genotypes of females received increased rapidly. Figure 2 shows the number of genotypes submitted as of June 2022 by global region. Genomic evaluations were rapidly accepted by the dairy industry as the basis for selecting service sires. In just a few years, the majority of breedings were to bulls with only genomic evaluations (Figure 3).

The genotyping chips used for dairy cattle also have evolved. To reduce the cost of genotyping, chips with fewer single nucleotide polymorphism (SNP) markers were introduced after the initial 50 K chip (54,001 SNPs). As technology advanced, higher density chips were offered. Typically, bulls that are marketed, have two genotypes: the first to determine if they rank high enough to be marketable and the second with higher density to maximize the accuracy of their evaluation by minimizing imputation errors. Figure 4 shows the distribution of chip densities for genotypes received in 2021.

Generation interval is the average age of parents when offspring are born and impacts genetic improvement. The shorter the generation interval, the faster progress can be made so long as accuracy is not compromised excessively. A few years after genomic evaluations became official in 2009, parent ages for bulls began to drop dramatically and are now near the biological minimum (Figure 5).

Genomic selection has produced a large increase in genetic trend as indicated by the average net merit of marketed Holstein bulls in the United States (Figure 6). The $85 annual trend for Holstein bulls that entered artificial-insemination service since 2011 is more than double that for the period from 2005 through 2009 period ($40), which was already a substantial improvement over the $13 average gain for the period from 2000 through 2004.

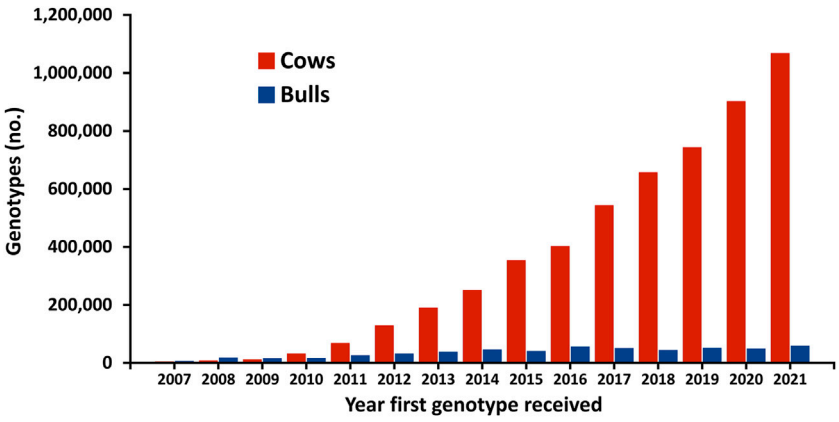


FIGURE 1
Number of dairy cattle genotypes submitted in the United States by year that first genotype was received.

2.1 Traits evaluated

The national genetic evaluations calculated by the CDCB currently include over 50 traits (Council on Dairy Cattle Breeding, 2022a). Production traits include milk yield, fat yield and percentage, and protein yield and percentage. Milking speed is collected only through the Brown Swiss type program so is only evaluated for that breed. Eighteen conformation (type) traits are included for non-Holstein breeds; Holstein conformation evaluations are calculated by Holstein Association USA (2022). Longevity traits include productive life, cow livability, and heifer livability (birth to first calving, added in 2020). Fertility traits include daughter pregnancy rate, cow conception rate, calving to first insemination, gestation length, and early first calving (added in 2019); male fertility is evaluated phenotypically as service-sire relative conception rate. The calving traits of dystocia (calving ease) and stillbirth rate are combined into a calving ability index. In addition to traditional evaluations for somatic cell score as a measure of mastitis resistance, evaluations for other health traits were introduced in 2018: displaced abomasum, ketosis, mastitis, metritis, milk fever (hypocalcemia), and retained placenta. In 2020, the trait feed saved was added as a measure of genetic merit for feed efficiency; it combines evaluations of body weight composite and residual feed intake (Council on Dairy Cattle Breeding, 2020). Most traits make a direct contribution to economic value. Gestation length is not included in the economic indexes; however, it is correlated with calving traits and may be useful in pasture-based systems to assist in determining calving date.

2.2 Genetic-economic indices

Lifetime genetic-economic indices are provided to the dairy industry for net merit, fluid merit, cheese merit, and grazing merit (Vanraden et al., 2021). Those indices rank animals based on their combined genetic merit for economically important traits. Multiple indexes are provided to support selection in a range of management and milk payment schemes. The indices are updated periodically to include new traits and to reflect prices expected in the next few years. The most recent update was in August 2021 and included information for the newly evaluated traits of feed saved, heifer livability, and early first calving (Figure 7).

2.3 Evaluation calculation features

Genomic evaluations are based on estimation of allele substitution effects for 78,964 SNPs selected considering minor allele frequency, distribution across the genome, linkage to genes of particular interest, and reliability

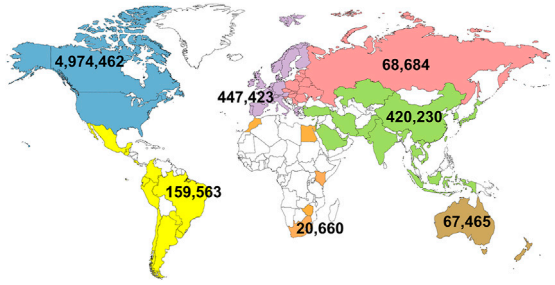


FIGURE 2
Total number of genotypes that have been submitted in the United States by global region as of June 2022.

considering call rate and Mendelian consistency. The sum of SNP effects, called direct genomic value, is combined with an estimate of polygenetic effects and the traditional evaluation to create the genomic evaluation (Vanraden, 2008). This is called the two-step method because traditional evaluations are calculated without the genomic data, which prevents consideration of selection based on genomic information and could cause selection bias. A single-step approach has been developed to allow simultaneous consideration of the genotypes and trait observations (Legarra and Ducrocq, 2012). Adapting the one-step method to the massive US dataset is an ongoing research project.

2.3.1 Estimation of breed composition

Breed composition is estimated with the same set of 78,964 SNPs used to determine the direct genomic values for other traits. The predictor population is purebred bulls, and the data are defined as one for the animal's breed and 0 otherwise. Solutions (as percentages) are forced to add to 100, and portions of less than 2% are distributed to remaining breeds. The percentages are called breed base representation (BBR). They are used to validate the breed of the identification data for an animal and weight individual breed contribution to the evaluations of crossbreds.

2.3.2 Evaluation of crossbreds

Because SNP effects differ by breed, genomic evaluations are calculated separately by breed. To provide evaluations for crossbreds, SNP effects from the individual breeds are combined. Animals with a highest BBR of less than 89.5% are evaluated as crossbreds, and their evaluations are a blend of their direct genomic values weighted by their BBRs for each breed. This blending is possible for traits that are initially evaluated on an all-breed base. Type traits and traits that are not calculated for all breeds are not blended; therefore, the animal receives an evaluation for the breed with the highest BBR. For first-generation crosses, the breed from the preferred identification

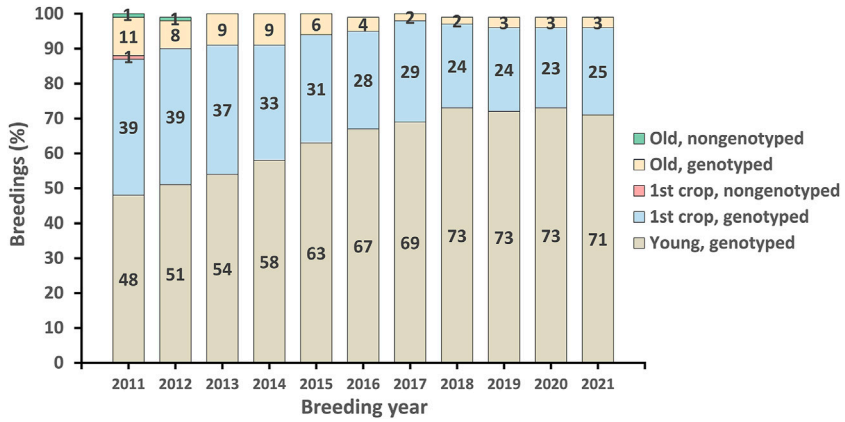


FIGURE 3
Genomic profile of Holstein service sires used for artificial-insemination breeding in the United States since 2011.

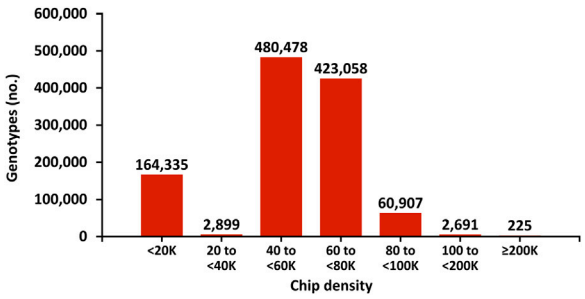


FIGURE 4
Marker density of genotyping chips for genotypes received in 2021 in the United States.

determines the breed of evaluation. Nearly 200,000 animals received blended evaluations in July 2022.

2.4 Pedigree validation and discovery

Each genotype is compared with the genotypes of parents and then with all existing genotypes that might have a parent-progeny relationship or be identical. To limit the time required, birth date limits are imposed. If both parents are confirmed, the search is limited to genotypes of animals born no more than 5 years before. That limit is increased to 12 years for other animals. Genotypes of bulls born more than 5 years ago without progeny born in the preceding 5 years are skipped unless their genotype was added to the evaluation system within the last year. No animals with genotyped progeny are skipped. Animals with conflicting parents or discovered relationships are not evaluated until the conflicts are resolved. In general, for a conflicting pair, the genotype with the less reliable information is the one designated as not usable. If a parent is not genotyped or not confirmed, the likelihood of the grandsire is determined. If the grandsire is unlikely, the animal is not evaluated.

Discovery of maternal grandsires (MGS) and maternal great-grandsires (MGGS) is done as part of the evaluation based on haplotypes in common. An imputation process is used to create genotypes with 78,964 SNPs from incoming genotypes of various lengths. The genome is divided into intervals and maternal or paternal origin is determined. Those haplotypes are compared, and bulls are designated as discovered ancestors based on the percentage of haplotypes in common. Crossing over reduces the expected haplotypes in common to 45% for MGS and 20% for MGGS. A bull's percentage must exceed the next highest bull's by 15% and have a percent matches greater than 35% for MGS and 15% for MGGS to be designated as discovered. The age of the bull at the birth of the grand progeny/great grand progeny is considered. The discoveries are used to remove an unlikely grandsire designation if the discovered grandsire is the same as the pedigree grandsire. If no pedigree information on a dam or granddam was provided, the discovered MGS and MGGS are added to the pedigree. Similarly, if the connecting dam is unknown, identification data will be constructed so that a pedigree record can be created to store the MGS or MGGS information.

To speed discovery, a set of 3,552 SNPs that are present on most genotyping chips and have high call rate and good Mendelian consistency were selected. Comparisons are ordered so that checking stops after 96 or 1,000 SNPs if the percentage of conflicts exceeds that likely for a parent-progeny relationship. The discovered closely related pairs are stored using a unique genotype identification so that the identity of close relatives is not affected by genotype reassignment.

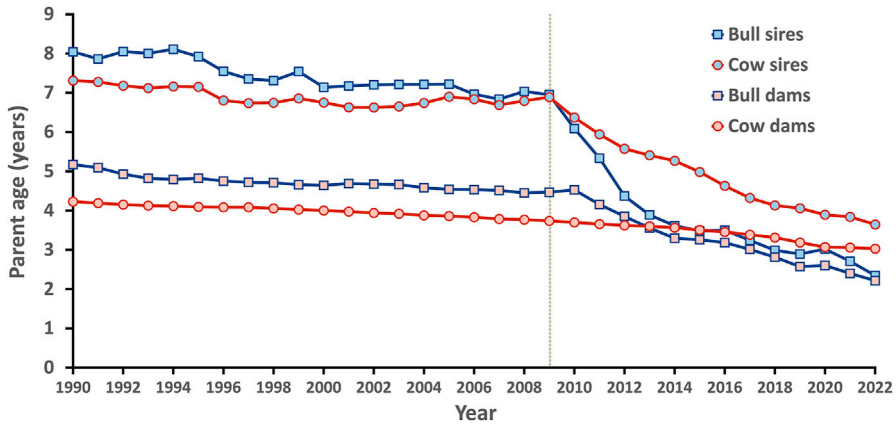


FIGURE 5
Generation intervals for Holsteins in the United States by sex and year.

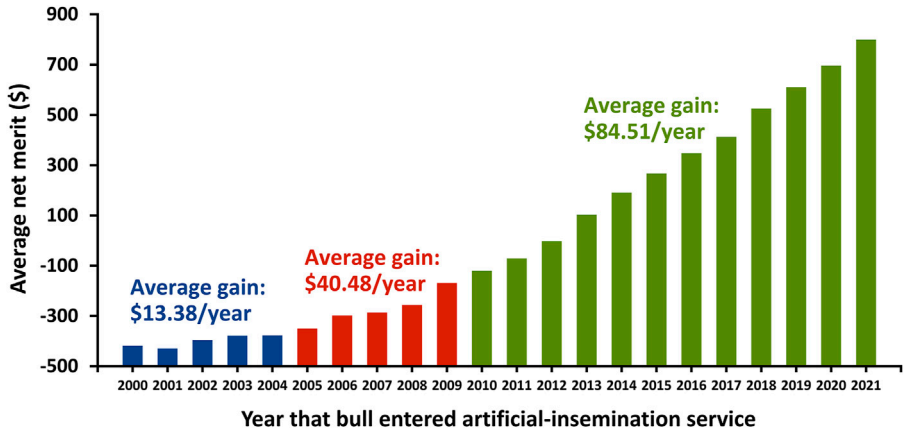


FIGURE 6
Gain in net merit for marketed Holstein bulls in the United States by entry year into artificial-insemination service.

2.5 Mating decisions

The CDCB provides information on the genomic relationships between potential dams and currently marketed bulls. Those data report the actual portion of genetic variants (alleles) in common, in contrast to pedigree analysis, which can only give the average based on relationships. This allows avoiding inbreeding more precisely.

The CDCB also provides predictions for a number of recessive conditions so that likely carrier-to-carrier matings can be avoided even without testing an animal. The haplotypes that affect fertility are conditions discovered through genomics and can now be considered in matings. Currently, 27 conditions are reported (Cole et al., 2022). Recently added recessives include JNS (neuropathy with splayed forelimbs in Jerseys), HH6 (early embryonic death in Holsteins), and AHC (curly calves in Ayrshires).

Future mating programs may also consider the effects of dominance, which causes some sire-MGS grandsire combinations to do better than expected and others to do worse.

3 Methods to increase evaluation accuracy

Genomic evaluation relies on having enough DNA markers to track the segments of chromosomes associated with high performance across generations. The SNPs are used as markers because of their reliability and low cost. However, crossing over (recombination) can disrupt the associations

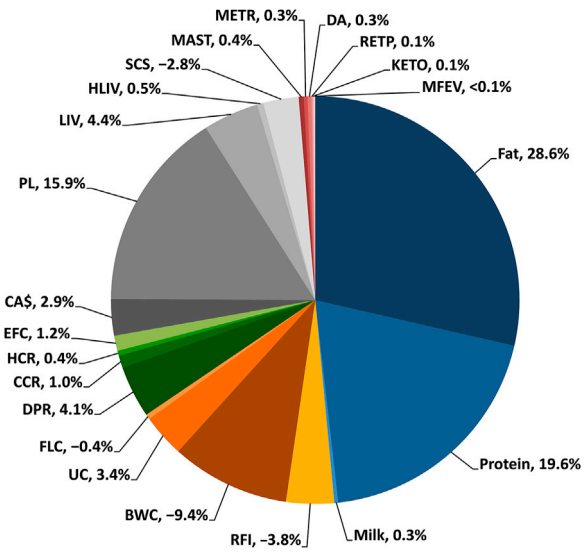


FIGURE 7
Relative emphasis on traits in the net merit (NM$) genetic-economic index revised in August 2021 by the Council on Dairy Cattle Breeding. Body weight composite, BWC; calving ability (calving ease and stillbirth rate), CA$; cow conception rate, CCR; cow livability, LIV; daughter pregnancy rate, DPR; displaced abomasum, DA; early first calving, EFC; feet-legs composite, FLC; heifer conception rate, HCR; heifer livability, HLIV; ketosis, KETO; mastitis, MAST; metritis, METR; milk fever, MFEV; productive life, PL; residual feed intake, RFI; retained placenta, RETP; somatic cell score, SCS; udder composite, UC.

between a marker and the causative variant, which results in a decay in the linkage between SNPs and genomic regions associated with high performance. To counteract this, new data are needed so that the SNP effects can be re-estimated to maintain or improve evaluation accuracy. Now that the cost of whole-genome sequencing has fallen and thousands of animals have been sequenced, research is focused on finding SNPs that are more closely associated with the causative genetic variants (or even the variants themselves). The closer the marker is to the causative variant, the lower the likelihood that recombination will disrupt the linkage.

3.1 More traits

For traditional genetic evaluations, an animal could only receive an evaluation for a trait that was observed for the animal or its offspring. Evaluation were limited to traits where large scale collection of data was possible such and milk and fat yield and type traits collected by breed associations. With genomics, evaluations can be generated for all genotyped animals if enough animals have genotypes and traditional evaluations for the trait (the reference population) to give reasonably accurate estimates of the SNP effects. Feed efficiency is an example of a small population of animals with feed intake measured providing the basis for feed saved evaluations for all genotyped animals. Efforts to collect data for more traits are ongoing. Foot health and milking speed are currently in the research phase. Mid-infrared spectroscopy of milk samples is expected to provide data related to traits of economic importance.

3.2 Larger reference population

Early in the development of the US genomic evaluation system, arrangements were made to share genotypes between countries to increase the size of the reference population (Wiggans et al., 2011). From the beginning, all US genotypes have been shared with Canada. For Holsteins, sharing is ongoing with Italy, the United Kingdom, Switzerland, and Germany. The reference population for Holsteins is so large that the value of adding older animals has declined; however, the addition of younger animals is still beneficial. The greatest benefit from sharing genotypes and phenotypic data may be for feed efficiency because of the high cost of data collection.

3.3 Better and more data

Herds with high levels of genotyping generally have fewer misidentified sires in their data that contribute to traditional evaluations. With better data, less information is lost from the extensive checks done by the CDCB to eliminate unreliable and inconsistent data. Genotypes are checked against all other genotypes to ensure they are assigned to the correct animal and that the parents are correctly identified. Data problems can include submitting the same identification number for different cows, not documenting that sexed semen was used for an insemination and reporting the transfer of an embryo to a recipient as a normal breeding. Greater knowledge of data usage should help providers better understand the importance and benefit of accurate data collection.

In addition to data accuracy, comprehensive reporting is important. Although genomics can be used to provide evaluations for animals without an observed trait, continued submission of phenotypic data still is needed to maintain accuracy. In recognition of the value of data, the CDCB makes payments to dairy records processing centers for providing data, supports the collection of feed efficiency data, and structures the fees for genomic evaluation to give a discount to data contributors.

The CDCB has a quality assurance program for the genotyping laboratories and nominators that includes monthly report cards and annual reviews. To become certified, the organizations must demonstrate the ability to provide data in

the required formats. The monthly reports for labs include SNP accuracy and completeness as well as the percentage of genotypes with nomination and animal genotypes with a low call rate. The labs receive reports on those characteristics for each submission for possible correction before submissions are to added to the database.

4 Conclusion

The popularity of the genomic evaluation program in the United States has resulted in a rapid growth in the genotyping of dairy cattle. When genomic evaluation began in 2008, the focus was on bulls, but genotyping of females has grown rapidly in recent years. Many dairy producers genotype all their heifers so that they can select among a range of breeding and management strategies.

Dairy genetics in the United States and worldwide have been transformed by the use of genomic information. Genomic evaluations determine the value of animals at a much earlier age and have contributed to a dramatic increase in the rate of genetic improvement. Bulls are used widely as sires based on the analysis of their DNA before they have any milking daughters. A continuing stream of improvements are planned to increase accuracy and comprehensiveness of genomic evaluations. Success requires a partnership between data suppliers and users to generate the most effective information for all.

Author contributions

GW drafted the manuscript, and JC provided updated values for the figures. Both authors reviewed and approved the submitted manuscript.

Funding

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

Acknowledgments

The authors thank Suzanne Hubbard for detailed technical editing, graphics preparation, and reference review. Dairy producers in the United States and allied industry and research groups including the United States Department of Agriculture, breed associations, genotyping laboratories, and artificial-insemination organizations are thanked for providing the data that support this project.

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

Cole, J. B., VanRaden, P. M., Null, D. J., Hutchison, J. L., and Hubbard, S. M. (2022). Haplotype tests for economically important traits of dairy cattle. AIP Res. Rep. Genomic5 (12-20). Available at: https://www.ars.usda.gov/ARSUserFiles/80420530/Publications/ARR/Haplotype%20tests_ARR-Genomic5.pdf. Accessed date July 10, 2022.

Council on Dairy Cattle Breeding (2020). Feed saved (FSAV). Trait reference sheet. Available at: https://www.uscdcb.com/wp-content/uploads/2020/11/CDCB-Reference-Sheet-Feed-Saved-12_2020.pdf. Accessed July 13, 2022.

Council on Dairy Cattle Breeding (2022a). Genetic evaluations. Available at: https://www.uscdcb.com/what-we-do/genetic-evaluations/. Accessed date July 13, 2022.

Council on Dairy Cattle Breeding (2022b). Genotype counts by chip type, breed code, and sex code in database as of 2022-06-27. Available at: https://queries.uscdcb.com/Genotype/cur_freq.html. Accessed date June 30, 2022.

Holstein Association USA (2022). Interpreting linear type trait STAs. Available at: https://www.holsteinusa.com/genetic_evaluations/ss_interpret_linear.html. Accessed date July 13, 2022.

Legarra, A., and Ducrocq, V. (2012). Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *J. Dairy Sci.* 95, 4629–4645. doi:10.3168/jds.2011-4982

Vanraden, P. M., Cole, J. B., Neupane, M., Toghiani, S., Gaddis, K. L., and Tempelman, R. J. (2021). Net merit as a measure of lifetime profit: 2021 revision. AIP Res. Rep. NM$8 (05-21). Available at: https://www.ars.usda.gov/ARSUserFiles/80420530/Publications/ARR/nmcalc-2021_ARR-NM8.pdf. Accessed date July 12, 2022.

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

Wiggans, G. R. (2017). *Forecasting the future of genomic selection*. Indianapolis, IN: DairyBusiness.

Wiggans, G. R., VanRaden, P. M., and Cooper, T. A. (2011). The genomic evaluation system in the United States: Past, present, future. *J. Dairy Sci.* 94, 3202–3211. doi:10.3168/jds.2010-3866

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 16 September 2022
DOI 10.3389/fvets.2022.984703

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Liu Jianbin,
Lanzhou Institute of Animal Science and Veterinary Medicine (CAAS), China
Hao Zhang,
China Agricultural University, China

*CORRESPONDENCE
Yanan Yang
yangyanan404@163.com
Shengguo Zhao
zhaosg@gsau.edu.cn

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

RECEIVED 02 July 2022
ACCEPTED 18 August 2022
PUBLISHED 16 September 2022

CITATION
Yuan H, Liu X, Wang Z, Ren Y, Li Y, Gao C, Jiao T, Cai Y, Yang Y and Zhao S (2022) Alternative splicing signature of alveolar type II epithelial cells of Tibetan pigs under hypoxia-induced. *Front. Vet. Sci.* 9:984703. doi: 10.3389/fvets.2022.984703

Alternative splicing signature of alveolar type II epithelial cells of Tibetan pigs under hypoxia-induced

Haonan Yuan[1], Xuanbo Liu[1], Zhengwen Wang[1], Yue Ren[2], Yongqing Li[3], Caixia Gao[4], Ting Jiao[1,5], Yuan Cai[1], Yanan Yang[1]* and Shengguo Zhao[1]*

[1]College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China, [2]Academy of Agriculture and Animal Husbandry Sciences, Institute of Animal Husbandry and Veterinary Medicine, Lhasa, China, [3]Xinjiang Academy of Animal Sciences, Xinjiang, China, [4]State Key Laboratory of Veterinary Biotechnology, Harbin Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Harbin, China, [5]College of Grassland Science, Gansu Agricultural University, Lanzhou, China

Alternative splicing (AS) allows the generation of multiple transcript variants from a single gene and affects biological processes by generating protein diversity in organisms. In total, 41,642 AS events corresponding to 9,924 genes were identified, and SE is the most abundant alternatively spliced type. The analysis of functional categories demonstrates that alternatively spliced differentially expressed genes (DEGs) were enriched in the MAPK signaling pathway and hypoxia-inducible factor 1 (HIF-1) signaling pathway. Proteoglycans in cancer between the normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups, such as *SLC2A1*, *HK1*, *HK2*, *ENO3*, and *PFKFB3,* have the potential to rapidly proliferate alveolar type II epithelial (ATII) cells by increasing the intracellular levels of glucose and quickly divert to anabolic pathways by glycolysis intermediates under hypoxia. *ACADL*, *EHHADH*, and *CPT1A* undergo one or two AS types with different frequencies in ATII cells between TN and TL groups (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups), and a constant supply of lipids might be obtained either from the circulation or *de novo* synthesis for better growth of ATII cells under hypoxia condition. *MCM7* and *MCM3* undergo different AS types between LN and LL groups (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups), which may bind to the amino-terminal PER-SIM-ARNT domain and the carboxyl terminus of *HIF-1α* to maintain their stability. Overall, AS and expression levels of candidate mRNAs between Tibetan pigs and Landrace pigs revealed by RNA-seq suggest their potential involvement in the ATII cells grown under hypoxia conditions.

KEYWORDS

alternative splicing, hypoxia, ATII cells, swine, MAPK signaling pathway, glycolysis/gluconeogenesis

Introduction

Tibetan pigs adapt well to hypoxic environments compared to other pigs, as the native breeds live in the Qinghai-Tibet Plateau (1). Studies have shown that Tibetan pigs have evolved typical characteristics to adapt to high-altitude hypoxia, especially with developed lungs, denser pulmonary arterioles, and larger alveoli (2, 3). Hypoxia could induce epithelial injury, influence alveolar homeostasis, and cause a series of pulmonary diseases, such as pulmonary hypertension (4, 5), chronic obstructive pulmonary disease (6, 7), and pulmonary fibrosis (8). Alveolar type I epithelial (ATI) and alveolar type II epithelial (ATII) cells have covered the alveolar surface. ATII can transform into ATI and is responsible for the lungs' repair, recycling, and production (9, 10). ATII could undergo cell death and replace by myofibroblasts in hypoxia-induced IPF, which prevents the repairing and renewal of the alveolar wall (11). The injury of regeneration and transdifferentiation in alveolar epithelial cells are vital points that lead to the disease under hypoxia-induced, which may result in breaks in epithelial basement membranes of alveoli (9). Activation of endoplasmic reticulum stress (12), a different expression of ROS (13), and hemoglobin (14) could involve in the oxygen-sensing pathway in alveolar epithelial cells. Alternative splicing (AS) is one of the essential mechanisms in post-transcriptional regulation and could be regulated by many biotic and abiotic stress factors, especially tightly associated with hypoxic adaptation of cells (15). For example, splicing targets of alternative first exon usage, exon skipping, and intron retention could potentially contribute to cancer cell hypoxic adaptation by promoting cancer cell proliferation, transcriptional regulation, and migration (16–18). Large-scale alterations in alternative splicing of ribosomal protein mRNAs were influenced by hypoxia (19). Promotes expression of the angiogenesis inhibitory alternatively spliced hypoxia-inducible factor-3α (*HIF-3α)* IPAS isoform, and *HIF-1α* splicing during angiogenesis could be regulated by hypoxia (18, 20). Recent studies have identified alternative splicing events that exist in lung (21–23), heat (24), and ovary (25). Until now, the analysis of alternative splicing in ATII was rarely reported. Here, we carried out a comparative study of AS in ATII during normoxic (21% O_2) and hypoxic (21% O_2) to explore the patterns and conservation of AS between Tibetan pigs and Landrace pigs. Our results supported further development of hypoxia-associated splicing events in ATII, representing one of the steps forward in the hypoxic adaptation of Tibetan pigs.

Abbreviations: DEGs, differentially expressed genes; TN, ATII cells of Tibetan pigs were cultured under 21% O_2; TL, ATII cells of Tibetan pigs were cultured under 2% O_2; LN, ATII cells of Landrace pigs were cultured under 21% O_2; LL, ATII cells of Landrace pigs were cultured under 2% O_2.

Materials and methods

Samples

Alveolar type II epithelial primary cells from newborn male Tibetan pigs and Landrace pigs were isolated and cultured as described previously (26) with minor modifications. ATII cells were collected at 48 h, which were cultured under normoxic conditions (21% O_2, 5% CO_2, and 79% N_2) between Tibetan pigs (TN, $n = 3$) and Landrace pigs (LN, $n = 3$), and under hypoxic conditions (2% O_2, 5% CO_2, and 98% N_2) between Tibetan pigs (TL, $n = 3$) and Landrace pigs (LL, $n = 3$), respectively.

RNA extraction, library construction, and sequencing

Total RNA was extracted from ATII cells using a TRIzol reagent kit (Invitrogen, Carlsbad, CA, USA), treated, removed, and precipitated using DNase I (NEB, Beijing, China) phenol-chloroform, and ethanol. Total RNA quality was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) and checked using Rnase-free agarose gel electrophoresis.

The mRNAs and non-coding RNAs (ncRNAs) were obtained by removing ribosome RNAs (rRNAs) from total RNA, fragmented into short fragments using fragmentation buffer and reverse transcribed into complementary DNA (cDNA) with random primers, and synthesized to second-strand cDNA. Next, the cDNA fragments were ligated to Illumina sequencing adapters by purifying with a QiaQuick PCR extraction kit (Qiagen, Venlo, The Netherlands), and the second-strand cDNA was digested. The twelve cDNA libraries were generated, purified, and sequenced using Illumina HiSeq^{TM} 4000 by Gene Denovo Biotechnology Co. (Guangzhou, China).

Relative abundance of mRNA

Clean, high-quality reads were obtained and filtered from raw reads using fastp (27) (version 0.18.0) and removing the rRNA mapped reads to the rRNA database. The RefSeq (*Sus scrofa* 11.1) databases were mapped using HISAT2 (28). Transcripts reconstruction was carried out with software Stringtie and HISAT2. HTSeq counted the number of reads aligned to each gene and exon. A fragment per kilobase of transcript per million mapped reads (FPKM) value was calculated to quantify its expression abundance. We carried out differentially expressed genes (DEGs) using a threshold of $|\log2(\text{fold\_change})| \geq 2$ and a false discovery rate (FDR) adjusted p-value $< 5\%$.

Identification of AS types and counts

Paired-end raw data were first evaluated using FastQC v0.11.8 (29), and quality control using the FASTX toolkit to trim bases in 5' sequences and trimmomatic to trim adaptor sequences and low-quality reads (30, 31). High-quality reads were aligned to the reference genome sequence (*Sus scrofa* 11.1) and merged using TopHat2 v2.1.1 (32) and Cufflinks v2.2.1 (33). Differential AS events were identified and analyzed using rMATS (version 4.0.1, http://rnaseq-mats.sourceforge.net/index.html) and AS variations of each transcription region by using StringTie software among four groups. The FDR < 0.05 in the comparison was used to identify different AS events. The classification of AS was as follows: alternative $5'$ splice sites (A5SSs), alternative 3' splice sites (A3SSs), retained introns (Ris), skipped exons (Ses), and mutually exclusive exons (MXEs) were the main categories of selective splicing.

Enrichment and integrative analysis of the alternatively spliced DEmRNAs regulatory network

We analyzed alternatively spliced DEGs using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) in the online tool database for annotation, visualization, and integrated discovery (DAVID, version 6.7, https://david.ncifcrf.gov/). GO was used to determine and explore the functions of the alternatively spliced DEGs as molecular function, biological process, and cellular component. KEGG analyzed alternatively spliced DEGs to reveal their roles, regulation processes, and enrichment in different biological pathways. The *p*-values <0.05 were considered significantly different enriched biological pathways. The co-expression regulatory network of alternatively spliced DEGs is generated using the PCC, and the diagram only shows the top 250. The potential regulatory network was constructed by Cytoscape (34).

qRT-PCR validation of AS events

The four groups randomly selected three alternatively spliced DEGs for Real-time quantitative reverse transcription polymerase chain reaction (qRT-PCR) verification. Total RNA was extracted from ATII cells to synthesize cDNA using a FastQuant cDNA first-strand synthesis kit (TianGen, China). The cDNA was subjected to qRT-PCR analysis. Transcript-specific primers (Supplementary Table S1 in Supplementary material 1) were designed based on the unique regions of selected alternatively spliced DEGs using Primer 5.0 software, *β-actin* was used as reference genes, and expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method. PCR conditions were performed as follows: 95°C for 30 s, forty cycles at 95°C for 5 s, 60°C for 30 s, and 72°C for 30 s.

Results

Forms of alternative splicing events

The average 11,245,202,725 bp clean data were obtained from 11,516,010,250 bp raw reads after filtering out low-quality data among twelve libraries (Supplementary Table S2 in Supplementary material 1). The 41,642 AS events corresponding to 9,924 genes were identified using genomic information and transcript data from the RNA-seq dataset (Figure 1, Supplementary material 2). Hypoxia-induced generally increased the number of AS events. Therefore, a total of five alternative splicing forms were obtained through data mapping analysis, such as A5SS, MXE, A3SS, SE, and RI, which revealed Ses as the most abundant event type (73.01%), followed by MXE (10.70%), A3SS (7.95%), and A5SS (5.02%); mutually RI occurred in only 3.32% of AS events among the four groups (Figure 2). Furthermore, 1,444, 2,192, 2,522, 954, and 9,238 alternatively spliced genes undergo A5SS, A3SS, MXE, RI, and SE events, respectively. The results demonstrated that almost all DEGs underwent at least one AS event. The frequencies of AS events were similar among different groups. The highest frequency (64) of AS events was in the *TNC* gene (ncbi_397460) in the TN group (Supplementary material 3), and the highest frequency of AS events was in the *OBSCN* gene among the four groups.

Alternatively, spliced DEGs in ATII cells response to hypoxia

The analysis found that most of the DEGs underwent AS events. Approximately, 33,985 AS events of the total expressed genes and 1,763 significant AS events were screened between TN and TL groups (Supplementary Table S3 in Supplementary material 1). We further selected the 1,470 intersection genes between normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups for significant hypoxia-related genes to identify their AS events, which revealed that 75.00% of them undergo diverse AS (Supplementary material 4). The AS of 901 intersection genes associated with hypoxia, such as *EPAS1*, *NREP*, and *VPS13B*, were only mediated by one or two events. Another 189 genes, such as *CCDC14*, *NKTR*, and *ATRX*, exhibited complex AS. For example, AS in *NFAT5*, *ECM1*, *ZBTB20*, *KMT2E*, *ZMYM1*, and *PLAGL1* were classified as five basic types between normoxic and hypoxic groups. We found that 233 differential splicing events of 4,514 AS circumstances were present in 1,470 differentially expressed

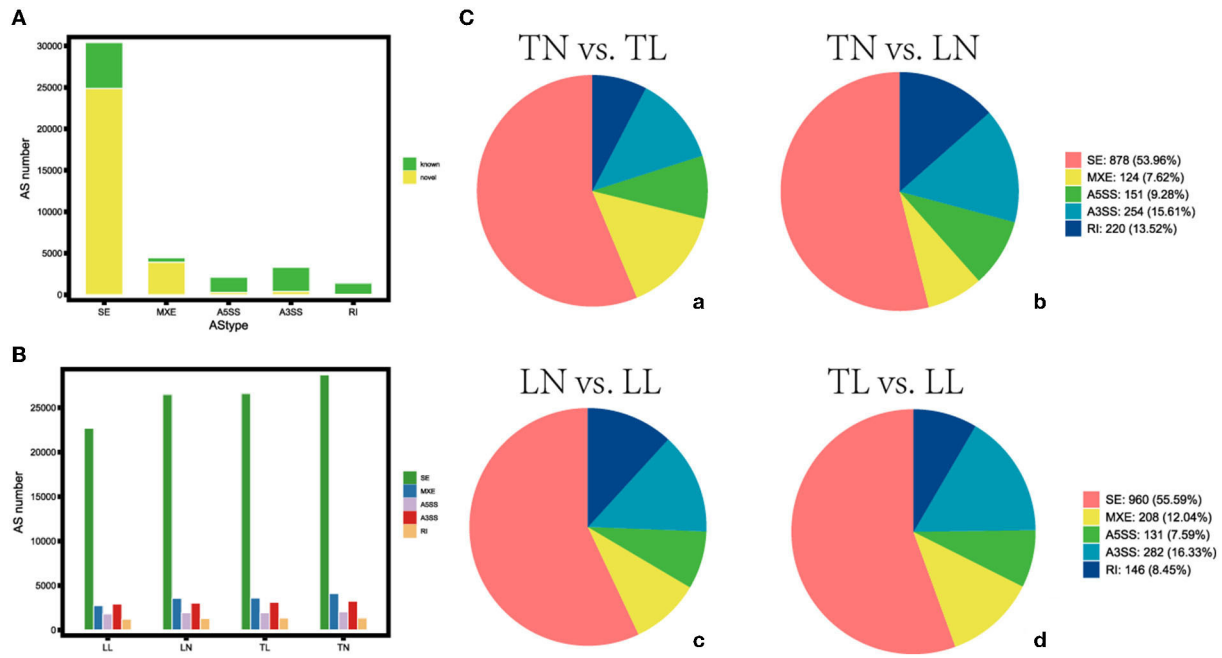


FIGURE 1
Distribution of the different types of alternative splicing (AS) events. **(A)** Distribution of the known and novel AS. **(B)** The number of AS events among four groups. **(C)** Different types of AS events (a) between TN and TL, (b) between TN and LN, (c) between LN and LL, and (d) between TL and LL groups.

intersection genes between the TN and TL groups (Figure 3, Supplementary Table S4 in Supplementary material 1).

GO and KEGG enrichment of alternatively spliced DEGs

The enrichment analyses of alternatively spliced DEGs were performed by GO analysis to investigate the biological function of AS events between normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups. The results showed that 817 biological processes, 163 molecular functions, and 114 cellular components were significantly enriched ($p < 0.05$) (Figures 4A,B). For AS genes of DEGs, biological processes were enriched considerably, such as regulation of nucleobase-containing compound metabolic process (GO: 0019219), nucleic acid metabolic process (GO: 0090304), and nucleobase-containing compound metabolic process (GO: 0006139). Several genes were significantly enriched in the nucleus (GO: 0005634), intracellular part (GO: 0044424), and intracellular (GO: 0005622) cellular component. Binding (GO: 0005488), heterocyclic compound binding (GO: 1901363), and organic cyclic compound binding (GO: 0097159) of molecular functions were most significantly enriched. In a comparison of TN and TL (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups), pyruvate metabolic process (GO: 0006090), binding (GO: 0005488), and intracellular part (GO: 0044424) of biological processes, molecular functions, and cellular components were most significantly enriched (Figures 4C,D). In a comparison of LN and LL (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups), cellular metabolic process (GO: 0044237), catalytic activity (GO: 0003824), and intracellular part (GO: 0044424) of biological processes, molecular functions, and cellular components were most significantly enriched (Figures 4E,F).

As the AS of mRNAs is directly related to functional characteristics, the function of alternatively spliced DEGs was analyzed by KEGG enrichment. A total of 279 pathways were enriched with 89 pathways significantly enriched ($p < 0.05$), of them MAPK signaling pathway (ko04010), HIF-1 signaling pathway (ko04066), and proteoglycans in cancer (ko05205) were most significantly enriched between normoxic (21% O_2, TN *vs.* LN) and hypoxic (2% O_2, TL *vs.* LL) groups (Figures 5A,B). When TL was compared with TN (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups) groups, alternatively spliced DEGs were found to be significantly enriched in carbon metabolism (ko01200), glycolysis/gluconeogenesis (ko00010), and fatty acid metabolism (ko01212) pathways (Figures 5C,D). Cell cycle (ko04110), metabolic pathways (ko01100), and RNA transport (ko03013) were most significantly enriched

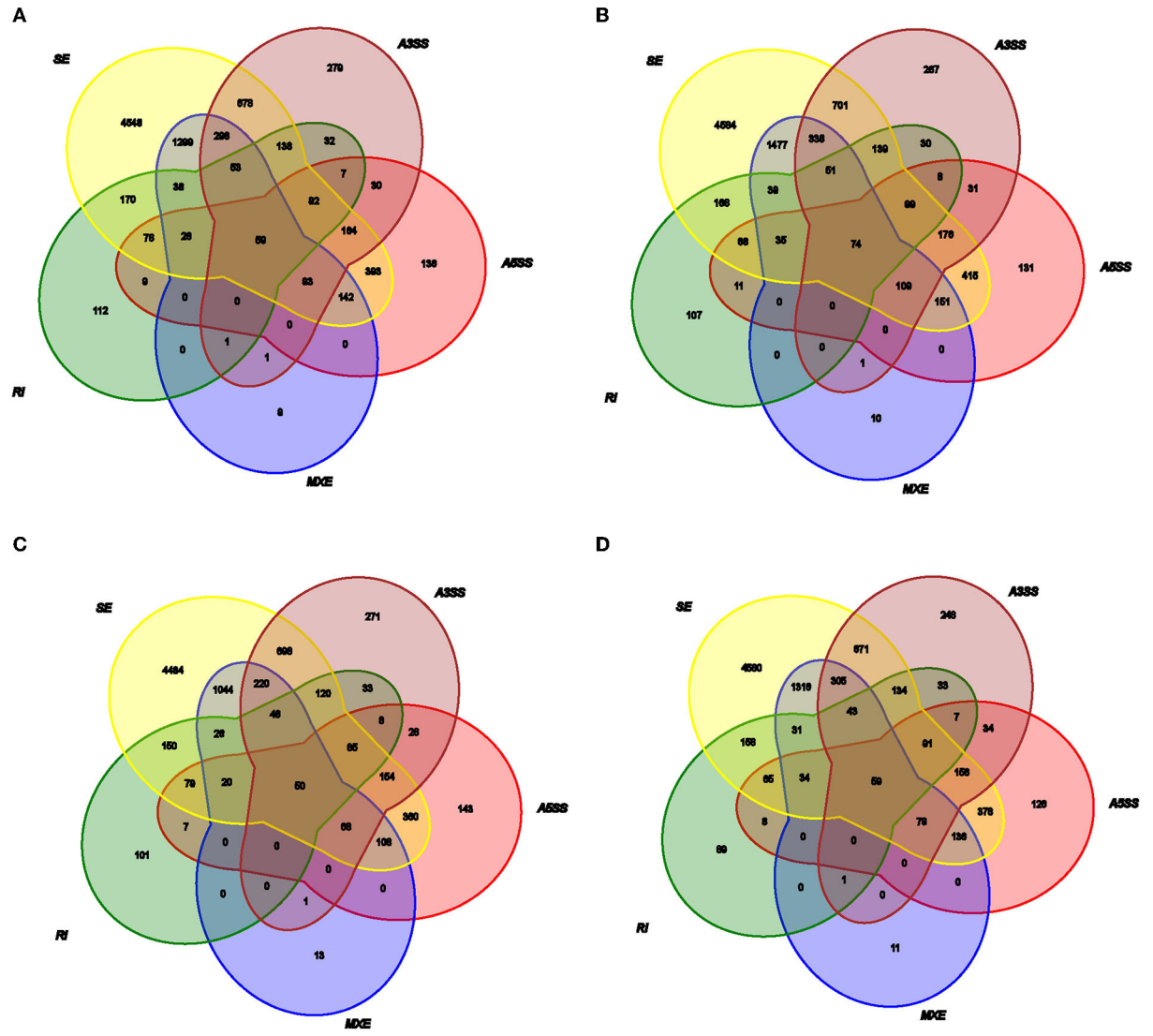


FIGURE 2
Relationship of the different types of AS events in **(A)** TL, **(B)** TN, **(C)** LL, and **(D)** LN.

by abundant genes between LN and LL groups (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups) (Figures 5E,F).

Coexpression network of alternatively spliced DEGs expression profiles

Three hypoxia-related co-expression networks of alternatively spliced DEGs were constructed. The top 250 relationship pair network diagrams are listed, such as comparison groups of normoxia and hypoxia, TN and TL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups), LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) (Figure 6, Supplementary Figures S1, S2). The intersection of comparisons between normoxic (TN, LN) and hypoxic (TL, LL) represented the main differences of ATII cells at different oxygen concentrations gradient. *ROCK2* (ncbi_397445), *KIF5B* (ncbi_595132), and *ZFP91* (ncbi_100525558) were selected as the most affected mRNAs, and there were strong correlations with several RNAs undergoing AS events between normoxic (TN, LN) and hypoxia (TL, LL) groups. Interestingly, *VCAN* (ncbi_397328), *HSD3B1* (ncbi_445539), and *FAM13C* (ncbi_100525364) were most significantly correlated with a large number of alternatively spliced DEGs between

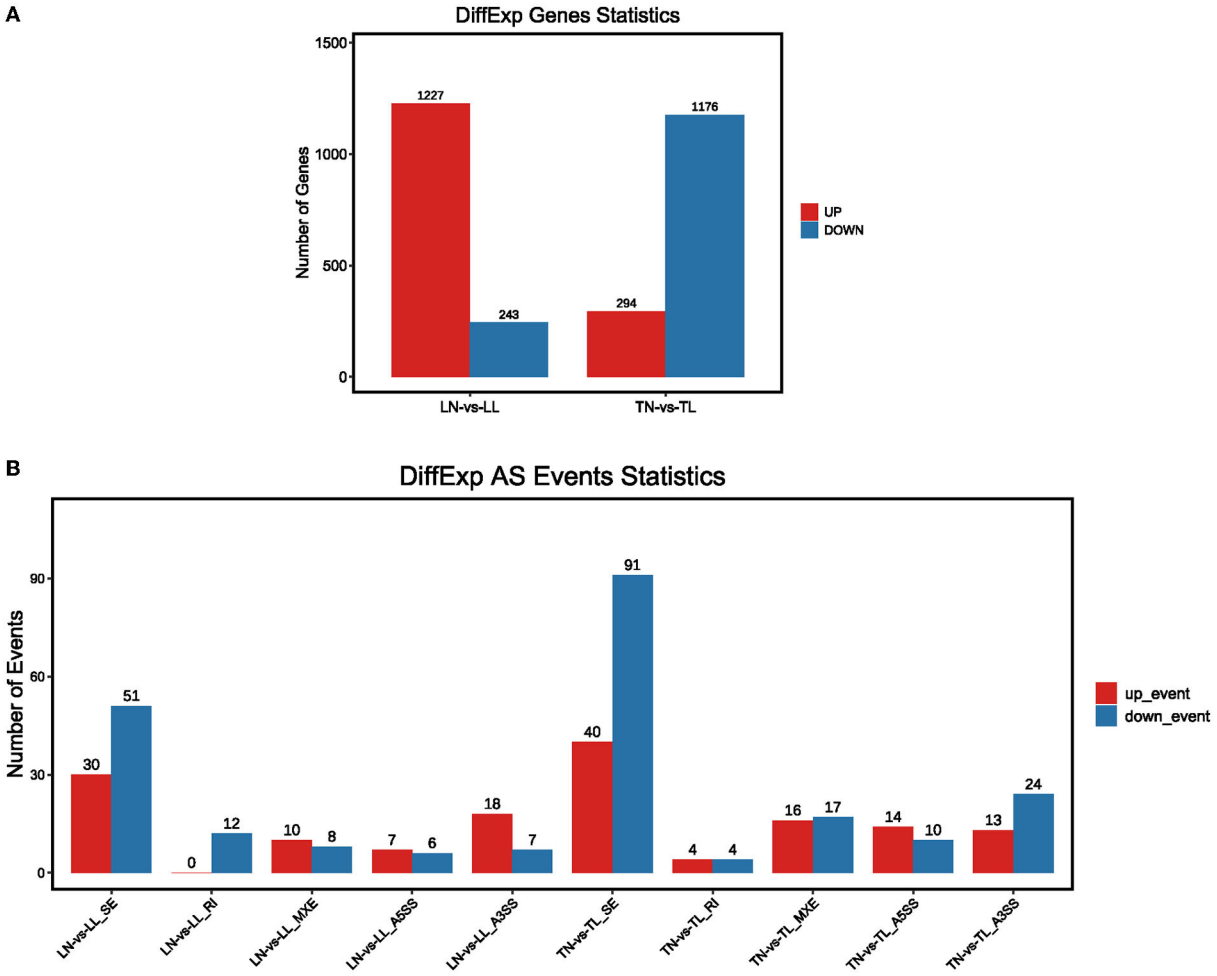


FIGURE 3
The presence of alternate spliced DEGs in the cross region of normoxic and hypoxic groups. **(A)** DEGs with alternate splicing were distributed in Tibetan pigs and landrace pigs in normoxia and hypoxia groups; **(B)** Distribution of DEGs with alternate splicing in five splicing events of Tibetan pigs and Landrace pigs in normoxia and hypoxia groups, respectively.

TN and TL groups (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups). Meanwhile, *ITGAV* (ncbi_397285), *ADAM9* (ncbi_397344), and *MYOF* (ncbi_100154898) were most significantly correlated with a large number of alternatively spliced DEGs between LN and LL groups (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups).

Verification of transcripts expression and AS events

Three alternatively spliced DEGs were randomly selected to further test the accuracy of RNA-seq data using qRT-PCR. *HP1BP3*, *NECTIN2*, and *DDX11* were predicted and identified as having two transcripts, and the type of alternative splicing is SE. The expression levels of the transcript with inclusion and skipping are higher than that of skipping transcript among four groups (Supplementary Figure S3), indicating that the alternative splicing prediction based on RNA-seq data was reliable.

Discussion

The identification, characterization, and post-transcriptional regulation of alternatively spliced DEGs were widely studied by attracting the interest of researchers (15, 35, 36), such as Xiang pig gilts, bovine, and human. Animals have a more complicated and more extensive intron than plants (37). Transcriptome survey reveals increased complexity

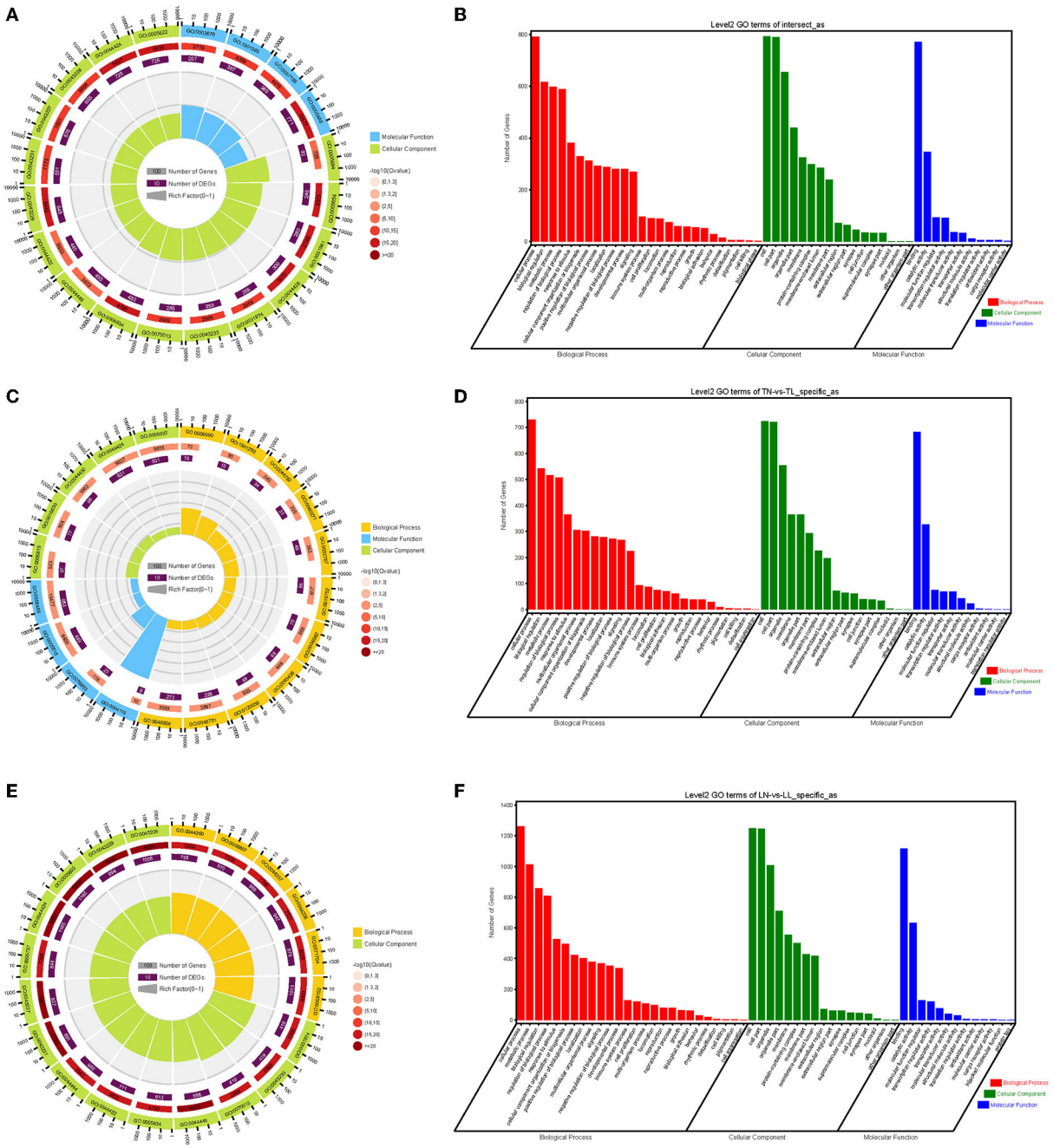


FIGURE 4
GO functional annotation **(A,B)** between normoxic (TN and LN) and hypoxic (TL and LL) groups, **(C,D)** between TN and TL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups), and **(E,F)** between LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) groups.

of the alternative splicing landscape in Arabidopsis (37–39), and their most common AS events were exon skipping and intron retention, respectively (40, 41). A5SS, MXE, A3SS, SE, and RI were components of five essential AS forms in our study, and this distribution pattern is also similar to that of other animals reported previously (15, 35, 42, 43), indicated that animals might possess similar alternative splicing forms. Alternative splicing events were numerous occur during organ development, tissue maturation, and cell differentiation, suggesting that alternative splicing supports proper development (15). The phenotype may be influenced by modification of gene transcription or translation induced

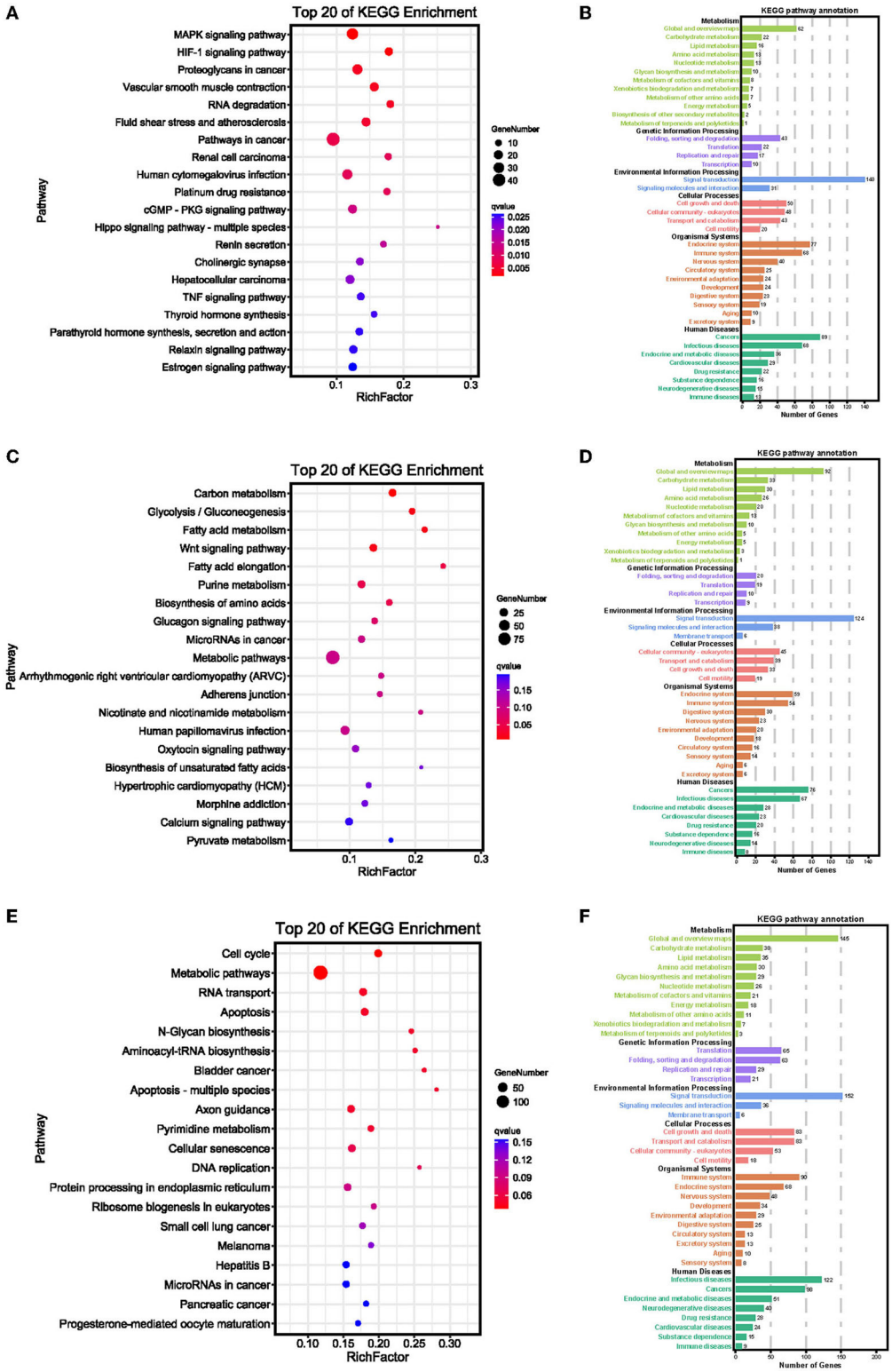


FIGURE 5
KEGG enrichment pathways of alternatively spliced DEGs **(A,B)** between normoxic (TN and LN) and hypoxic (TL and LL) groups, **(C,D)** between TN and TL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups), and **(E,F)** between LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) groups. The ordinate is the pathway, and the abscissa is the enrichment factor. Darker colors indicate smaller q-values.

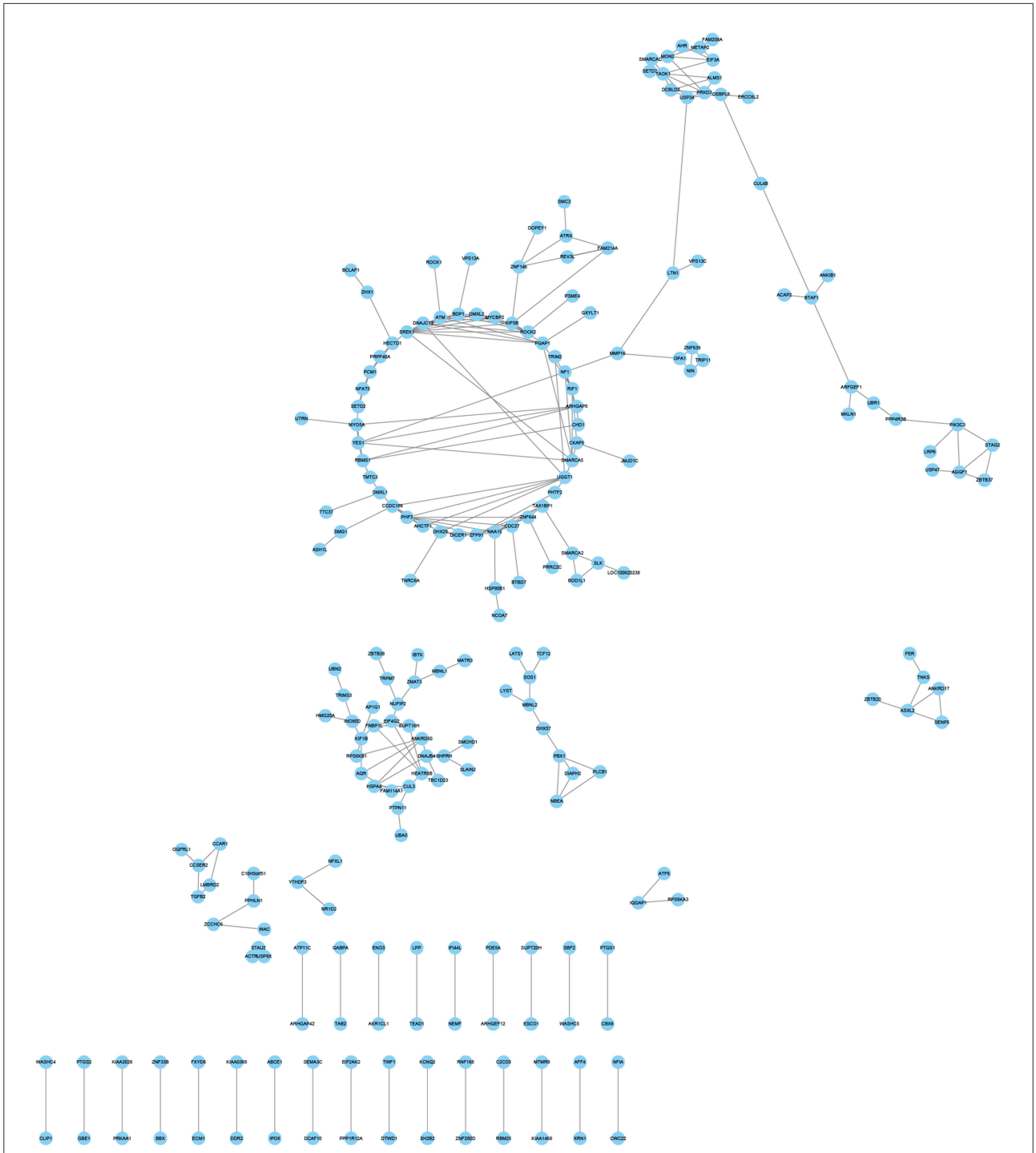


FIGURE 6
Coexpression network analyses of alternatively spliced DEGs between normoxic (TN and LN) and hypoxic (TL and LL) groups.

by a hypoxia condition (44). SE may be the primary source of proteomic and transcriptomic and plays a significant role in hypoxia response by regulating genes and determining phenotype as the most abundant event types (73.01%) in ATII cells among four groups (45, 46).

Regulation of AS in ATII cells response to hypoxia

According to the KEGG enrichment, a total of 1,088 alternatively spliced DEmRNAs were enriched in 279 pathways

between normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups, of which 35, 17, and 25 genes were enriched in MAPK signaling pathway, HIF-1 signaling pathway, and proteoglycans in cancer. MAPK may arise and upregulate the transcription of anti-apoptotic genes under exposure to hypoxia and play critical roles in opposing the inflammatory response and regulating cell proliferation, differentiation, and apoptosis, which may be a novel strategy for the treatment of chronic obstructive pulmonary fibrosis (6, 47–50). Insulin-like growth factors (*IGF1* and *IGF2*) enriched in 30 pathways and underwent one type of AS event between normoxic and hypoxic groups and might act as cross-talk between MAPK pathways and HIF signaling pathway (51), which may reduce ATII cell apoptosis under hypoxic conditions (52). The increase of HIF-1 transcriptional activity under a hypoxia environment is due to a decrease of cellular NAD+, which downregulates Sirt1 to enhance HIF-1α acetylation (53). As expected, we also discovered that several glycolysis-related genes (such as *SLC2A1*, *HK1*, *HK2*, *ENO3*, and *PFKFB3*) undergo one or two AS event types. The frequency of *HK2* was higher in normoxia than that of hypoxia groups, and the frequency of SE events in *ENO3* was lower in LN groups than any others. The frequency of SE events in *HK1* was lowest in LL groups, enriched in the HIF-1 signaling pathway, and may promote anaerobic metabolism by elevating interstitial pressure and alleviating cell damage through glucose metabolism under hypoxia conditions (54, 55). The energy and metabolic intermediates produced through cells rely on glycolysis by hypoxia availability. *HK1* and *HK2*, responsible for the initial steps of glycolysis, convert glucose to glucose-6-phosphate (G-6-P) through phosphorylation, initiating glycolysis and producing pyruvate and lactic acid as energy sources (56, 57). PFKFB enzymes catalyze the synthesis of fructose-2,6-bisphosphate (F-2,6-P2) as one of the numerous glycolytic regulators. *PFKFB3* plays a dominant role in vascular cells, leukocytes, and many transformed cells and catalyzes the conversion of fructose-6-phosphate to fructose-1,6-biphosphate as the number of the four isoforms of PFKFBs (58, 59). *PFKFB3* undergoes SE events and has a lower frequency in LN groups than in any other groups, may control the steady-state concentration of F-2,6-P2, and glycolysis also mediated the generation of growth factors and proinflammatory cytokines in ATII cells under hypoxia condition (59, 60). Thus, glycolysis intermediates can be increased in the intracellular levels of glucose and quickly diverted to anabolic pathways under hypoxia as substrates for lipid and protein biosynthesis and DNA replication to rapidly proliferate ATII cells (61, 62).

ROCK2, *KIF5B*, and Zinc finger protein 91 (*ZFP91*) regulated several mRNAs under hypoxia conditions stimulation as essential hypoxia-inducible genes between normoxia and hypoxia groups. Under hypoxia conditions, pulmonary arterial endothelial cells' proliferation and cell cycle *via* activation of the ROCK2 signaling pathway (63). Cell migration of macrophages and bladder cancer cells may inhibit *ROCK2* expression (64, 65). *ZFP91* could upregulate the expression of *HIF-1α via* binds to its promoter region and is involved in various biological processes (66, 67). In summary, the present study shows that the A3SS and SE AS events of *ZFP91* and higher frequency of SE events in *ROCK2* and under normoxia (LN and TN) groups may influence proliferation, apoptosis, and epithelial–mesenchymal transition of ATII cells (63–67).

Functional effects of alternatively spliced DEGs of Tibetan pigs and landrace pigs at hypoxia conditions

Although the alternatively spliced DEGs in the same oxygen concentration of Tibetan pigs and Landrace pigs should have similar alternative splicing, 18, 12, and 10 alternatively spliced DEGs were most significantly enriched in carbon metabolism, glycolysis/gluconeogenesis, and fatty acid metabolism among the TN and TL groups (excluding alternatively spliced DEGs shared between normoxic and hypoxic groups). *LDHA* undergoes SE and MXE events and is significantly enriched in the glycolysis/gluconeogenesis pathway of ATII cells in Tibetan pigs under normoxia and hypoxia, a net charge of −6, and preferentially converts pyruvate to lactate, and occupies plasma membrane and mitochondrial with *LDHB* isoforms (68). Previous research reveals that *CD36* and intracellular lipid expression and content were augmented in hypoxic hepatocytes. The membrane-bound sterol regulatory element-binding protein (*SREBP*) transcription factors could respond to lipid availability and regulate lipid uptake and synthesis as central regulators of lipid homeostasis (69). Fatty acids were identified as a physiological modulator of HIF and have similar functions to oxygen, defining a mechanism for lipoprotein regulation (70). Several alternatively spliced DEGs, such as *ACADL*, *EHHADH*, and *CPT1A*, undergo one or two AS types of ATII cells, and different types and frequencies of AS event may be a constant supply of lipids were obtained either from the circulation or *de novo* synthesis for ATII cells growth better under hypoxia condition (71, 72).

In contrast to the results obtained from the comparison of ATII cells of Landrace pigs under normoxic and hypoxic conditions, the pathways related to the alternatively spliced DEGs identified from the comparison between LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) were associated with cell cycle, metabolic pathways, RNA transport, and apoptosis. Available evidence suggests hypoxia compensates for cell cycle arrest with decreased S-phase entry in mature ECs and progenitor differentiation during angiogenesis (73). The *p53* is a cell cycle regulator and

apoptosis in the white shrimp in response to hypoxia (74), and the miR-493-STMN-1 pathway could promote hypoxia-induced epithelial cell cycle arrest in G_2/M phase (75). *CDK2* (cyclin-dependent kinase 2) undergo AS event between LN and LL groups, which could be activated by either *CCNE* (cyclin E) or *CCNA* (cyclin A) at the G1/S phase transition or S phase, and mediates degradation of *HIF-1α* at the G1/S change (76). *MCM7* and *MCM3* undergo AS events between LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) groups, bind to the amino-terminal PER-SIM-ARNT (PAS) domain, and the carboxyl terminus of *HIF-1α* to maintain their stability (77).

Conclusion

In this study, we disclosed features of AS events in ATII cells through RNA-seq data. The results indicated that different types of AS and regulatory networks might partially contribute to the significant variance in ATII cells of Tibetan pigs and Landrace pigs under different oxygen concentrations. *ACADL*, *EHHADH*, and *CPT1A* may be a constant supply of lipids were obtained either from the circulation or *de novo* synthesis for ATII cells of Tibetan pigs growth better under hypoxia conditions. Therefore, this study provided a better understanding of the effects of different AS of candidate functional genes on ATII cells' response to hypoxia.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

The animal study was reviewed and approved by Livestock Care Committee of Gansu Agricultural University.

Author contributions

SZ and YY were the overall project leader who provided financial support and experimental conception. HY was involved in data analyses, statistical analyses, language revisions, journal selection, and manuscript submissions and revisions. XL and ZW contributed to the experimental design and implementation. CG contributed to the supervision and assistance of students in managing animals and collecting and analyzing samples. YL and YR were responsible for the trial implementation, supervision of students collecting and analyzing samples, and manuscript preparation. YC and TJ contributed to supervision of sample collection and analysis and manuscript editing. All authors contributed to the article and approved the submitted version.

Funding

The study was supported by the National Natural Science Foundation of China (32060730).

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/fvets.2022.984703/full#supplementary-material

SUPPLEMENTARY MATERIAL 1

SUPPLEMENTARY TABLE S1
Primers used to detect alternatively spliced differentially expressed genens (DEGs) in ATII cells of pigs by qRT-PCR.

SUPPLEMENTARY TABLE S2
Overview of the reads and quality filtering of mRNA libraries.

SUPPLEMENTARY TABLE S3
The number of genes underwent alternative splicing (AS) events.

SUPPLEMENTARY TABLE S4
AS events of Tibetan pigs and Landrace pigs were present in DEGs between normoxic and hypoxic groups.

SUPPLEMENTARY MATERIAL 2

2.1.
The A3SS events corresponding to genes were identified.

2.2.
The A5SS events corresponding to genes were identified.

2.3.
The MXE events corresponding to genes were identified.

2.4.
The RI events corresponding to genes were identified.

2.5.
The SE events corresponding to genes were identified.

SUPPLEMENTARY MATERIAL 3

3.1.
The frequency of AS events in the TN group.

3.2.
The frequency of AS events in the TL group.

3.3.
The frequency of AS events in the LN group.

3.4.
The frequency of AS events in the LL group.

SUPPLEMENTARY MATERIAL 4

4.1.
Differentially expressed genes between normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups.

4.2.
Alternatively spliced DEGs between normoxic (21% O_2, TN and LN) and hypoxic (2% O_2, TL and LL) groups.

SUPPLEMENTARY MATERIAL 5

SUPPLEMENTARY FIGURE S1
Coexpression network analyses of alternatively spliced differentially expressed genes (DEGs) between TN and TL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) groups.

SUPPLEMENTARY FIGURE S2
Coexpression network analyses of alternatively spliced DEGs between LN and LL (excluding alternatively spliced DEGs shared between normoxia and hypoxia groups) groups.

SUPPLEMENTARY FIGURE S3
(A) Expression patterns of three alternatively spliced DEGs. **(B)** Expression patterns of eight randomly selected DEGs. Histogram represents the change in transcript level according to the FPKM value of RNA-seq (left y-axis), and broken line indicates that relative expression level defense by RT-PCR (right y-axis).

References

1. Ma YF, Han XM, Huang CP, Zhong L, Adeola AC, Irwin DM, et al. Population genomics analysis revealed origin and high-altitude adaptation of tibetan pigs. *Sci Rep.* (2019) 9:11463. doi: 10.1038/s41598-019-47711-6

2. Yang Y, Gao C, Yang T, Sha Y, Cai Y, Wang X, et al. Characteristics of Tibetan pig lung tissue in response to a hypoxic environment on the Qinghai-Tibet Plateau. *Arch Anim Breed.* (2021) 64:283–92. doi: 10.5194/aab-64-283-2021

3. Yang Y, Yuan H, Yang T, Li Y, Gao C, Jiao T, et al. The expression regulatory network in the lung tissue of tibetan pigs provides insight into hypoxia-sensitive pathways in high-altitude hypoxia. *Front Genet.* (2021) 12:691592. doi: 10.3389/fgene.2021.691592

4. Sydykov A, Mamazhakypov A, Maripov A, Kosanovic D, Weissmann N, Ghofrani HA, et al. Pulmonary Hypertension In Acute And Chronic High Altitude Maladaptation Disorders. *Int J Environ Res Public Health.* (2021) 18:1692. doi: 10.3390/ijerph18041692

5. Nathan SD, Barbera JA, Gaine SP, Harari S, Martinez FJ, Olschewski H, et al. Pulmonary hypertension in chronic lung disease and hypoxia. *Eur Respir J.* (2019) 53:1801914. doi: 10.1183/13993003.01914-2018

6. Rabe KF, Watz H. Chronic obstructive pulmonary disease. *Lancet.* (2017) 389:1931–40. doi: 10.1016/S0140-6736(17)31222-9

7. Labaki WW, Rosenberg SR. Chronic obstructive pulmonary disease. *Ann Intern Med.* (2020) 173:ITC17–32. doi: 10.7326/AITC202008040

8. Tanguy J, Goirand F, Bouchard A, Frenay J, Moreau M, Mothes C, et al. [18F]FMISO PET/CT imaging of hypoxia as a non-invasive biomarker of disease progression and therapy efficacy in a preclinical model of pulmonary fibrosis: comparison with the [18F]FDG PET/CT approach. *Eur J Nucl Med Mol Imaging.* (2021) 48:3058–74. doi: 10.1007/s00259-021-05209-2

9. Guillot L, Nathan N, Tabary O, Thouvenin G, Le Rouzic P, Corvol H, et al. Alveolar epithelial cells: master regulators of lung homeostasis. *Int J Biochem Cell Biol.* (2013) 45:2568–73. doi: 10.1016/j.biocel.2013.08.009

10. Aspal M, Zemans RL. Mechanisms of ATII-to-ATI cell differentiation during lung regeneration. *Int J Mol Sci.* (2020) 21:3188. doi: 10.3390/ijms21093188

11. Alvarez-Palomo B, Sanchez-Lopez LI, Moodley Y, Edel MJ, Serrano-Mollar A. Induced pluripotent stem cell-derived lung alveolar epithelial type II cells reduce damage in bleomycin-induced lung fibrosis. *Stem Cell Res Ther.* (2020) 11:213. doi: 10.1186/s13287-020-01726-3

12. Delbrel E, Uzunhan Y, Soumare A, Gille T, Marchant D, Planès C, et al. Stress is involved in epithelial-to-mesenchymal transition of alveolar epithelial cells exposed to a hypoxic microenvironment. *Int J Mol Sci.* (2019) 20:1299. doi: 10.3390/ijms20061299

13. Sherman MA, Suresh MV, Dolgachev VA, McCandless LK, Xue X, Ziru L, et al. Molecular characterization of hypoxic alveolar epithelial cells after lung contusion indicates an important role for HIF-1α. *Ann Surg.* (2018) 267:382–91. doi: 10.1097/SLA.0000000000002070

14. Grek CL, Newton DA, Spyropoulos DD, Baatz JE. Hypoxia up-regulates expression of hemoglobin in alveolar epithelial cells. *Am J Respir Cell Mol Biol.* (2011) 44:439–47. doi: 10.1165/rcmb.2009-0307OC

15. Ule J, Blencowe BJ. alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol Cell.* (2019) 76:329–45. doi: 10.1016/j.molcel.2019.09.017

16. Han J, Li J, Ho JC, Chia GS, Kato H, Jha S, et al. Hypoxia is a key driver of alternative splicing in human breast cancer cells. *Sci Rep.* (2017) 7:4108. doi: 10.1038/s41598-017-04333-0

17. Liu Z, Sun L, Cai Y, Shen S, Zhang T, Wang N, et al. Hypoxia-induced suppression of alternative splicing of MBD2 promotes breast cancer metastasis via activation of FZD1. *Cancer Res.* (2021) 81:1265–78. doi: 10.1158/0008-5472.CAN-20-2876

18. Farina AR, Cappabianca L, Sebastiano M, Zelli V, Guadagni S, Mackay AR, et al. Hypoxia-induced alternative splicing: the 11th Hallmark of Cancer. *J Exp Clin Cancer Res.* (2020) 39:110. doi: 10.1186/s13046-020-01616-9

19. Brumwell A, Fell L, Obress L, Uniacke J. Hypoxia influences polysome distribution of human ribosomal protein S12 and alternative splicing of ribosomal protein mRNAs. *RNA.* (2020) 26:361–71. doi: 10.1261/rna.070318.119

20. Heikkilä M, Pasanen A, Kivirikko KI, Myllyharju J. Roles of the human hypoxia-inducible factor (HIF)-3α variants in the hypoxia response. *Cell Mol Life Sci.* (2011) 68:3885–901. doi: 10.1007/s00018-011-0679-5

21. Wu DD, Yang CP, Wang MS, Dong KZ, Yan DW, Hao ZQ, et al. Convergent genomic signatures of high-altitude adaptation among domestic mammals. *Natl Sci Rev.* (2020) 7:952–63. doi: 10.1093/nsr/nwz213

22. Yan JQ, Liu M, Ma YL, Le KD, Dong B, Development LiGH, et al. of alternative splicing signature in lung squamous cell carcinoma. *Med Oncol.* (2021) 38:49. doi: 10.1007/s12032-021-01490-1

23. Xu Z, Wei J, Qin F, Sun Y, Xiang W, Yuan L, et al. Hypoxia-associated alternative splicing signature in lung adenocarcinoma. *Epigenomics.* (2021) 13:47–63. doi: 10.2217/epi-2020-0399

24. Weeland CJ, van den Hoogenhof MM, Beqqali A, Creemers EE. Insights into alternative splicing of sarcomeric genes in the heart. *J Mol Cell Cardiol.* (2015) 81:107–13. doi: 10.1016/j.yjmcc.2015.02.008

25. Tang LT, Ran XQ, Mao N, Zhang FP, Niu X, Ruan YQ, et al. Analysis of alternative splicing events by RNA sequencing in the ovaries of Xiang pig at estrous and diestrous. *Theriogenology.* (2018) 119:60–8. doi: 10.1016/j.theriogenology.2018.06.022

26. Wang X, Zhang L, Sun B. Neonatal type II alveolar epithelial cell transplant facilitates lung reparation in piglets with acute lung injury and extracorporeal life support. *Pediatr Crit Care Med.* (2016) 17:e182–92. doi: 10.1097/PCC.0000000000000667

27. Singh S, Gupta M, Pandher S, Kaur G, Goel N, Rathore P. Using *de novo* transcriptome assembly and analysis to study RNAi in *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae). *Sci Rep.* (2019) 9:13710. doi: 10.1038/s41598-019-49997-y

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

29. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data, Babraham Bioinformatic.* Cambridge, United Kingdom: Babraham Institute (2010).

30. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* (2010) 38:e131. doi: 10.1093/nar/gkq224

31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics30.* (2014) 2114–20. doi: 10.1093/bioinformatics/btu170

32. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* (2013) 14:R36. doi: 10.1186/gb-2013-14-4-r36

33. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks [published correction appears in Nat Protoc. 2014 Oct;9:2513]. *Nat Protoc.* (2012) 7:562–78. doi: 10.1038/nprot.2012.016

34. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic* Acids Res. (2015) 43:D447–52. doi: 10.1093/nar/gku1003

35. Bhadra M, Howell P, Dutta S, Heintz C, Mair WB. Alternative splicing in aging and longevity. *Hum Genet.* (2020) 139:357–69. doi: 10.1007/s00439-019-02094-6

36. Sciarrillo R, Wojtuszkiewicz A, Assaraf YG, Jansen G, Kaspers GJL, Giovannetti E, et al. The role of alternative splicing in cancer: from oncogenesis to drug resistance. *Drug Resist Updat.* (2020) 53:100728. doi: 10.1016/j.drup.2020.100728

37. Iwata H, Gotoh O. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics.* (2011) 12:45. doi: 10.1186/1471-2164-12-45

38. Song H, Wang L, Chen D, Li F. The function of pre-mRNA alternative splicing in mammal spermatogenesis. *Int J Biol Sci.* (2020) 16:38–48. doi: 10.7150/ijbs.34422

39. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* (2012) 22:1184–95. doi: 10.1101/gr.134106.111

40. Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet.* (2003) 34:177–80. doi: 10.1038/ng1159

41. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* (2008) 456:470–6. doi: 10.1038/nature07509

42. Wang X, Du L, Wei H, Zhang A, Yang K, Zhou H, et al. Identification of two Stat3 variants lacking a transactivation domain in grass carp: new insights into alternative splicing in the modification of teleost Stat3 signaling. *Fish Shellfish Immunol.* (2018) 77:13–21. doi: 10.1016/j.fsi.2018.03.022

43. Fang X, Xia L, Yu H, He W, Bai Z, Qin L, et al. Comparative genome-wide alternative splicing analysis of longissimus dorsi muscles between Japanese Black (Wagyu) and Chinese Red Steppes Cattle. *Front Vet Sci.* (2021) 8:634577. doi: 10.3389/fvets.2021.634577

44. Singh RS. Darwin's legacy II: why biology is not physics, or why it has taken a century to see the dependence of genes on the environment. *Genome.* (2015) 58:55–62. doi: 10.1139/gen-2015-0012

45. Cossins AR, Crawford DL. Fish as models for environmental genomics. *Nat Rev Genet.* (2005) 6:324–33. doi: 10.1038/nrg1590

46. Xia JH, Li HL, Li BJ, Gu XH, Lin HR. Acute hypoxia stress induced abundant differential expression genes and alternative splicing events in heart of tilapia. *Gene.* (2018) 639:52–61. doi: 10.1016/j.gene.2017.10.002

47. Singh M, Yadav S, Kumar M, Saxena S, Saraswat D, Bansal A, et al. The MAPK-activator protein-1 signaling regulates changes in lung tissue of rat exposed to hypobaric hypoxia. *J Cell Physiol.* (2018) 233:6851–65. doi: 10.1002/jcp.26556

48. Shologu N, Scully M, Laffey JG, O'Toole D. Human mesenchymal stem cell secretome from bone marrow or adipose-derived tissue sources for treatment of hypoxia-induced pulmonary epithelial injury. *Int J Mol Sci.* (2018) 19:2996. doi: 10.3390/ijms19102996

49. Estaras M, Gonzalez-Portillo MR, Fernandez-Bermejo M, Mateos JM, Vara D, Blanco-Fernandez G, et al. Melatonin induces apoptosis and modulates cyclin expression and MAPK Phosphorylation in pancreatic stellate cells subjected to hypoxia. *Int J Mol Sci.* (2021) 22:5555. doi: 10.3390/ijms22115555

50. Dong Q, Jie Y, Ma J, Li C, Xin T, Yang D, et al. Renal tubular cell death and inflammation response are regulated by the MAPK-ERK-CREB signaling pathway under hypoxia-reoxygenation injury. *J Recept Signal Transduct Res.* (2019) 39:383–91. doi: 10.1080/10799893.2019.1698050

51. Zeng Y, Zhang L, Hu Z. Cerebral insulin, insulin signaling pathway, and brain angiogenesis. *Neurol Sci.* (2016) 37:9–16. doi: 10.1007/s10072-015-2386-8

52. Fan J, Shi S, Qiu Y, Zheng Z, Yu L. MicroRNA-486-5p down-regulation protects cardiomyocytes against hypoxia-induced cell injury by targeting IGF-1. *Int J Clin Exp Pathol.* (2019) 12:2544–51.

53. Majmundar AJ, Wong WJ, Simon MC. Hypoxia-inducible factors and the response to hypoxic stress. *Mol Cell.* (2010) 40:294–309. doi: 10.1016/j.molcel.2010.09.022

54. Nagao A, Kobayashi M, Koyasu S, Chow CCT, Harada H. HIF-1-dependent reprogramming of glucose metabolic pathway of cancer cells and its therapeutic significance. *Int J Mol Sci.* (2019) 20:238. doi: 10.3390/ijms20020238

55. Li X, Wang M, Li S, Chen Y, Wang M, Wu Z, et al. HIF-1-induced mitochondrial ribosome protein L52: a mechanism for breast cancer cellular adaptation and metastatic initiation in response to hypoxia. *Theranostics.* (2021) 11:7337–59. doi: 10.7150/thno.57804

56. Xu S, Catapang A, Doh HM, Bayley NA, Lee JT, Braas D, et al. Hexokinase 2 is targetable for HK1 negative, HK2 positive tumors from a wide variety of tissues of origin. *J Nucl Med.* (2018) 60:212–7. doi: 10.2967/jnumed.118.212365

57. Xu S, Zhou T, Doh HM, Trinh KR, Catapang A, Lee JT, et al. An HK2 antisense oligonucleotide induces synthetic lethality in HK1-HK2+ multiple myeloma. *Cancer Res.* (2019) 79:2748–60. doi: 10.1158/0008-5472.CAN-18-2799

58. Okar DA, Wu C, Lange AJ. Regulation of the regulatory enzyme, 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Adv Enzyme Regul.* (2004) 44:123–54. doi: 10.1016/j.advenzreg.2003.11.006

59. Chesney J. 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase and tumor cell glycolysis. *Curr Opin Clin Nutr Metab Care.* (2006) 9:535–9. doi: 10.1097/01.mco.0000241661.15514.fb

60. Cao Y, Zhang X, Wang L, Yang Q, Ma Q, Xu J, et al. PFKFB3-mediated endothelial glycolysis promotes pulmonary hypertension. *Proc Natl Acad Sci USA.* (2019) 116:13394–403. doi: 10.1073/pnas.1821401116

61. Lottes RG, Newton DA, Spyropoulos DD, Baatz JE. Alveolar type II cells maintain bioenergetic homeostasis in hypoxia through metabolic and molecular adaptation. *Am J Physiol Lung Cell Mol Physiol.* (2014) 306:L947–55. doi: 10.1152/ajplung.00298.2013

62. Arthur SA, Blaydes JP, Houghton FD. Glycolysis regulates human embryonic stem cell self-renewal under hypoxia through HIF-2α and the glycolytic sensors CTBPs. *Stem Cell Rep.* (2019) 12:728–42. doi: 10.1016/j.stemcr.2019.02.005

63. Qiao F, Zou Z, Liu C, Zhu X, Wang X, Yang C, et al. ROCK2 mediates the proliferation of pulmonary arterial endothelial cells induced by hypoxia in the development of pulmonary arterial hypertension. *Exp Ther Med.* (2016) 11:2567–72. doi: 10.3892/etm.2016.3214

64. Luo J, Lou Z, Zheng J. Targeted regulation by ROCK2 on bladder carcinoma via Wnt signaling under hypoxia. *Cancer Biomark.* (2019) 24:109–16. doi: 10.3233/CBM-181949

65. Chen H, Du J, Zhang S, Tong H, Zhang M. Ghrelin suppresses migration of macrophages via inhibition of ROCK2 under chronic intermittent hypoxia. *J Int Med Res.* (2020) 48:300060520926065. doi: 10.1177/0300060520926065

66. Ma J, Mi C, Wang KS, Lee JJ, Jin X. Zinc finger protein 91 (ZFP91) activates HIF-1α via NF-κB/p65 to promote proliferation and tumorigenesis of colon cancer. *Oncotarget.* (2016) 7:36551–62. doi: 10.18632/oncotarget.9070

67. Hanafi M, Chen X, Neamati N. Discovery of a napabucasin PROTAC as an effective degrader of the E3 ligase ZFP91. *J Med Chem.* (2021) 64:1626–48. doi: 10.1021/acs.jmedchem.0c01897

68. Hussien R, Brooks GA. Mitochondrial and plasma membrane lactate transporter and lactate dehydrogenase isoform expression

in breast cancer cell lines. *Physiol Genomics.* (2011) 43:255–64. doi: 10.1152/physiolgenomics.00177.2010

69. Ye J, DeBose-Boyd RA. Regulation of cholesterol and fatty acid synthesis. *Cold Spring Harb Perspect Biol.* (2011) 3:a004754. doi: 10.1101/cshperspect.a004754

70. Shao W, Hwang J, Liu C, Mukhopadhyay D, Zhao S, Shen MC, et al. Serum lipoprotein-derived fatty acids regulate hypoxia-inducible factor. *J Biol Chem.* (2020) 295:18284–300. doi: 10.1074/jbc.RA120.015238

71. Menendez JA, Lupu R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer.* (2007) 7:763–77. doi: 10.1038/nrc2222

72. Baenke F, Peck B, Miess H, Schulze A. Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. *Dis Model Mech.* (2013) 6:1353–63. doi: 10.1242/dmm.011338

73. Acosta-Iborra B, Tiana M, Maeso-Alonso L, Hernández-Sierra R, Herranz G, Santamaria A, et al. Hypoxia compensates cell cycle arrest with progenitor differentiation during angiogenesis. *FASEB J.* (2020) 34:6654–74. doi: 10.1096/fj.201903082R

74. Nuñez-Hernandez DM, Felix-Portillo M, Peregrino-Uriarte AB, Yepiz-Plascencia G. Cell cycle regulation and apoptosis mediated by p53 in response to hypoxia in hepatopancreas of the white shrimp Litopenaeus vannamei. *Chemosphere.* (2018) 190:253–9. doi: 10.1016/j.chemosphere.2017.09.131

75. Liu T, Liu L, Liu M, Du R, Dang Y, Bai M, et al. MicroRNA-493 targets STMN-1 and promotes hypoxia-induced epithelial cell cycle arrest in G2/M and renal fibrosis. *FASEB J.* (2019) 33:1565–77. doi: 10.1096/fj.201701355RR

76. Hubbi ME, Semenza GL. An essential role for chaperone-mediated autophagy in cell cycle progression. *Autophagy.* (2015) 11:850–1. doi: 10.1080/15548627.2015.1037063

77. Semenza GL. Hypoxia. Cross talk between oxygen sensing and the cell cycle machinery. *Am J Physiol Cell Physiol.* (2011) 301:C550–2. doi: 10.1152/ajpcell.00176.2011

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 September 2022
DOI 10.3389/fgene.2022.974787

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Guangxin E,
Southwest University, China
Chuzhao Lei,
Northwest A&F University, China

*CORRESPONDENCE
Xiangdong Ding,
xiangdongding@hotmail.com
Gábor Mészáros,
gabor.meszaros@boku.ac.at

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

RECEIVED 21 June 2022
ACCEPTED 29 August 2022
PUBLISHED 27 September 2022

CITATION
Naji MM, Jiang Y, Utsunomiya YT, Rosen BD, Sölkner J, Wang C, Jiang L, Zhang Q, Zhang Y, Ding X and Mészáros G (2022), Favored single nucleotide variants identified using whole genome Re-sequencing of Austrian and Chinese cattle breeds. *Front. Genet.* 13:974787. doi: 10.3389/fgene.2022.974787

Favored single nucleotide variants identified using whole genome Re-sequencing of Austrian and Chinese cattle breeds

Maulana M. Naji[1], Yifan Jiang[2], Yuri T. Utsunomiya[3], Benjamin D. Rosen[4], Johann Sölkner[1], Chuduan Wang[2], Li Jiang[2], Qin Zhang[2], Yi Zhang[2], Xiangdong Ding[2]* and Gábor Mészáros[1]*

[1]University of Natural Resources and Life Sciences, Vienna, Austria, [2]China Agricultural University, Beijing, China, [3]Department of Production and Animal Health, School of Veterinary Medicine, São Paulo State University (Unesp), Araçatuba, Brazil, [4]Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, United States

Cattle have been essential for the development of human civilization since their first domestication few thousand years ago. Since then, they have spread across vast geographic areas following human activities. Throughout generations, the cattle genome has been shaped with detectable signals induced by various evolutionary processes, such as natural and human selection processes and demographic events. Identifying such signals, called selection signatures, is one of the primary goals of population genetics. Previous studies used various selection signature methods and normalized the outputs score using specific windows, in kbp or based on the number of SNPs, to identify the candidate regions. The recent method of iSAFE claimed for high accuracy in pinpointing the candidate SNPs. In this study, we analyzed whole-genome resequencing (WGS) data of ten individuals from Austrian Fleckvieh (*Bos taurus*) and fifty individuals from 14 Chinese indigenous breeds (*Bos taurus, Bos taurus indicus,* and admixed). Individual WGS reads were aligned to the cattle reference genome of ARS. UCD1.2 and subsequently undergone single nucleotide variants (SNVs) calling pipeline using GATK. Using these SNVs, we examined the population structure using principal component and admixture analysis. Then we refined selection signature candidates using the iSAFE program and compared it with the classical iHS approach. Additionally, we run Fst population differentiation from these two cattle groups. We found gradual changes of taurine in north China to admixed and indicine to the south. Based on the population structure and the number of individuals, we grouped samples to Fleckvieh, three Chinese taurines (Kazakh, Mongolian, Yanbian), admixed individuals (CHBI_Med), indicine individuals (CHBI_Low), and a combination of admixed and indicine (CHBI) for performing iSAFE and iHS tests. There were more significant SNVs identified using iSAFE than the iHS for the candidate of positive selection and more detectable signals in taurine than in indicine individuals. However, combining admixed and indicine individuals decreased

the iSAFE signals. From both within-population tests, significant SNVs are linked to the olfactory receptors, production, reproduction, and temperament traits in taurine cattle, while heat and parasites tolerance in the admixed individuals. Fst test suggests similar patterns of population differentiation between Fleckvieh and three Chinese taurine breeds against CHBI. Nevertheless, there are genes shared only among the Chinese taurine, such as PAX5, affecting coat color, which might drive the differences between these yellowish coated breeds, and those in the greater Far East region.

KEYWORDS

cattle, whole-genome sequence (WGS), selection signature, *Bos taurus*, bos indicus, iSAFE, IHS, fst

Introduction

Cattle are vital livestock for humans providing meat and milk for consumption, leather for protection, and power for plowing and transportation (FAO, 2015; Xu et al., 2015). Using available genetic evidence, there were two primary independent events postulated for the initial domestication of cattle, i.e., between 10,000 and 8,000 years ago for *Bos taurus* (*B. taurus*) in the Fertile Crescent and 8,000–6,000 years ago for *Bos taurus indicus* (*B. indicus*) in the Indus valley (Loftus et al., 1994). Since then, following human migration and trade, cattle have spread across the globe and undergone further evolutionary events for adaptation to local environments due to natural or selective breeding shaping each breed's morphology, physiology, and behavior from its initial attributes (FAO, 2015; Xu et al., 2015; Wu et al., 2018). Currently, there are more than a thousand distinctive cattle breeds recognized worldwide (FAO, 2015).

The study of footprints in genes or genomics regions of livestock species due to the continuous evolutionary process is one of the main interests of population genetics (de Simoni Gouveia et al., 2014; Randhawa et al., 2016). With the development of genomics, these signals can be identified using single nucleotide polymorphisms (SNPs) arrays and whole-genome resequencing (WGS) data (Flori et al., 2009; Utsunomiya et al., 2013; Qanbari et al., 2014; Randhawa et al., 2014). These signals are inferred as they deviate from the neutral expectations in the patterns of genomic variations despite possible recombination events (Utsunomiya et al., 2013; de Simoni Gouveia et al., 2014). There are various proposed methods to detect these signals. Based on its approaches, they can be grouped into methods using local genetic diversity depression within a population, changes in allele frequency spectrum within and cross-populations, population allele differentiation across-populations, and haplotype homozygosity within and cross-populations (Utsunomiya et al., 2013; Randhawa et al., 2016).

Estimated for the first importation of *B. taurus* from West Asia around 3,900 years ago, there are ~90 million cattle of various breeds in China, of which fifty-three of it are indigenous (Chen et al., 2018; National Bureau of Statistics, 2018). A previous study reported gradual transitions in cattle breed composition found across the country. *B. taurus is* predominantly found in the northern part, gradually admixed of *B. taurus* and *B. indicus* population in the central part, and pure *B. indicus* breeds to the southern part of the country (Chen et al., 2018). Another study (Zhang et al., 2020) using copy number variations (CNVs) supported that most Chinese breeds were hybrids of *B. taurus* and *B. indicus*.

Fleckvieh is a prominent dual-purpose breed in Austria with a population of around 1.5 million heads, corresponding to 76% of the total cattle population in the country (Kalcher et al., 2018). Also internationally known as Simmental, Fleckvieh genome was reported as one of the most studied *B. taurus* cattle after Frisian-Holstein (Randhawa et al., 2016). A previous study (Qanbari et al., 2014) utilized sequencing data of German Fleckvieh for selection signatures analysis. They employed ~15 million autosomal SNPs inferred from the sequence data and found 106 candidates of selection regions linked to genes with the functionality of neuro-behavioral, sensory perception, and coat coloring patterns.

Most of the previous studies were limited in pinpointing exact locations of selection signatures, as they proposed large chunks of genomic regions in the size of a few kilobases to megabases as the candidates, containing many genes and thousands of polymorphisms (Randhawa et al., 2014; Xu et al., 2015; Bhati et al., 2020). They considered the region within linkage disequilibrium proximate as the candidate regions and reducing spurious effects of many SNPs signals as the reasons for using large scanning windows.

Recently developed methods of integrated Selection of Allele Favoured by Evolution - iSAFE are suggested to pinpoint the best candidate SNPs in selection signature regions (Akbari et al., 2018). iSAFE is designed to exploit signals from ongoing selective sweeps as it scores are based on the rank-order of the mutation in SNP candidates. Using phased genotype, this tool assigns intermediate score for each mutation based on the number of times it appears in different haplotypes weighted by total of all mutations found in the haplotype and its frequency. Then, overlapped scanning window is applied on these intermediate-scores of all mutations to find the best candidate SNP driving the

selection, see Methods for details. iSAFE outperformed other tools, such as integrated Haplotype Score—iHS (Voight et al., 2006), in detecting favorable SNPs within large loci of 5 Mb without knowledge of demography, phenotype under selection, or functional annotations (Akbari et al., 2018). Thus, in this study, we aim to examine the candidate SNPs that drive the selection identified by iSAFE in genome-wide level, with no prior knowledge of candidate regions in selection, compared to the classical approach of integrated Haplotype Score—iHS (Voight et al., 2006; Szpiech and Hernandez, 2014) using sequence data of several Chinese breeds and Austrian Fleckvieh.

Materials and methods

Ethics statement

For this study, DNA was previously extracted from commercial AI bull semen straws. Thus, no ethical statement was further required.

Alignment, variant calling, and phasing genotypes

In this study, we utilized whole genome re-sequencing of sixty individuals from fourteen Chinese and one Austrian cattle breeds, namely: Dabieshan (n = 2), Dehong (n = 2), Dengchuan (n = 2), Fujian (n = 2), Guanling (n = 2), Kazakh (n = 6), Liping (n = 2), Luxi (n = 2), Mongolian (n = 12), Nanyang (n = 2), Qinchuan (n = 2), Wenling (n = 2), Tibetan (n = 2), Yanbian (n = 10), and Fleckvieh (n = 10). In the analysis, we applied the alignment to SNV calling pipeline in China Agricultural University computational cluster for all Chinese breeds and Vienna Scientific Cluster for Austrian Fleckvieh.

BWA-mem v.0.7.17 (Li and Durbin, 2010) aligned paired-end reads of FASTQ against cattle reference genome ARS_UCD1.2 (Rosen et al., 2020), resulting in a sequence alignment map (SAM) file. Subsequently, samtools v.1.10 (Li et al., 2009) sorted SAM file by chromosomes and converted to binary alignment map (BAM). Picard (https://broadinstitute.github.io/picard/) functions of MarkDuplicates flagged duplicate reads in BAM files and function of AddOrReplaceReadGroups modified read groups information accordingly. For subsequent steps, GATK v.4.1 (McKenna et al., 2010) was used. GATK functions of BaseRecalibrator and ApplyBQSR detected and corrected base quality scores of mapped reads nearby known variants. GATK HaplotypeCaller with–ERC GVCF option called individual genotype for each BAM file.

Individual GVCFs files were combined using the GenomicsDB function of GATK, allowing combinations of GVCFs called using different versions of GATK. The joint cohort of GenotypeGVCFs called the final VCF file using parameter–allow-old-rms-mapping-quality-annotation-data since individual GVCFs were called using a different version of GATK in two different computational clusters. Subsequently, we retained single nucleotide variants using GATK SplitVcfs function and filter variants with the following parameters "QD < 2.0, QUAL <30, SOR >3, FS > 60, MQ < 40, MQRankSum <12.5, ReadPosRankSUm < -8.0" following general GATK's recommendation.

We added ancestral alleles (Naji et al., 2021a) in the info column of the vcf file separately for each autosome using–fill-aa script of VCFTools v.0.1.15 (Danecek et al., 2011). Subsequently, Bcftools v.1.7 (Li et al., 2009) retained the biallelic SNPs in the VCF. Then, genotypes in the VCF file were phased using Beagle v.5.1 (Browning et al., 2018) and indexed using tabix v.1.7-2 (Li et al., 2009) resulting final phased data for the analysis.

Principal component and admixture analysis

Before phasing steps, the multi-sample VCF file containing all autosomes was converted to binary plink format using VCFtools (Danecek et al., 2011). Plink1.9 (Chang et al., 2015) merged the dataset with additional individuals of Angus, Brahman, Gir, Holstein, Indian Zebu, Jersey, Kenana, Mangshi, Nelore, and Simmental breeds from the publicly available SRA NCBI database used in the previous study (Naji et al., 2021b). We filtered out variants with missing call rates exceeding 0.2. We used the–pca function with five eigenvectors for PCA on ~4.5 million variants that were shared by all individuals. Admixture v.1.3 (Alexander et al., 2009) assessed population structures using the same input file as the PCA with K numbers of three to five. Outputs of PCA and admixture analysis were plotted using R (R Core Team, 2020).

Scanning for SNVs driving positive selection

The iSAFE test ranks all SNVs within linkage-disequilibrium (LD) regions with selective sweep signals based on their contribution to the selection signal. The program (Akbari et al., 2018) scans for signals up to 5 Mb using statistics derived solely from haplotypes and ancestral allele information. Under the hood, iSAFE used two steps; first, it searched for the best candidate mutations using selection of allele favored by evolution (SAFE) and then combined those signals for the final iSAFE score for the maximum region spanning 5 Mb.

In the first step, haplotype allele frequency (HAF) is used to distinguish haplotypes based on the sum of derived allele counts. Haplotypes are considered 'distinct' once they have different HAF scores and 'carries a mutation *e*' if they have derived allele at site *e* with *f* mutation frequency. When a particular haplotype is a putative carrier of a favored allele, its HAF score increases due to carrying more derived alleles.

$$k(e) = \frac{number\ of\ distinct\ haplotypes\ carrying\ mutation\ e}{number\ of\ distinct\ haplotypes\ in\ sample}$$

$k(e)$ denotes a fraction of distinct haplotypes carrying mutation e, while $\phi(e)$ denotes the normalized sum of HAF scores carrying the mutation e.

$$\phi(e) = \frac{sum\ of\ HAF\ scores\ of\ haplotypes\ carrying\ mutation\ e}{sum\ of\ HAF\ scores\ of\ all\ haplotypes}$$

Based on these calculations, SAFE-score is defined as

$$SAFE(e) = \frac{\phi - k}{\sqrt{f(1-f)}}$$

Theoretically, selective sweeps will reduce the $k(e)$ score as the number of distinct haplotypes carrying favored mutations is reduced. Increasing HAF scores in carrier haplotypes will reduce the ratio of total HAF-score contributed by non-carrier haplotypes, consequently higher ϕ value. Thus, the mutation with the highest SAFE score is expected as a candidate of favored mutation.

As the k score reduces its power to pinpoint favored mutation due to most haplotypes becoming unique in larger windows, thus it applies a set of half-overlapped windows (W) with a fixed size of 300 SNPs on the second step. δ denotes a list of selected mutations e in each window with the highest SAFE score. For mutation e in window w, $\psi_{e,w'}$ denotes the larger SAFE score of e and 0 when e is inserted into window w'. $\psi_{e,w'}$ will be relatively high when e is a favored mutation and the genealogies of w and w' are very similar. $\alpha(w)$ denotes the weight of each window w which would have a high value corresponding to favored mutations contained. The iSAFE score for mutation e is calculated by the higher SAFE score of e and weight of all scanning windows.

$$\alpha(w) = \frac{\sum_{e\epsilon\delta}\psi_{e,w}}{\sum_{w'\epsilon W}\sum_{e\epsilon\delta}\psi_{e,w'}}$$

$$iSAFE(e) = \sum_{w\epsilon W}\psi_{e,w}\,.\,\alpha(w)$$

We used the built-in program to pinpoint favored mutations for a non-overlapped window of 4 Mb in autosomes for each pool of individuals. Then, we concatenated iSAFE scores for all SNPs of all autosomes and applied the normal distribution's right-tail probability density function (PDF) to infer the p-values as provided in R (R Core Team, 2020).

The iHS test was first proposed in 2006 and used for many studies to identify positive selections in livestock populations (Qanbari et al., 2014; Randhawa et al., 2016; Vatsiou et al., 2016). We used selscan (Szpiech and Hernandez, 2014) to perform the iHS test using phased data for each chromosome of individual pools. In the notation below, for each queried SNV (x_i), integrated haplotype homozygosity (iHH) of ancestral (0) and derived 1) haplotypes (C: = {0,1}) was calculated from the extended haplotype homozygosity (EHH) (Sabeti et al., 2002) from both upstream (U) and downstream (D) set of markers of each query site (x_i). $g(x_{i-1}, x_i)$ represents the genetic distance between two markers created with an arbitrary value of one centiMorgan per megabase.

$$iHH_c = \sum_{i=1}^{|D|}\frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i) + \sum_{i=1}^{|U|}\frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i)$$

The unstandardized iHS score was calculated as $iHS = ln\frac{iHH1}{iHH0}$ where a positive value indicates unusual long haplotypes carrying derived alleles compared to the neutral model indicating recent positive selection. We applied the normal distribution's right-tail probability density function (PDF) to infer the p-values as provided in R (R Core Team, 2020).

We used–weir-fst-pop in vcftools (Danecek et al., 2011) based on Fst estimation (Weir and Cockerham, 1984) to analyze selection signature between the population of each taurine breeds (Fleckvieh, Kazakh, Mongolian, and Yanbian) against a combination of all indicine and admixed Chinese individuals. Fst values were averaged with 10 Kb non-overlapping windows. The probability density function of normal distribution inferred the p-values considering the right tail only. Genome-wide significance -$\log_{10}$(p) of 7.301 was set as the threshold. For all the analysis, manhattan plots were built using the qqman R-package (Turner, 2018).

Functional annotation and gene expression

SnpEff (Cingolani et al., 2012) annotated SNVs and windows positions above the threshold using Ensembl version of ARS_UCD1.2 annotation file. For functions of individual genes, we referred to the one listed in https://www.genecards.org/and https://www.ncbi.nlm.nih.gov/gene/. We further considered only genes indicated by SNVs within coding regions. We used pantherdb.org to classify the functionality of associate genes listed by different statistical methods to its gene ontology (GO) terms. We annotated significant genes indicated by the tests for their expression level using cattle gene atlas (Fang et al., 2020). We retained the information of maximum expression level in fragments per kilo base per million mapped reads (FPKM) and its corresponding tissue where the maximum FPKM is found for each indicated gene. In this repository, the mean and standard deviation for FPKM across 91 tissues and 447 individuals are 26.79 and 730.56, respectively.

Results

Whole-genome sequence data of 60 individuals from 14 breeds of Chinese cattle and Austrian Fleckvieh were aligned against the ARS_UCD1.2 reference genome. On average, there were ~316 million paired reads per individual FASTQ, with length

TABLE 1 Alignment summary of the dataset.

Breeds	Species[a]	N animals	Total reads in million[b]	Read length[c]	Mapped reads[d]	Depth[b]
Fleckvieh	*B. taurus*	10	326.52	101	0.977	5.823
Kazakh	*B. taurus*	6	250.32	131	0.998	8.859
Mongolian	*B. taurus*	12	210.78	139	0.998	8.259
Yanbian	*B. taurus*	10	226.90	137	0.998	8.513
Dabieshan	*B. indicus*	2	345.20	96	0.992	10.257
Dehong	*B. indicus*	2	342.82	93	0.997	10.052
Dengchuan	*B. indicus*	2	346.43	94	0.997	10.346
Fujian	*B. indicus*	2	332.11	95	0.997	9.988
Guanling	*B. indicus*	2	340.23	96	0.997	10.269
Liping	*B. indicus*	2	338.78	96	0.997	10.097
Wenling	*B. indicus*	2	339.14	90	0.997	9.387
Luxi	*Admixed*	2	324.72	95	0.998	10.133
Nanyang	*Admixed*	2	369.14	96	0.998	10.333
Tibetan	*Admixed*	2	301.14	96	0.998	9.636
Qinchuan	*Admixed*	2	347.76	93	0.998	9.846

[a]Assigned species were based on the principal component analysis carried out in this study—Admixed of *Btaurus* and *B. indicus*; Superscript b, c, d, e were the average values from individuals in each respective breed.
[b]Depth values inferred from SNVs, in the final VCF, file.

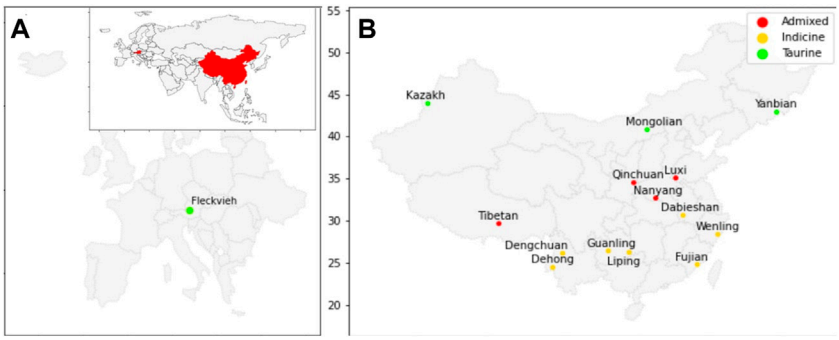


FIGURE 1
Map of origin for cattle used in this study; **(A)** Austrian Fleckvieh origin on European map with an inset of China and Austria in the world map; **(B)** Position of Chinese cattle breeds on the maps.

varies from 90 to 148 bases per read. In total, ~60 million SNVs passed the set of hard filtration for all autosomes with an average depth of ~9 × Table 1 indicated alignments summary for each breed with details provided in Supplementary Table S1. Figure 1 depicted the origin of Chinese cattle and Austrian Fleckvieh on the world map.

Principal component and admixture analysis

We inferred the population structure using PCA from ~4.5 million SNVs shared by all individuals in the dataset. The PCA explains 47.37, 10.20, 6.68, 6.52, and 5.89 percent of variance for components one to five, respectively. Figure 2 depicted the clustering of all individuals based on the first and second components. We observed a clear separation between the taurine and indicine cattle by the first component regardless of its origin. The three Chinese breeds, Kazakh, Mongolian, and Yanbian, were clustered together with Austrian Fleckvieh and other renowned *B. taurus* breeds, such as Angus, Holstein, and Jersey. Nanyang, Luxi, Qinchuan, and Tibetan were admixed as they were between clusters of *B. taurus* and *B. indicus*. While Dabieshan, Dehong, Dengchuan, Fujian, Guanling, Liping, and Wenling were clustered together with other *B. indicus* breeds such as Brahman, Nelore, and Gir.

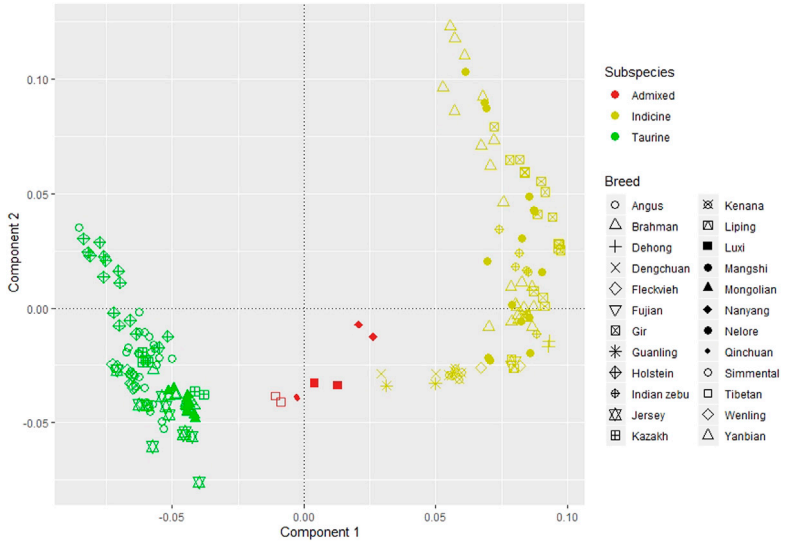


FIGURE 2
Principal component analysis; component 1 explains 47.37 percent of variants and component 2 for 10.20 percent.

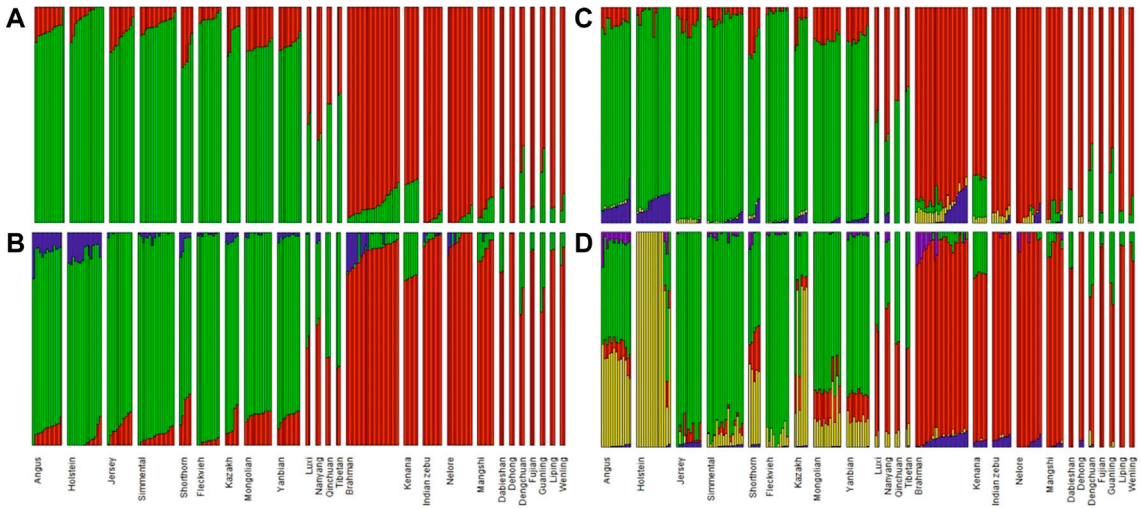


FIGURE 3
Admixture analysis using **(A)** K = 22; **(B)** K = 3; **(C)** K = 4; and **(D)** K = 5.

Admixture analysis for those breeds using K from two to five supported the results of PCA, see Figure 3. Thus, based on these results and considering the number of individuals in each breed, we pooled individuals into separate groups for further selection signature analysis, i.e. Fleckvieh, Kazakh, Mongolian, Yanbian, CHBI_Low (seven Chinese *B. indicus* breeds), CHBI_Med (four Chinese admixed breeds), and CHBI(combination of CHBI_Low and CHBI_Med).

The PCA and admixture results matched the geographical origin of individuals as in Figure 1. Kazakh, Mongolian, and Yanbian were sampled from the northern part of Chinese. CHBI_Med individuals were from the middle latitude of the country. While the CHBI_Low individuals were originated from

TABLE 2 Summary of output scores of SNVs and windows from iHS, iSAFE, and Fst tests.

Pools[a]	iSAFE test				iHS test				Fst test[f]			
	Mean ± SD[b]	Sign. SNVs[c]	Intergenic[d]	Gene[e]	Mean ± SD[b]	Sign. SNVs[c]	Intergenic[d]	Gene[e]	Mean ± SD[g]	Sign. Windows[h]	Intergenic[i]	Gene[j]
Fleckvieh	0.07 ± 0.05	2,264	1,764	161	-0.01 ± 0.70	6	3	2	0.08 ± 0.05	301	177	167
Kazakh	0.10 ± 0.08	1,502	1,039	56	-0.53 ± 1.07	0	0	0	0.06 ± 0.05	336	190	185
Mongolian	0.06 ± 0.04	3,446	2,656	111	-0.15 ± 0.64	41	18	7	0.08 ± 0.05	334	185	179
Yanbian	0.05 ± 0.03	5,068	3,681	258	-0.09 ± 0.65	91	36	8	0.07 ± 0.05	334	188	182
CHBI	0.11 ± 0.07	0	0	0	-0.01 ± 0.60	30	23	4	NA	NA	NA	NA
CHBI_Med	0.10 ± 0.07	469	368	3	-0.49 ± 0.75	59	43	11	NA	NA	NA	NA
CHBI_Low	0.09 ± 0.06	1,648	881	28	0.02 ± 0.70	5	4	1	NA	NA	NA	NA

[a]Pools: grouping of individuals - first four are specific *B. taurus*, CHBI_Low (seven Chinese *B. indicus*), CHBI_Med (four Chinese admixed), and CHBI (combination of CHBI_Low and CHBI_Med).
[b]Mean and SD, of raw values for SNVs, reported in each respective test.
[c]Number of SNVs, passing the threshold of genome-wide significance ($-\log_{10}(p) = 7.301$).
[d]Number of SNVs, annotated to intergenic regions.
[e]Number of SNVs, annotated to coding regions of genes.
[f]Fst test of respective breed against CHBI.
[g]Mean and SD, of Fst values for all windows.
[h]Number of windows passing the threshold of genome-wide significance ($-\log_{10}(p) = 7.301$).
[i]Number of windows annotated to intergenic regions.
[j]Number of windows annotated to coding regions of gene.

10.3389/fgene.2022.974787

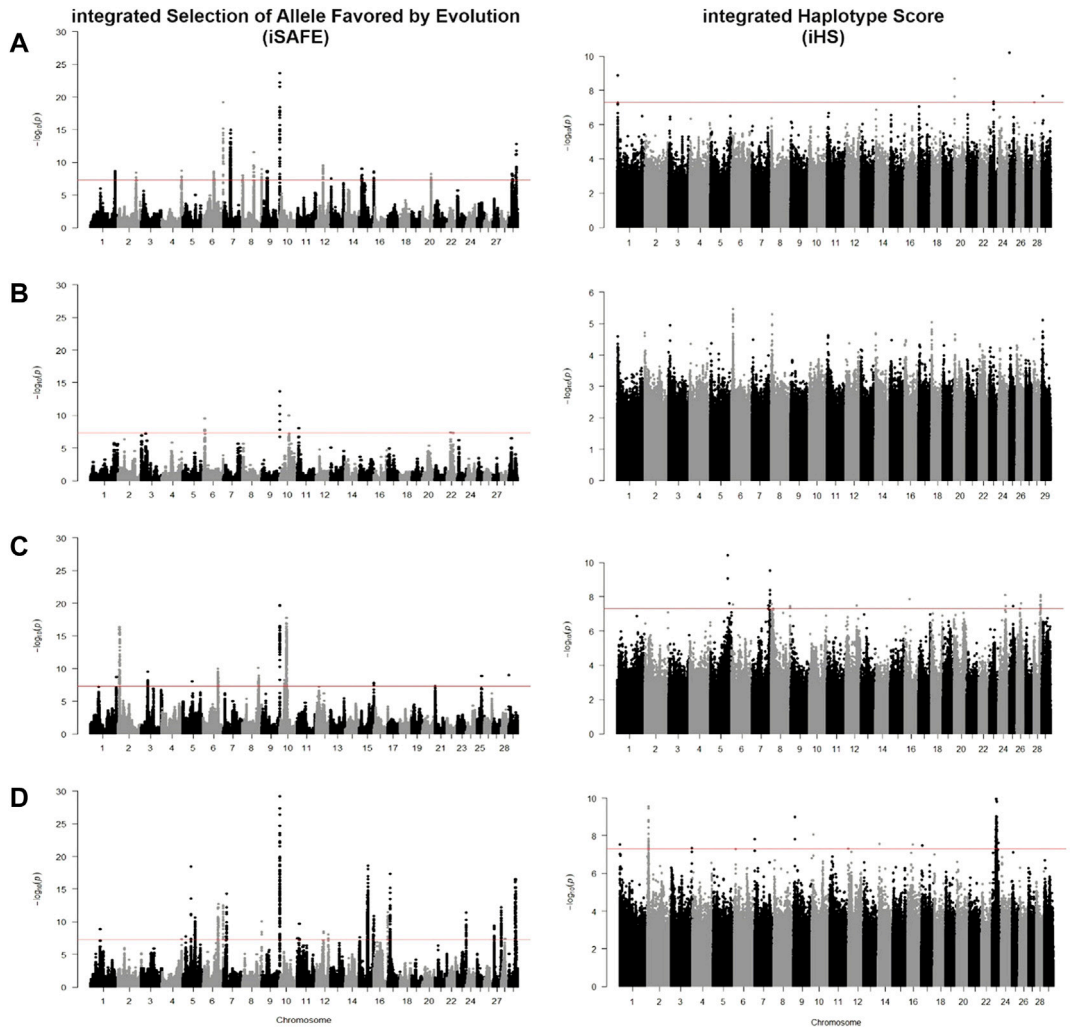


FIGURE 4
Manhattan plot for iSAFE and iHS tests for **(A)** Fleckvieh; **(B)** Kazakh; **(C)** Mongolian; and **(D)** Yanbian.

the southern part of the country. Coincidentally, Austrian Fleckvieh originated from a region with latitude around 48°, while three Chinese taurine breeds were also coming from a similar temperate climate of 42° latitudes.

Comparison of methods in scanning for positive selection

We identified candidate SNVs for positive selection using two within-population tests of iSAFE and iHS. Additionally, we performed one cross-population test of Fst between taurine breeds against CHBI individuals. Phased vcf files were separated for each pool of individuals and underwent both tests, respectively. Table 2 indicated the descriptive statistics of significant SNVs of respective tests and individual pools. Figure 4 and Figure 5 depicted the manhattan plots of *B. taurus* and other CHBI groups, respectively.

Using iSAFE, we found several peaks of SNVs for *B. taurus* breeds. For all *B. taurus*, the strongest signals come from chromosome nine around 104.4 Mb. CHBI_Med and CHBI_Low had independently signals on chromosome 20 and 1, respectively. We did not find significant SNVs for CHBI. Around 71 percent of SNVs indicated as significant in the iSAFE test were annotated to intergenic regions, as stated in Table 2. A list of significant genes indicated by iSAFE, iHS, and Fst is provided in Supplementary Tables S2-S4.

In Fleckvieh, Kazakh, Mongolian, and Yanbian, two genes of *ENSBTAG00000045624* and *ENSBTAG00000047934*, known also as *OR10D1M*, were genes indicated by the most significant iSAFE score. The later gene belongs to the olfactory receptor family, which interacts with odorant molecules in the nose, initiating the

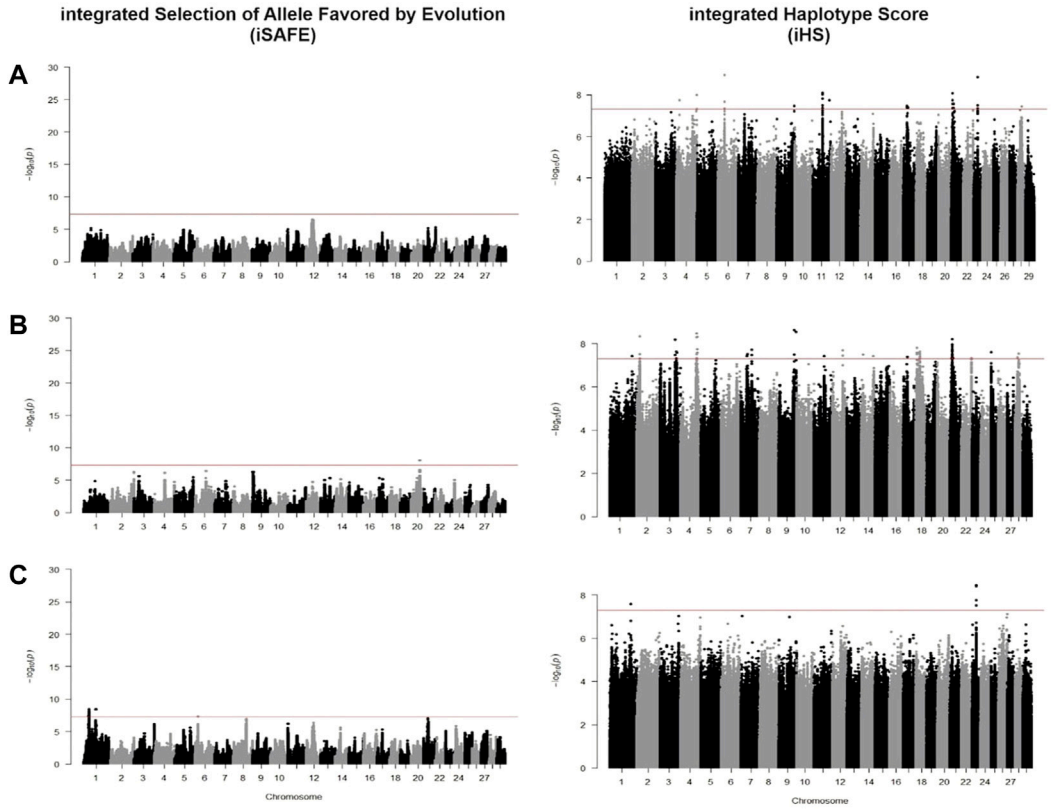


FIGURE 5
Manhattan plot for iSAFE and iHS tests for **(A)** CHBI; **(B)** CHBI_Med; and **(C)** CHBI_Low.

neuronal response that starts a sense of smell. This gene was neither found in significant regions of CHBI_Low nor CHBI_Med. *ENSBTAG00000053225* (*OR8B60*) and *ENSBTAG00000050546* (*OR8AR1*) were olfactory receptor genes indicated in Fleckvieh, Kazakh, and Yanbian. All these top indicated genes are located at chr 9, around 104.3 Mb. Within 100 Kb vicinity of these olfactory genes, we found *FAM120B, DLL1, PSMB1*, and *PDCD2. FAM120B* has several associations of twinning rate in mammals, fat deposition in chicken and inflects pig sperm maturation during spermatogenesis due to its function in adipogenesis regulation of *PPARG* (Vinet et al., 2012; Moreira et al., 2015; Gòdia et al., 2020). In human, *DLL1* plays role in Notch signaling pathway regulating cell differentiation and proliferation in embryonic development and maintenance of adult stem cells (Jaleco et al., 2001). In cattle, activation of Notch pathway by miRNA targeting *DLL1* leads to restrain adipose differentiation which might lead to different subcutaneous adipose tissue between Wagyu and Holstein (Guo et al., 2017). While in embryo development, *in vitro* expression of *PSMB1* is significantly reduced after bovine oocyte maturation (Adona et al., 2011). Similarly, *PDCD2* plays also role in embryo development as indicated of its activation during bovine 16-cell stage (Graf et al., 2014).

Overlapped genes found in Fleckvieh and Yanbian were *ACP1, ALKAL2, ENSBTAG00000045328, ENSBTAG00000045624, ENSBTAG00000047934, ENSBTAG00000050546, ENSBTAG00000051204, ENSBTAG00000053225, POLN, SH3YL1* and *U6*. *ALKAL2* is associated with reproduction function and upregulated in granulosa cell of bacteria-infected uterus in Holstein heifers (Horlock et al., 2020) while *POLN* was reported to influence mature body size in US sheep population (Posbergh and Huson, 2021). For CHBI_Med, *CDH12* is associated with longevity and desaturation of milk fatty acids as reported in few dairy cattle (Mészáros et al., 2014; Cecchinato et al., 2019). For CHBI_Low, *CYP2U1* is linked to milk fat secretion in Sahiwal cattle in India (Illa et al., 2021).

Using iHS, we did not find significant SNVs for Kazakh. In contrary, Yanbian had 91 significant SNVs. These SNVs were observed as a peak in chromosome 23. Significant SNV at chr23: 26, 067, 413 was detected both in Yanbian and Fleckvieh. For Mongolian, we observed several peaks on chromosomes five and 7. A total of 30, 59, and five SNVs were above the threshold for

CHBI, CHBI_Med, and CHBI_Low, respectively, with no overlaps among them. For all groups, the mean iHS score was generally in a negative value except for the CHBI_Low. 58 percent of significant SNVs in iHS were annotated to intergenic regions, as indicated in Table 2.

For Fleckvieh, SNVs with significant iHS at chromosome 20 around 3.8 Mb overlapped to *STK10* gene, which is significantly associated with slaughter weight and carcass quality in several beef cattle breeds (Karisa et al., 2013; Hay and Roberts, 2018). For Mongolian, SNVs with significant iHS scores were overlapped with the novel gene of *ENSBTAG00000050324, PTPRM, GRID1, CACNA1C, SORCS3, NRG3*, and *TXNDC2. PTPRM* has extended function in regulating cellular growth, differentiation, mitotic cycle, and is associated with scrotal circumference in Nellore and Brahman cattle (Melo et al., 2019). *GRID1* is known for its function in the central nervous system and is down-regulated in fetuses carrying deletion variants in *PEG3* domain leading to stillbirth (Flisikowski et al., 2012). *CACNA1C* is linked to immune defense and was hyper-methylated in Angus during stress of high-temperature high-humidity period (Del Corvo et al., 2021). *SORCS3* was highly associated with temperament trait and average daily gain (Xu et al., 2019; Shen et al., 2022). *NRG3* is associated with fat yield component in sheep production (García-Gámez et al., 2012). While *TXNDC2* is linked to average daily gain and age at puberty in Korean cattle (Edea et al., 2020).

For Yanbian, the top genes indicated by significant iHS score were *ENSBTAG00000026163, ENSBTAG00000007075, C2H2orf88, TBCA, HIBCH, TMEM71, SMYD3, ARFIP1,* and *HDAC4*, which all these genes play a role in cellular proliferation and transcription factors. *TBCA* is associated with sire conception rate in the US Jersey cattle (Rezende et al., 2018). *HIBCH* is one of candidate genes in association study of calving performance in Charolais population (Purfield et al., 2015). While *ARFIP1* is associated with milk production traits in Holstein (Lee et al., 2016).

For CHBI, significant SNVs were in novel genes of *ENSBTAG00000026163, ENSBTAG00000053922, PBLD*, and *AFDN*, which encodes a multi-domain protein involved in signaling and organization of cell junctions during embryogenesis. For CHBI_Med, *ENSBTAG00000020723, PDE10A, AKAP13, KLHL25, ENSBTAG00000054043, FAM234A, PODXL, RUVBL1*, and *ABHD1* were indicated. *AKAP13* and *KLHL25* were also reported as selection candidates in North African cattle (Ben-Jemaa et al., 2020). *RUVBL1* is associated with tolerance of African cattle towards heat and parasites stress (Taye et al., 2017a; Yougbaré et al., 2021). While in CHBI_Low, *ENSBTAG00000026163* is a gene indicated by significant SNV.

Using Fst test, we found a similar pattern among the *B. taurus* cattle as shown in Figure 6. There were 224 similar significant windows in Fleckvieh Kazakh, 181 in Fleckvieh-Mongolian, 208 in Fleckvieh-Yanbian, 185 in Kazakh-Mongolian, 217 in Kazakh-Yanbian and 222 in Mongolian-Yanbian. Among these, 132 windows were significant in all of Fleckvieh, Kazakh, Mongolian, and Yanbian. Out of 132, 73 windows were annotated to intergenic regions and the rest to 60 genes. Within these genes, *LCT* was reported as selection candidates in several Italian cattle and its mutations in humans, irrespective of location and mutation type, are linked to congenital lactase deficiency (Torniainen et al., 2009; Sorbolini et al., 2016). *DHRS3* is known for its importance in retinoic acid metabolism and is essential in regulating body axis formation during embryonic development (Kam et al., 2013). *PRKCZ*, where young calves are exposed to hypoxia, leads to anti-replication activity of cells related to this gene in pulmonary artery adventitia (Das et al., 2008).

Several genes indicated by Fst were found exclusively in three Chinese taurine breeds and not in Fleckvieh. These genes might be related to the adaptation process to the local habitat. For example, *ANXA10* was detected as a selection candidate in Kholmogory cattle and deletion-type CNV of 34 kb identified in this gene was linked to embryonic mortality in Japanese Black cattle (Sasaki et al., 2016; Yurchenko et al., 2018). *C14H8orf34* is associated with claims on epinephrine hormone excretion in the urinary due to many pathways of metabolism acceleration under stress situations (de Camargo et al., 2015). *CACNA2D3* plays role in active calcium ion transport and was highly expressed in *Longissimus Lumborom* than *Psoas Major* muscles after postmortem in Chinese Jinjiang cattle (Yu et al., 2019). *PAX5* is associated with the proportion of black color in Holstein (Hayes et al., 2010). *PLAG1* is a known gene with pleiotropic effects on body weight and fertility traits (Fortes et al., 2013; de Camargo et al., 2015; Yurchenko et al., 2018). *TAC3* was associated with reproduction process as highly expressed in non-pregnant heifers compared to heifers that later became pregnant (Dickinson et al., 2018).

Go classification and expression level

GO classification for genes indicated by iHS, iSAFE, and Fst tests is provided in Supplementary Table S5. Cellular process (GO:0009987) was the top GO term indicated by significant genes irrespective of the test and individuals pool. Metabolic process (GO:0008152) was the second top GO term for iSAFE test in Fleckvieh and Mongolian, while for Kazakh and Yanbian, the second top was biological regulation (GO:0065007). There were 10, 5, 2, and 13 genes for developmental process (GO:0032502) indicated by iSAFE test for Fleckvieh, Kazakh, Mongolian, and Yanbian, respectively. In general, Yanbian had more coverage to broader GO terms like reproduction (GO:0000003), reproductive process (GO:0022414), multi-organism process (GO:0051704), growth (GO:0040007) perhaps due to the higher number of genes detected by iSAFE compared to Fleckvieh, Kazakh, and Mongolian.

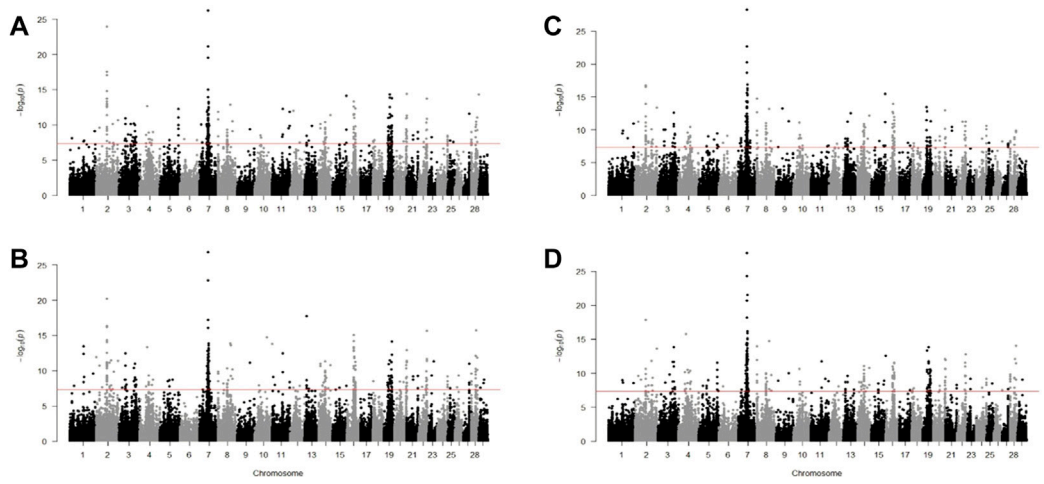


FIGURE 6
Manhattan plot for Fst test for **(A)** Fleckvieh; **(B)** Kazakh; **(C)** Mongolian; and **(D)** Yanbian against CHBI.

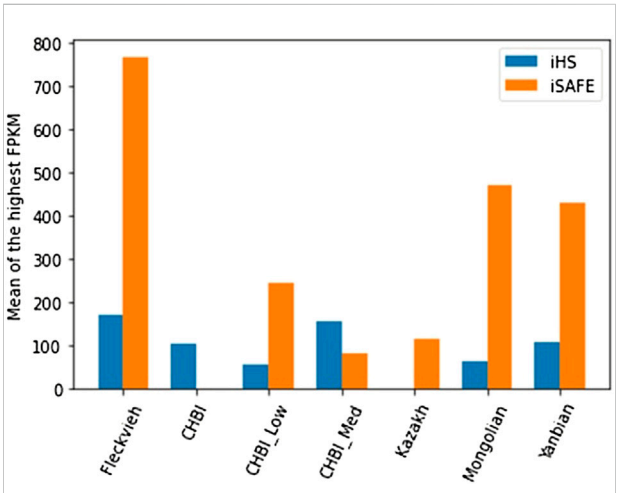


FIGURE 7
Mean of the highest FPKM for all significant genes indicated by iHS and iSAFE tests in each respective individual pool.

We annotated significant genes indicated by iHS and iSAFE toward their maximum expression level (FPKM) and the corresponding tissue as listed in the repository of cattle gene atlas (Fang et al., 2020). Rectangular bar in Figure 7 depicted the average FPKM of significant genes for the corresponding test and individual pool. In general, the mean FPKM values by both tests were higher than the mean of FPKM records of the full repository (26.79). iSAFE indicated higher mean of FPKM than the iHS except for CHBI_Med and CHBI where no SNVs were significant in iSAFE tests. Supplementary Figures S1, S2 depicted cloud plots for the associated tissues with the FPKM for iSAFE and iHS tests. Genes indicated by iHS were mostly highly expressed in the ileum tissue as indicated in Fleckvieh, Yanbian, CHBI, and CHBI_Low. For iSAFE, the significant genes for all individual pools were all highly expressed in the sperm, see Supplementary Figure S1.

For Fleckvieh, *TFF1* was the gene listed by iSAFE with the highest FPKM (10,413) in abomasum tissue. *TFF1* was reported to be associated with mammary development and secretion of minerals to the bovine milk (Stella et al., 2010; Gao et al., 2017). Ten modifiers and one low impact were estimated for the SNVs indicated within *TFF1* in Fleckvieh. For Kazakh, the highest expressed gene was *ALDH1A2* with 250 FPKM from stalk median eminence tissue and is associated with carcass weight in beef cattle (Willing et al., 2012). Functional modifiers' impact were annotated for all 13 SNVs in *ALDH1A2*. For Mongolian cattle, *ALDOA* was the highest expressed gene with 12,960 FPKM in chorid plexus tissue in the brain. *ALDOA* is primarily related to glycolytic and energy metabolism (Wærp et al., 2019). *TNNT2* identified in Yanbian was highly expressed in heart tissue with 5006 FPKM. Seven SNVs with modifier impact were associated with this gene, which is related to the striated muscle contraction due to intracellular calcium ion concentration and found in a previous study as selection candidates in Holstein cattle (Taye et al., 2017b).

Discussion

This study indicated a gradual shifting of taurine cattle in northern China to admixed and pure indicine cattle towards the southern part of China, similar to the report from the previous studies (Chen et al., 2018; Zhang et al., 2020). Zhang et Al (2020)

indicated that Mongolian and Kazakh, two Chinese taurine in our study, were well adapted to cold winters. They suggested that the admixture and introgression of taurine and indicine from north to south can be affiliated to loci in the genome, which might help individuals adapt to the local environment (Wu et al., 2018). A previous study (Gao et al., 2017) suggested that Chinese taurine cattle in the north shared the same genetic ancestry to several Central Asia, Russian-Yakutstk, Korean and Japanese cattle (Turano-Mongolian) in the greater region due to past activities of nomads and the Mongolian empire.

Previous studies used various selection signature methods and normalized the outputs score using specific windows, in basepairs or based on the number of SNPs, to identify the candidate regions (Qanbari et al., 2014; Xu et al., 2015; Yurchenko et al., 2018; Bhati et al., 2020). We applied a similar approach for the Fst test using non-overlapping 10 Kb windows. However, for within-population tests of iHS and iSAFE, we did an experimental analysis to point out the causal SNV mutations in coding regions that significantly drive selective sweeps in genome-wide level. We found that iHS indicated fewer signals passing the genome-wide significant threshold than iSAFE in any breed. This is in line with simulations in the original paper where iSAFE could detect almost double the signals for favoured mutations than iHS (Akbari et al., 2018). Both methods were associated with declining performance in detecting mutations in regions that are closed to fixation, yet we found no overlapped genes indicated by these two tests.

Generally, our study suggested higher selection signals for taurine than indicine cattle in both iSAFE and iHS tests. For example, in the iHS test, Yanbian had 91 significant SNVs while CHBI_Med had only 59. Similarly, in the iSAFE test, Yanbian had around five thousand significant SNVs while CHBI_Med had a far less, around 469 SNVs. Our finding is similar to previous study where indicine cattle of Gyr and Nelore had substantially fewer regions proposed as selection evidence compared to taurine cattle of Brown Swiss and Angus (Utsunomiya et al., 2013). Moreover, pools of indicine cattle in our study were a combination of several breeds due to limited number of individuals to a suggested minimum of six individuals for better population genetic analysis (Willing et al., 2012). Thus as the results, we observed decayed of the iSAFE signal in CHBI as the combination of CHBI_Low and CHBI_Med, compared to the scenario when both groups were tested independently. We assumed that the signals for each indicine breed would be more significant and apparent if the sample size were equal to the taurine breed. However, due to the circumstances, we could not do it for the current study.

As indicated in the results section, more than half of the signals fall under intergenic regions. We did not consider SNVs found in those regions and retained only SNPs in the coding regions. Within these SNVs, several were without official gene ID names. For example, 15 SNVs creating a peak in iHS test chromosome 23 of Yanbian were in an active transcription of ENSBTAG00000007075 gene. According to https://bgee.org/?page=gene&gene_id=ENSBTAG00000007075, this gene was described as a major histocompatibility complex, class I, A-like precursor and has paralogs to BOLA-A and JSP. 1 genes. And has an association with feeding efficiency in Norwegian Red heifers where it is upregulated during diet changes from low-protein-high-energy to low-protein-low-energy feed (Wærp et al., 2019). Yet, for GO classification, we considered only genes with official ID names overlooking functions of genes with prefix ENBST names.

KIT was indicated as one of selection candidate genes affecting coat colors (Flori et al., 2009; Stella et al., 2010; Xu et al., 2015). In chromosome six around 70 Mb, where KIT is located, there were 236 SNPs in the phased genotypes. Yet, we did not find any significant SNVs passing the genome-wide threshold, though the maximum iSAFE scores ranges from 0.04 to 0.20 among the *B. taurus,* see Supplementary Figure S3. Apparently, the threshold for iSAFE on genome-wide level has biased the findings to SNVs within highly-scores segments. Meanwhile each genome segment may have different significance level for assigning SNV as the best candidate of selection. This was demonstrated in the original manuscript where a SNV with score of 0.10 was the best candidate in *HBB* while score of 0.61 was the best candidate for *EDAR* (Akbari et al., 2018). However, genes indicated by genome-wide threshold of iSAFE might act as the driver of selection within the LD segments as they had the highest scores. Though the functionality of these genes were quite spurious, generally they had higher expression in tissues, particularly in sperm, compared to ones indicated by iHS.

In the Fst test, we found *PAX5* as a candidate gene in three Chinese taurine breeds, not shared with Fleckvieh, which function is associated with black color patterns (Hayes et al., 2010). In general, Chinese indigenous cattle, including these three breeds, are considered as 'yellow' cattle, thought the they are actually in different level of brownish colors. A specific *PAX5* might affect the color pattern of these breeds, separating them from other Turano-Mongolian cattle, such as the Mongolian and Korean cattle, which still retain their original dark-brown coat color pattern (Gao et al., 2017).

Our findings suggested that three Chinese taurine cattle breeds shared a considerable amount of candidate regions with Fleckvieh. Though we can confirm that there was no recorded genetic material exchange between Austria and China, it was reported that there were programs for improving the productivity of local breeds by crossing to European breeds in the last decades (Gao et al., 2017). As those European breeds might have similar characteristics as Fleckvieh, thus we cannot attribute similarity between Austrian Fleckvieh and Chinese taurine solely due to independent co-selection of nature, but also possibly due to recent crossing with other European breeds.

Conclusion

Our study confirmed a gradient of taurine and indicine admixed cattle from north to south of China. More significant SNVs were identified using iSAFE than the iHS for the candidate of positive selection and more detectable signals in taurine than in indicine individuals. However, combining individuals of different breeds decaying the iSAFE signals. From both tests, significant SNVs are linked to the olfactory receptors, production, reproduction, and temperament traits in taurine cattle, while heat and parasites tolerance in the admixed individuals. Fst test suggests similar patterns of population differentiation between Fleckvieh and three Chinese taurine breeds against Chinese indicine breeds. However, there are genes shared only among the Chinese taurine, such as PAX5, affecting black coat color, which might underlying differences of these breeds to other Turano-Mongolian cattle.

Data availability statement

The datasets presented in this article are not readily available because the animals and subsequently their genomic data are property of respective breeding organizations. Requests to access these datasets should be directed to GM (gabor.meszaros@boku.ac.at) for the data set from Austria or to XD (xiangdongding@hotmail.com) for the data sets from China.

Author contributions

XD and GM conceived and designed the study. YJ collected the samples. MMN and YJ ran the analysis. MMN drafted the manuscript. YTU, BDR, JS, CW, LJ, QZ, YZ, XD, and GM interpreted the analysis results and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Funding

Collaborative project of University of Natural Resources and Life Sciences, Vienna and Chinese Agricultural University, Beijing is supported by the National Key Research and Development Project (2019YFE0106800) and WTZ project. MMN is supported by the Ernst Mach Grant, ASEA UNINET (OeAD Austria).

Acknowledgments

Computations were carried out in Chinese Agricultural University HPC and Vienna Scientific Cluster (VSC3)

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.2022.974787/full#supplementary-material

References

Adona, P., de Bem, T., Mesquita, L., Rochetti, R., and Leal, C. (2011). Embryonic development and gene expression in oocytes cultured *in vitro* in supplemented pre-maturation and maturation media. *Reproduction Domest. Animals* 46, e31–e38. doi:10.1111/j.1439-0531.2010.01618.x

Akbari, A., Vitti, J. J., Iranmehr, A., Bakhtiari, M., Sabeti, P. C., Mirarab, S., et al. (2018). Identifying the favored mutation in a positive selective sweep. *Nat. Methods* 15, 279–282. doi:10.1038/nmeth.4606

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

Ben-Jemaa, S., Mastrangelo, S., Lee, S.-H., Lee, J. H., and Boussaha, M. (2020). Genome-wide scan for selection signatures reveals novel insights into the adaptive capacity in local North African cattle. *Sci. Rep.* 10, 19466. doi:10.1038/s41598-020-76576-3

Bhati, M., Kadri, N. K., Crysnanto, D., and Pausch, H. (2020). Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics* 21, 27. doi:10.1186/s12864-020-6446-y

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

Cecchinato, A., Macciotta, N. P. P., Mele, M., Tagliapietra, F., Schiavon, S., Bittante, G., et al. (2019). Genetic and genomic analyses of latent variables related to the milk fatty acid profile, milk composition, and udder health in dairy cattle. *J. Dairy Sci.* 102, 5254–5265. doi:10.3168/jds.2018-15867

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat. Commun.* 9, 2337. doi:10.1038/s41467-018-04737-0

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly. (Austin)* 6, 80–92. doi:10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., et al.1000 Genomes, and Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Das, M., Burns, N., Wilson, S. J., Zawada, W. M., and Stenmark, K. R. (2008). Hypoxia exposure induces the emergence of fibroblasts lacking replication repressor signals of PKCzeta in the pulmonary artery adventitia. *Cardiovasc. Res.* 78, 440–448. doi:10.1093/cvr/cvn014

de Camargo, G., Aspilcueta-Borquis, R., Fortes, M., Porto-Neto, R., Cardoso, D., Santos, D., et al. (2015). Prospecting major genes in dairy buffaloes. *BMC Genomics* 16, 872. doi:10.1186/s12864-015-1986-2

de Simoni Gouveia, J. J., da Silva, M. V. G. B., Paiva, S. R., and de Oliveira, S. M. P. (2014). Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37, 330–342. doi:10.1590/s1415-47572014000300004

Del Corvo, M., Lazzari, B., Capra, E., Zavarez, L., Milanesi, M., Utsunomiya, Y. T., et al. (2021). Methylome patterns of cattle adaptation to heat stress. *Front. Genet.* 12, 633132. doi:10.3389/fgene.2021.633132

Dickinson, S. E., Griffin, B. A., Elmore, M. F., Kriese-Anderson, L., Elmore, J. B., Dyce, P. W., et al. (2018). Transcriptome profiles in peripheral white blood cells at the time of artificial insemination discriminate beef heifers with different fertility potential. *BMC Genomics* 19, 129. doi:10.1186/s12864-018-4505-4

Edea, Z., Jung, K. S., Shin, S.-S., Yoo, S.-W., Choi, J. W., and Kim, K.-S. (2020). Signatures of positive selection underlying beef production traits in Korean cattle breeds. *J. Anim. Sci. Technol.* 62, 293–305. doi:10.5187/jast.2020.62.3.293

Fang, L., Cai, W., Liu, S., Canela-Xandri, O., Gao, Y., Jiang, J., et al. (2020). Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* 30, 790–801. doi:10.1101/gr.250704.119

FAO (2015). *The second report on the state of the world's animal genetic resources for food and agriculture*. Rome, Italy: Food and Agriculture Organization.

Flisikowski, K., Venhoranta, H., Bauersachs, S., Hänninen, R., Fürst, R. W., Saalfrank, A., et al. (2012). Truncation of MIMT1 gene in the PEG3 domain leads to major changes in placental gene expression and stillbirth in cattle. *Biol. Reprod.* 87, 140. doi:10.1095/biolreprod.112.104240

Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S., et al. (2009). The genome response to artificial selection: A case study in dairy cattle. *PLoS One* 4, e6595. doi:10.1371/journal.pone.0006595

Fortes, M. R. S., Reverter, A., Kelly, M., McCulloch, R., and Lehnert, S. A. (2013). Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. *Andrology* 1, 644–650. doi:10.1111/j.2047-2927.2013.00101.x

Gao, Y., Gautier, M., Ding, X., Zhang, H., Wang, Y., Wang, X., et al. (2017). Species composition and environmental adaptation of indigenous Chinese cattle. *Sci. Rep.* 7, 16196. doi:10.1038/s41598-017-16438-7

García-Gámez, E., Gutiérrez-Gil, B., Sahana, G., Sánchez, J.-P., Bayón, Y., and Arranz, J.-J. (2012). GWA analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the LALBA gene. *PLOS ONE* 7, e47782. doi:10.1371/journal.pone.0047782

Gòdia, M., Casellas, J., Ruiz-Herrera, A., Rodríguez-Gil, J. E., Castelló, A., Sánchez, A., et al. (2020). Whole genome sequencing identifies allelic ratio distortion in sperm involving genes related to spermatogenesis in a swine model. *DNA Res.* 27, dsaa019. doi:10.1093/dnares/dsaa019

Graf, A., Krebs, S., Heininen-Brown, M., Zakhartchenko, V., Blum, H., and Wolf, E. (2014). Genome activation in bovine embryos: Review of the literature and new insights from RNA sequencing experiments. *Anim. Reprod. Sci.* 149, 46–58. doi:10.1016/j.anireprosci.2014.05.016

Guo, Y., Zhang, X., Huang, W., and Miao, X. (2017). Identification and characterization of differentially expressed miRNAs in subcutaneous adipose between Wagyu and Holstein cattle. *Sci. Rep.* 7, 44026. doi:10.1038/srep44026

Hay, E. H., and Roberts, A. (2018). Genome-wide association study for carcass traits in a composite beef cattle breed. *Livest. Sci.* 213, 35–43. doi:10.1016/j.livsci.2018.04.018

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.* 6, e1001139. doi:10.1371/journal.pgen.1001139

Horlock, A. D., Piersanti, R. L., Ramirez-Hernandez, R., Yu, F., Ma, Z., Jeong, K. C., et al. (2020). Uterine infection alters the transcriptome of the bovine reproductive tract three months later. *Reproduction* 160, 93–107. doi:10.1530/REP-19-0564

Illa, S. K., Mukherjee, S., Nath, S., and Mukherjee, A. (2021). Genome-wide scanning for signatures of selection revealed the putative genomic regions and candidate genes controlling milk composition and coat color traits in sahiwal cattle. *Front. Genet.* 12, 699422. doi:10.3389/fgene.2021.699422

Jaleco, A. C., Neves, H., Hooijberg, E., Gameiro, P., Clode, N., Haury, M., et al. (2001). Differential effects of Notch ligands Delta-1 and Jagged-1 in human lymphoid differentiation. *J. Exp. Med.* 194, 991–1002. doi:10.1084/jem.194.7.991

Kalcher, L., Fürst, C., and Egger-Danner, C. (2018). *Jahresbericht 2017* Vienna: ZuchtData Austria.

Kam, R. K. T., Shi, W., Chan, S. O., Chen, Y., Xu, G., Lau, C. B.-S., et al. (2013). Dhrs3 protein attenuates retinoic acid signaling and is required for early embryonic patterning. *J. Biol. Chem.* 288, 31477–31487. doi:10.1074/jbc.M113.514984

Karisa, B. K., Thomson, J., Wang, Z., Bruce, H. L., Plastow, G. S., and Moore, S. S. (2013). Candidate genes and biological pathways associated with carcass quality traits in beef cattle. *Can. J. Anim. Sci.* 93, 295–306. doi:10.4141/cjas2012-136

Lee, Y.-S., Shin, D., Lee, W., Taye, M., Cho, K., Park, K.-D., et al. (2016). The prediction of the expected current selection coefficient of single nucleotide polymorphism associated with Holstein milk yield, fat and protein contents. *Asian-Australas. J. Anim. Sci.* 29, 36–42. doi:10.5713/ajas.15.0476

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., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M., and Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proc. Natl. Acad. Sci. U. S. A.* 91, 2757–2761. doi:10.1073/pnas.91.7.2757

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

Melo, T. P., Fortes, M. R. S., Fernandes Junior, G. A., Albuquerque, L. G., and Carvalheiro, R. (2019). Rapid communication: Multi-breed validation study unraveled genomic regions associated with puberty traits segregating across tropically adapted breeds. *J. Anim. Sci.* 97, 3027–3033. doi:10.1093/jas/skz121

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

Moreira, G. C. M., Godoy, T. F., Boschiero, C., Gheyas, A., Gasparin, G., Andrade, S. C. S., et al. (2015). Variant discovery in a QTL region on chromosome 3 associated with fatness in chickens. *Anim. Genet.* 46, 141–147. doi:10.1111/age.12263

Naji, M. M., Utsunomiya, Y. T., Sölkner, J., Rosen, B. D., and Mészáros, G. (2021b). Assessing *Bos taurus* introgression in the UOA *Bos indicus* assembly. *Genet. Sel. Evol.* 53, 96. doi:10.1186/s12711-021-00688-1

Naji, M., Utsunomiya, Y., Sölkner, J., Rosen, B., and Mészáros, G. (2021a). Investigation of ancestral alleles in the Bovinae subfamily. *BMC Genomics* 22, 108. doi:10.1186/s12864-021-07412-9

National Bureau of Statistics (2018). *China statistical yearbook 2018*. Xicheng District, Beijing: National Bureau of Statistics of China.

Posbergh, C. J., and Huson, H. J. (2021). All sheeps and sizes: A genetic investigation of mature body size across sheep breeds reveals a polygenic nature. *Anim. Genet.* 52, 99–107. doi:10.1111/age.13016

Purfield, D. C., Bradley, D. G., Evans, R. D., Kearney, F. J., and Berry, D. P. (2015). Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. *Genet. Sel. Evol.* 47, 47. doi:10.1186/s12711-015-0126-4

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T. M., Fries, R., et al. (2014). Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.* 10, e1004148. doi:10.1371/journal.pgen.1004148

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation.

Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2016). A meta-assembly of selection signatures in cattle. *PLoS One* 11, e0153013. doi:10.1371/journal.pone.0153013

Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 15, 34. doi:10.1186/1471-2156-15-34

Rezende, F. M., Dietsch, G. O., and Peñagaricano, F. (2018). Genetic dissection of bull fertility in US Jersey dairy cattle. *Anim. Genet.* 49, 393–402. doi:10.1111/age.12710

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 9, giaa021. doi:10.1093/gigascience/giaa021

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi:10.1038/nature01140

Sasaki, S., Ibi, T., Akiyama, T., Fukushima, M., and Sugimoto, Y. (2016). Loss of maternal ANNEXIN A10 via a 34-kb deleted-type copy number variation is associated with embryonic mortality in Japanese Black cattle. *BMC Genomics* 17, 968. doi:10.1186/s12864-016-3312-z

Shen, J. F., Chen, Q. M., Zhang, F. W., Hanif, Q., Huang, B. Z., Chen, N. B., et al. (2022). Genome-wide association study identifies quantitative trait loci affecting cattle temperament. *Zool Res.* 43 (1), 14–25. doi:10.24272/j.issn.2095-8137.2021.176

Sorbolini, S., Gaspa, G., Steri, R., Dimauro, C., Cellesi, M., Stella, A., et al. (2016). Use of canonical discriminant analysis to study signatures of selection in cattle. *Genet. Sel. Evol.* 48, 58. doi:10.1186/s12711-016-0236-7

Stella, A., Ajmone-Marsan, P., Lazzari, B., and Boettcher, P. (2010). Identification of selection signatures in cattle breeds selected for dairy production. *Genetics* 185, 1451–1461. doi:10.1534/genetics.110.116111

Szpiech, Z. A., and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827. doi:10.1093/molbev/msu211

Taye, M., Lee, W., Caetano-Anolles, K., Dessie, T., Hanotte, O., Mwai, O. A., et al. (2017a). Whole genome detection of signature of positive selection in African cattle reveals selection for thermotolerance. *Animal Sci. J.* 88, 1889–1901. doi:10.1111/asj.12851

Taye, M., Lee, W., Jeon, S., Yoon, J., Dessie, T., Hanotte, O., et al. (2017b). Exploring evidence of positive selection signatures in cattle breeds selected for different traits. *Mamm. Genome* 28, 528–541. doi:10.1007/s00335-017-9715-6

Torniainen, S., Freddara, R., Routi, T., Gijsbers, C., Catassi, C., Höglund, P., et al. (2009). Four novel mutations in the lactase gene (LCT) underlying congenital lactase deficiency (CLD). *BMC Gastroenterol.* 9, 8. doi:10.1186/1471-230X-9-8

Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* 3 (25), 731. doi:10.21105/joss.00731

Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., et al. (2013). Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLOS ONE* 8, e64280. doi:10.1371/journal.pone.0064280

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Mol. Ecol.* 25, 89–103. doi:10.1111/mec.13360

Vinet, A., Drouilhet, L., Bodin, L., Mulsant, P., Fabre, S., and Phocas, F. (2012). Genetic control of multiple births in low ovulating mammalian species. *Mamm. Genome* 23, 727–740. doi:10.1007/s00335-012-9412-4

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

Wærp, H. K. L., Waters, S. M., McCabe, M. S., Cormican, P., and Salte, R. (2019). Long-term effects of prior diets, dietary transition and pregnancy on adipose gene expression in dairy heifers. *PLOS ONE* 14, e0218723. doi:10.1371/journal.pone.0218723

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38 (6), 1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x

Willing, E.-M., Dreyer, C., and van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE* 7, e42649. doi:10.1371/journal.pone.0042649

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., et al. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat. Ecol. Evol.* 2, 1139–1145. doi:10.1038/s41559-018-0562-y

Xu, L., Bickhart, D. M., Cole, J. B., Schroeder, S. G., Song, J., Tassell, C. P. V., et al. (2015). Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.* 32, 711–725. doi:10.1093/molbev/msu333

Xu, L., Yang, L., Wang, L., Zhu, B., Chen, Y., Gao, H., et al. (2019). Probe-based association analysis identifies several deletions associated with average daily gain in beef cattle. *BMC Genomics* 20, 31. doi:10.1186/s12864-018-5403-5

Yougbaré, B., Soudré, A., Ouédraogo, D., Zoma, B. L., Tapsoba, A. S. R., Sanou, M., et al. (2021). Genome-wide association study of trypanosome prevalence and morphometric traits in purebred and crossbred Baoulé cattle of Burkina Faso. *PLOS ONE* 16, e0255089. doi:10.1371/journal.pone.0255089

Yu, Q., Tian, X., Sun, C., Shao, L., Li, X., and Dai, R. (2019). Comparative transcriptomics to reveal muscle-specific molecular differences in the early postmortem of Chinese Jinjiang yellow cattle. *Food Chem.* 301, 125262. doi:10.1016/j.foodchem.2019.125262

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

Zhang, Y., Hu, Y., Wang, X., Jiang, Q., Zhao, H., Wang, J., et al. (2020). Population structure, and selection signatures underlying high-altitude adaptation inferred from genome-wide copy number variations in Chinese indigenous cattle. *Front. Genet.* 10, 1404. doi:10.3389/fgene.2019.01404

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fvets.2022.993773

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Laing Chunnian,
Lanzhou Institute of Husbandry and Pharmaceutical Sciences (CAAS), China
Shanhe Wang,
Yangzhou University, China

*CORRESPONDENCE
Yanhong Zhao
947196432@163.com

†These authors share first authorship

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

RECEIVED 14 July 2022
ACCEPTED 12 September 2022
PUBLISHED 29 September 2022

CITATION
Liu Z, Liu Z, Mu Q, Zhao M, Cai T, Xie Y, Zhao C, Qin Q, Zhang C, Xu X, Lan M, Zhang Y, Su R, Wang Z, Wang R, Wang Z, Li J and Zhao Y (2022) Identification of key pathways and genes that regulate cashmere development in cashmere goats mediated by exogenous melatonin. *Front. Vet. Sci.* 9:993773. doi: 10.3389/fvets.2022.993773

Identification of key pathways and genes that regulate cashmere development in cashmere goats mediated by exogenous melatonin

Zhihong Liu[1,2,3,4†], Zhichen Liu[1†], Qing Mu[1], Meng Zhao[5], Ting Cai[6], Yuchun Xie[7], Cun Zhao[1], Qing Qin[1], Chongyan Zhang[1], Xiaolong Xu[1], Mingxi Lan[1], Yanjun Zhang[1], Rui Su[1], Zhiying Wang[1], Ruijun Wang[1], Zhixin Wang[1], Jinquan Li[1,2,3,4] and Yanhong Zhao[1,2,3,4*]

[1]College of Animal Science, Inner Mongolia Agricultural University, Hohhot, China, [2]Key Laboratory of Animal Genetics, Breeding and Reproduction, Inner Mongolia Agricultural University, Hohhot, China, [3]Key Laboratory of Mutton Sheep Genetics and Breeding, Ministry of Agriculture, Hohhot, China, [4]Goat Genetics and Breeding Engineering Technology Research Center, Inner Mongolia Agricultural University, Hohhot, China, [5]Inner Mongolia Autonomous Region Agriculture and Animal Husbandry Technology Extension Center, Hohhot, China, [6]Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot, China, [7]Hebei Normal University of Science and Technology, Qinhuangdao, China

The growth of secondary hair follicles in cashmere goats follows a seasonal cycle. Melatonin can regulate the cycle of cashmere growth. In this study, melatonin was implanted into live cashmere goats. After skin samples were collected, transcriptome sequencing and histological section observation were performed, and weighted gene co-expression network analysis (WGCNA) was used to identify key genes and establish an interaction network. A total of 14 co-expression modules were defined by WGCNA, and combined with previous analysis results, it was found that the blue module was related to the cycle of cashmere growth after melatonin implantation. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis showed that the first initiation of exogenous melatonin-mediated cashmere development was related mainly to the signaling pathway regulating stem cell pluripotency and to the Hippo, TGF-beta and MAPK signaling pathways. *Via* combined differential gene expression analyses, 6 hub genes were identified: *PDGFRA*, *WNT5A*, *PPP2R1A*, *BMPR2*, *BMPR1A*, and *SMAD1*. This study provides a foundation for further research on the mechanism by which melatonin regulates cashmere growth.

KEYWORDS

melatonin, WGCNA, hair follicle cycle, Inner Mongolia cashmere goats, PDGF

Introduction

The Inner Mongolia cashmere goat is an economically important animal with diverse uses and is an important economic breeding stock in China. Cashmere is one of the finest and lightest fibers produced by animals; it is noted for its outstanding properties and rarity and is much more expensive than wool of the same fineness. It is used especially in the production of high-end textiles. Cashmere is produced by hair follicles, an accessory organ of the skin (1, 2). Cashmere goat hair can be divided into two types of fibers: unmedullated cashmere and well-medullated coarse hair. Hair follicles can also be divided into primary hair follicles and secondary hair follicles (3). In Inner Mongolia cashmere goats, cashmere growth exhibits a seasonal pattern and a periodic change influenced by the natural photoperiod (2). Each year, cashmere starts growing in late summer, stops growing in the middle of winter and falls off naturally in spring. These periodic changes occur through three phases of cellular activity: growth, regression, and repose (4). Secondary hair follicles grow from April to November, regress from December to January, and rest from February to March.

The growth cycle of cashmere is affected by many factors, such as sunshine duration, melatonin (MT; N-acetyl-5-methoxy-tryptamine), nutrition, genetics, endocrine status, etc. (5). MT is an endogenous hormone produced by the pineal gland. Its secretion has a distinct circadian rhythm, with inhibition during the day and active secretion at night. MT has long been recognized as an effective regulatory neuroendocrine substance associated with hair growth and the hair cycle, dependent on photoperiods, seasonal rhythms and environmental factors. Notably, MT performs unique biological functions in regulating hair growth in goats (6). Studies have shown that MT is a key mediator between the photoperiod and cashmere growth and that the level of circulating MT directly affects cashmere growth. Exogenous MT has been confirmed to have a positive effect on cashmere growth. In the non-growing period, the use of exogenous MT can stimulate the growth of cashmere but can also lead to early shedding of cashmere followed by another cashmere growth period (7). Considering the above effects of MT, we sought to determine whether the economic benefit of cashmere goats can be increased by artificially implanting MT to change the reception of light signals. Our previous study showed that continuous MT implantation promoted the entry of cashmere into hair follicles 2 months in advance of the seasonal cycle and promoted the development of secondary hair follicles. MT significantly affected the expression of *WNT10B*, *β-catenin* and other proteins in the skin tissue of Inner Mongolia cashmere goats. With the rapid development of high-throughput sequencing technology, through the analysis of differentially expressed genes (DEGs) and functional enrichment analysis, scientists have identified many regulatory factors and signaling pathways that may influence the hair follicle cycle: the WNT signaling pathway, fibroblast growth factor (FGF) family, bone morphogenetic protein (BMP) family, Sonic hedgehog (SHH) signaling pathway, transforming growth factor (TGF) family, Notch signaling pathway, etc. (8–10).

Bioinformatics methods are increasingly used for the exploration and analysis of target genes or proteins (11). Weighted gene co-expression network analysis (WGCNA) is a method used to study gene modules related to traits at the whole-transcriptome level. This mathematical method is used to transform expression data into scale-free distribution network information and then perform clustering on the target gene set and identify the gene modules of interest. In contrast to traditional gene clustering approaches, WGCNA clustering can reveal biological significance and has been widely used in gene mining, gene function prediction and other aspects. Regarding WGCNA in cashmere goats, a research group found through WGCNA that *WNT10A* is a key gene in the early stage of development and maturation of fetal skin hair follicles in Inner Mongolia cashmere goats (12). To date, WGCNA has been used in many studies on diseases, especially tumors (13–16), but the use of WGCNA to evaluate the effects of MT implantation in Inner Mongolia cashmere goats has not been reported. Therefore, to explore the genes and pathways through which MT promotes early growth of cashmere, we conducted this experiment.

Materials and methods

Sample collection

The research samples were collected from the Inner Mongolia White Cashmere Goat Breeding Farm in Etoke Banner, Ordos City, Inner Mongolia. 61-year-old ewes with similar body weights, with no relation to each other and in good growth condition were selected and divided into two groups. One group was implanted with MT, and the other group was used as a control. In the implanted group, MT was implanted subcutaneously behind the ears of the cashmere goats at a dose of 2 mg/kg BW every 2 months. Skin samples of 1 cm × 1 cm were collected from the scapula side of the ewes at the beginning of each month, and the skin samples were stored in an ultralow temperature refrigerator at −80°C for later use.

Preparation of frozen sections and HE staining

Spare tissue samples were removed from −80°C, thawed in equilibrium at 4°C, placed in 4% paraformaldehyde solution

and fixed at 4°C overnight. The next day, the fixed tissues were washed three times with PBS for 3 min each time. After the filter paper dried, the tissues were placed in 30% sucrose solution and dehydrated overnight at 4°C until the tissues sank to the bottom. The excess water was absorbed with filter paper, the appropriate angle was adjusted according to the sectioning direction, and the tissue was placed on the precooled section substrate; then, the tissues were embedded with OTC embedding agent before cryosectioning. The microtome was prechilled, and the tissues were sectioned after being completely frozen. The sections were stained using a hematoxylin eosin (HE) staining kit (G1120, Solarbio) according to the instructions.

Total RNA extraction from skin

Total RNA was extracted from the skin of three cashmere goats using an RNAiso Plus Kit (TRIzol method). The purity and integrity of total RNA were measured using a sterile UV–VIS spectrophotometer and an Agilent 2,100 bioanalyzer, respectively, and the three RNA samples were then mixed. Total RNA was stored in a freezer at −80°C.

Construction of the sequencing library

The cDNA library for transcriptome sequencing was constructed according to the operating instructions of an Illumina TruSeqTM RNA Sample Preparation Kit. Total RNA extracted from three cashmere goats per group in each month was mixed together in equal amounts. The mRNA was purified with oligo(dT) magnetic beads and sheared into 100–400 bp mRNA fragments. Double-stranded cDNA was synthesized using the mRNA fragments as templates, an exonuclease and a polymerase. To obtain the sequencing library, the double-stranded cDNA was phosphorylated, ligated to sequencing adapters and subjected to poly(A) tailing prior to PCR amplification.

Transcriptome sequencing and splicing

The Illumina HiSeqTM 2000 sequencing platform was used for paired-end sequencing of cDNA. A 2 × 100 bp sequencing protocol was used for sequencing by Shanghai Meggie Biopharmaceutical Co., Ltd. After sequencing, referenced and unreferenced genomes were used to compare the reads, and assembly was conducted with the Trinity and Velvet methods. The filtered reads were aligned to the goat (*Capra hircus*) genome. The FPKM value of each gene was calculated to estimate the gene expression level.

Construction of the gene co-expression network

Before WGCNA, the selected gene set was screened and filtered to remove low-quality genes or samples with unstable effects on the results to improve the accuracy of network construction. A weighted gene co-expression network was constructed using WGCNA (V. 147) package in R software (17). The Pearson linear correlation coefficient of each pair of genes in the gene set was calculated to construct the correlation coefficient matrix ($S_{ij} = |cor(X_i, X_j)|$). To remove the influence of low correlation coefficients, an appropriate threshold (β) was selected for exponential weighting of the correlation coefficient matrix and construction of the adjacency matrix. Then, the adjacency matrix was transformed into the topological overlap matrix ($TOM_{ij} = \frac{l_{ij}+a_{ij}}{\min\{k_i,k_j\}+1-a_{ij}} (u \neq i,j)$). The dissimilarity matrix D ($d_{ij} = 1 - TOM_{ij}$) was then constructed, and hierarchical clustering was carried out for this matrix. The dynamic cutting tool in WGCNA was used to prune the cluster tree, and the gene modules were preliminarily divided. Based on the similarity of the eigengenes of the original modules, these modules were combined to form the final modules.

Identification of the module and hub genes related to cashmere development

As shown in our previous publication (18), the cycle of cashmere growth could be divided into three distinct periods: a growth period (March–September), a regression period (September–December) and a resting period (December–March). March was considered to be the beginning of the cycle. In this study, among all 14 co-expression modules, the expression pattern of the blue module from 1 to 12 months was consistent with the growth cycle of cashmere. The expression pattern of the blue module gradually increased beginning in the growth period and showed a downward trend in the regression period and resting period. Moreover, after melatonin implantation, cashmere goats exhibited cashmere growth in May, which was also consistent with the blue module expression pattern. Therefore, the blue module was selected for relevant analysis in this study. The protein–protein interaction (PPI) network of the differentially expressed genes in the blue module was predicted using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING, version 11.5) (19), and a comprehensive score >0.4 was considered to indicate a statistically significant difference. The results of STRING analysis were imported into Cytoscape (20) to visualize the interaction network. The maximal clique centrality (MCC) metric was used to identify hub genes in the PPI network. MCC performed better than CytoHubba (21) in 11 other available

methods (22), and the genes with the top 15 MCC scores were considered hub genes.

Enrichment analysis of key module genes

To explore the biological functions of the differentially expressed genes, Gene Ontology (GO)/Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed on the differentially expressed genes between the implanted group and the control group in May. GO enrichment analysis was performed using the online tool g: Profiler (23). The GO categories included biological process (BP), cellular component (CC), and molecular function (MF). Significantly enriched GO terms were selected according to $P < 0.05$. The goat database was selected for KEGG pathway enrichment analysis with KOBAS 3.0 (24), and $P < 0.05$ was set as the screening criterion for significant enrichment.

Combined analysis of hub genes and DEGs

To identify the genes expressed in May that play a role in activating cashmere growth in advance of the seasonal cycle, we screened for the genes differentially expressed in May between the implantation group and the control group. The thresholds for differential gene expression were | log2-fold change | > 1 and $P < 0.05$. Then, Venny 2.1.0 was used to determine the overlap among the three sets of genes.

Quantitative real-time PCR

In this study, an Applied Biosystems QuantStudio 3 real-time fluorescence quantitative PCR system was used to reverse transcribe the extracted total RNA into cDNA according to the instructions of a PrimeScript^{TM} RT Master Mix Kit (RR036A, TAKARA). Then, based on the cDNA sequences of the goat *PDGFRA*, *WNT5A*, *BMPR2,* and *BMPR1A* genes published by NCBI, specific primers were designed with Primer 5.0 (Table 3) and synthesized by Shanghai Bioengineering Co., Ltd. Finally, a TB Green "Premix Ex Taq" II Kit (RR820A, TAKARA) was used for qRT–PCR. Six samples were tested for each month, and three technical replicates were performed.

Statistical analysis

The $2^{-\triangle\triangle}CT$ method was used for the analysis of all qRT–PCR data, and SPSS software (version 22.0) was used for statistical analysis. Values are expressed as the means ± standard deviations. A significance level of 0.05 was used.

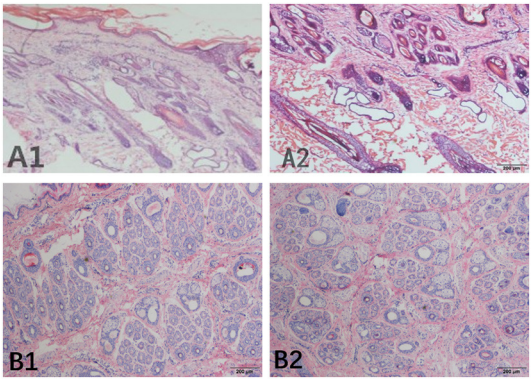


FIGURE 1
Skin tissue section of Inner Mongolia cashmere goats collected in May. **(A)** Represents the longitudinal section, **(B)** represents the transverse section, and numbers 1 and 2 represent the implanted group and the control group, respectively.

Results

Histological examination of goat skin

The S:P (ratio of secondary hair follicles to primary hair follicles) was used as an indicator of the number of hair follicles in the skin of the cashmere goats. In this study, we performed histological staining of sample sections from the control group and the implanted group in May (Figure 1), and six fields of view were observed for each sebaceous gland sample collected at different depths. The number of hair follicles in each counting area was calculated, and the results showed that the hair follicles of the control group were located deeply in the epidermal layer, while the hair follicles of the implanted group were located near the epidermal layer. The ratio of secondary hair follicles to primary hair follicles in the implanted group was significantly higher than that in the control group ($P < 0.01$).

Quality of sequencing data and splicing results

The sequencing data for the 24 samples are shown in Table 1. The Trinity and Velvet assembly methods were used for comparison. When the Velvet assembly method was used, the effect was best when the kmer (bp) value was 57; 323,630 transcripts were obtained, and 67.27% of the transcripts were successfully mapped to the genome. With the Trinity method, 511,110 transcripts were obtained, and 90.06% of the transcripts were successfully mapped to the goat genome. Comparison of the genome mapping results indicated that the Trinity method performed better.

TABLE 1 Results of high-quality raw data.

Sample name	Number of reads	Total length (bp)	Percentage (%)
LZH-1	11901322	2088403976	86.87
LZH-2	11799342	2093444071	87.83
LZH-3	13129996	2254870449	85.02
LZH-4	12177094	2048134833	83.27
LZH-5	12151639	2130929746	86.81
LZH-6	11984821	2104532776	86.93
LZH-7	15948330	2710213186	84.13
LZH-8	11420123	2016743197	87.42
LZH-9	12990697	2298341553	87.59
LZH-10	12851175	2269169813	87.41
LZH-11	15025883	2513464188	82.81
LZH-12	13299627	2221588767	82.69
LZH-A	21997027	3736758621	84.10
LZH-B	17376784	2939183676	83.73
LZH-C	19578410	3294127588	83.29
LZH-D	13946831	2350977514	83.45
LZH-E	10365856	1623233029	77.52
LZH-F	13348217	2233152344	82.82
LZH-G	18423325	3104576795	83.42
LZH-H	13380685	2283690673	84.49
LZH-I	11977439	2114244179	87.39
LZH-J	12751089	2262360512	87.83
LZH-K	12775926	2249682980	87.17
LZH-M	13890435	2452180440	87.39

TABLE 2 Number of genes in the co-expression module.

Module color	Gene numbers	Module color	Gene numbers
Blue	4530	Midnightblue	76
Cyan	84	Pink	212
Green	291	Purple	119
Greenyellow	117	Salmon	103
Lightcyan	61	Tan	105
Lightgreen	41	Turquoise	4574
Magenta	402	Yellow	301

Construction of the weighted gene co-expression network

RNA-seq was performed on 12-month transcriptome data for goat skin samples from the implantation group and the control group. Average linkage hierarchical clustering based on the computed topological overlap was then used to identify genes with very similar expression patterns in each module.

TABLE 3 Primer sequences specific for down-producing goat *PDGFRA*, *WNT5A*, *BMPR2*, *BMPR1A* and *GAPDH* and PCR product size.

Gene name	Sequence of primer	Products size
PDGFRA	F: GAGTGCCATTGAGACAGGTTCCAG	143 bp
	R: CCGAATCTGCCAGTTACAGGAAGC	
WNT5A	F: TCCAAGATTCAAAGAGCCTGCTTCC	145 bp
	R: AGCGTCCACTCCTGCCTACTTC	
BMPR2	F: CAACACCACTCAGTCCACCTCATTC	139 bp
	R: CCTTGTTTGCGGTCTCCTGTCAG	
BMPR1A	F: CCAGGTCAGCAATACAGCAACTCC	112 bp
	R: CTCCACACAGAAATCTACGGCACTC	
GAPDH	F: TGGACATCGTTGCCATCAATGACC	100 bp
	R: TTGACTGTGCCGTGGAACTTGC	

A total of 14 co-expression modules were identified in the WGCNA network and represented by different colors. The resulting blue, cyan, green, green–yellow, light cyan, light green, magenta, midnight blue, pink, purple, salmon, tan, turquoise, and yellow modules contained 4,530, 84, 291, 117, 61, 41, 402, 76, 212, 119, 103, 105, 4,574, and 301 genes, respectively (a total of 11,016 genes) (Table 2). In the gene clustering results, Dynamic Tree Cut represents each group of genes, and Merged Dynamic represents the merged tree obtained by the dynamic cutting method (Figure 2A). In the topological overlap heatmaps, more topological overlap was observed within the modules than across the modules (Figure 2B). Based on correlation coefficient analysis, we clustered the modules associated with cashmere growth and development (positive or negative) (Figure 2C).

Identification of the module and hub genes related to cashmere development

As shown in Figure 3, among all 14 co-expression modules, the expression pattern of the blue module from 1 to 12 months was consistent with the growth cycle of cashmere, as previously shown by our group. Moreover, after the implantation of melatonin, cashmere goats exhibited cashmere growth in May, which was also consistent with the blue module expression pattern. Therefore, the blue module was selected for relevant analysis in this study. We constructed PPI networks for the genes in the blue module. The genes with the top 15 MCC scores in CytoHubba were identified as hub genes.

Enrichment analysis of key module genes

We further performed GO/KEGG enrichment analyses with 316 differentially expressed genes upregulated in the

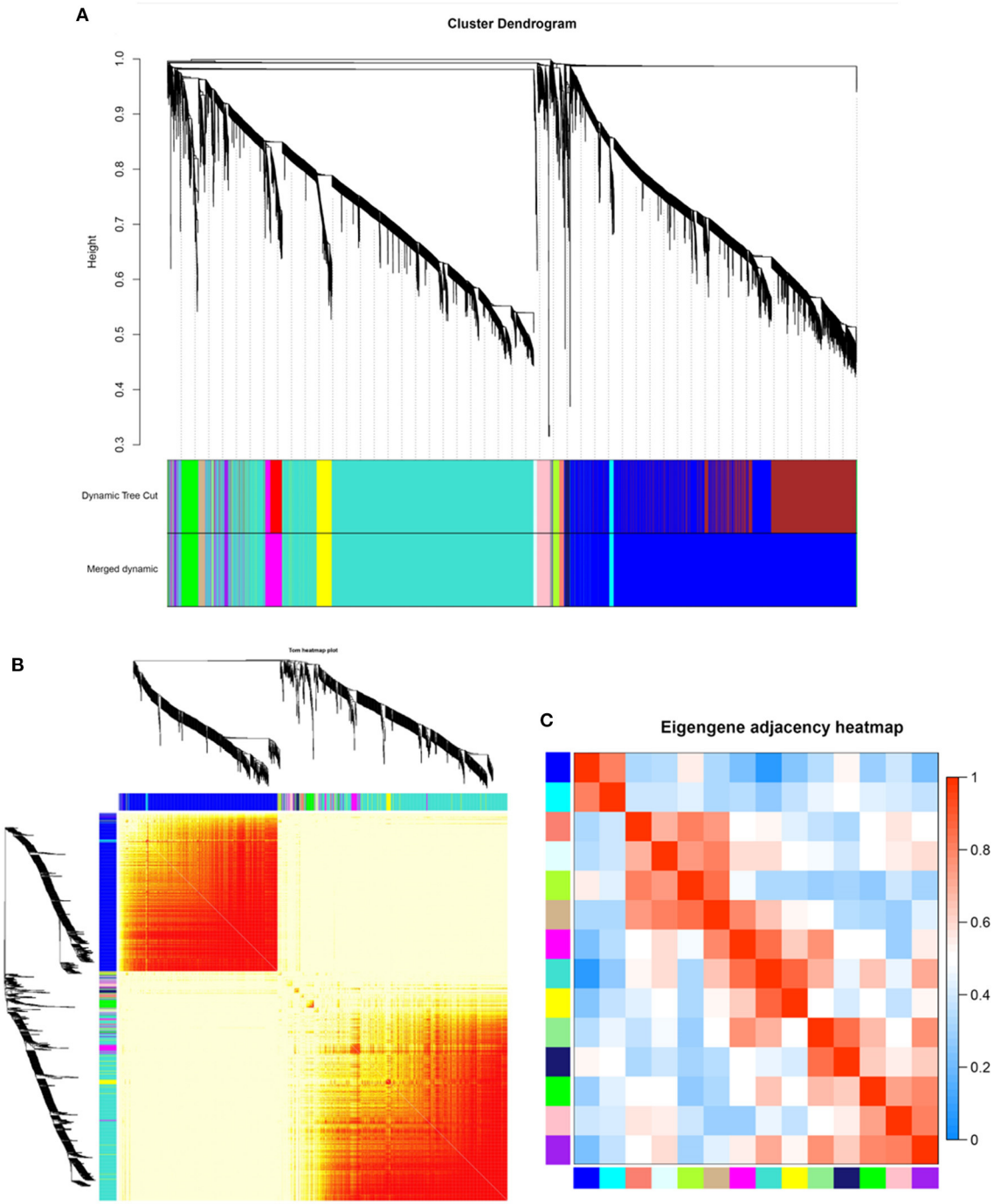


FIGURE 2
Identification of co-expression modules by WGCNA. **(A)** Dendrogram of gene clustering. **(B)** Network heatmap of module genes. Each tree represents a module, and each branch represents a gene. The darker the color of each dot is, the stronger the connectivity between the two genes in the row and column. **(C)** Correlation heatmap of each module; colors indicate correlations between modules. The closer the color is red, the higher the correlation, and the closer the color is blue, the lower the correlation.

blue module in May. According to GO classification statistics, 148 terms were grouped into three GO categories: cellular component, molecular function, and biological process. Among the three GO categories, the most significantly enriched were cellular protein modification process, desmosome and protein binding. Notably, the cellular components were

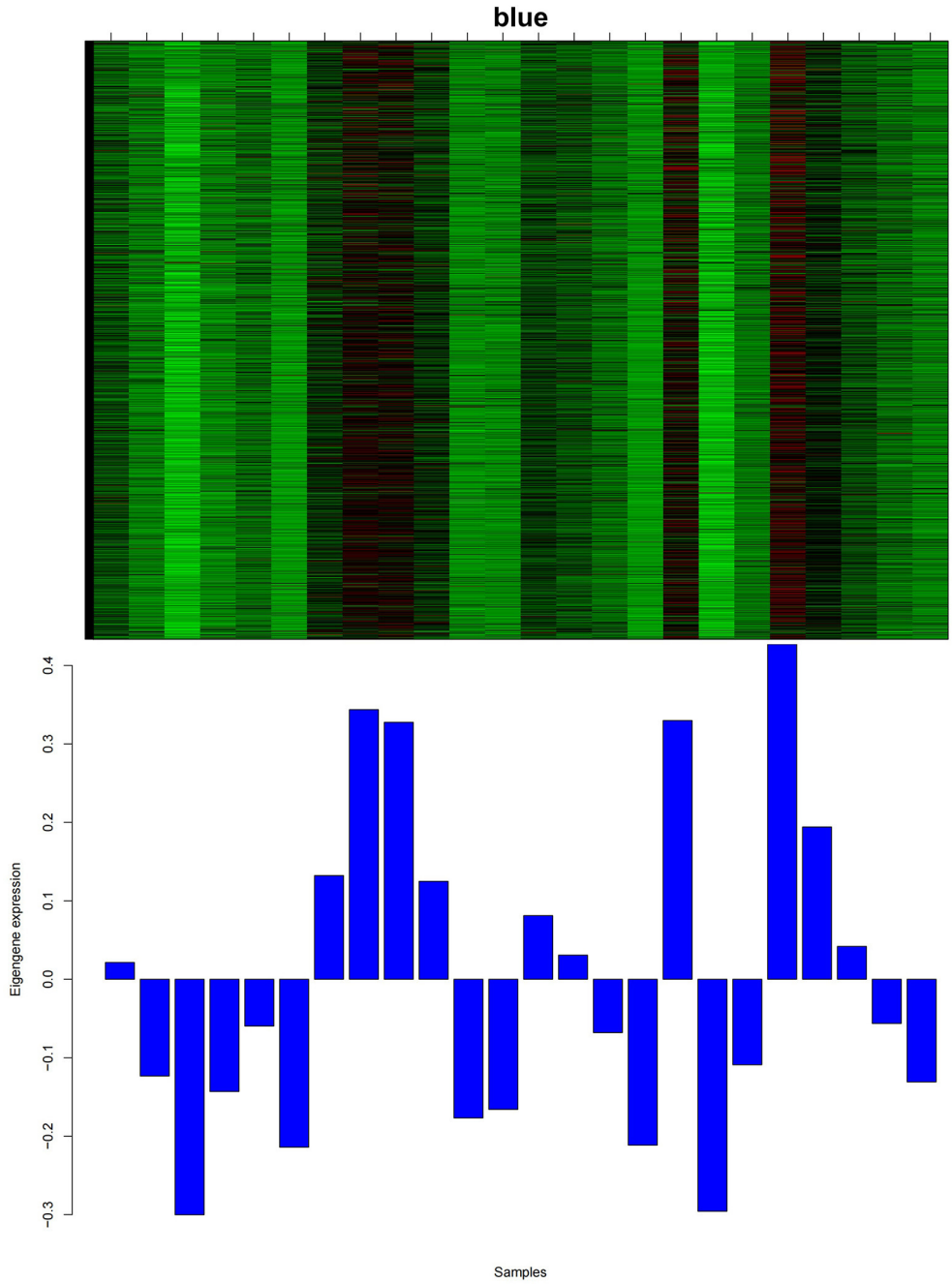


FIGURE 3
Heatmap of the gene expression pattern of the blue module. The abscissa is the name of the sample. The top figure shows the heatmap of gene expression in different samples of the module, and the bottom figure shows the expression pattern of the feature value of the module in different samples.

significantly enriched in desmosomes, which are the most important intercellular adhesion junctions, directly adhering to desmosomal cadherins on adjacent keratinocytes to form the epidermal layer.

KEGG pathway analysis (Figure 4) showed that in May, the differentially expressed genes between the control group and the implanted group were mainly enriched in the signaling pathway that regulates the pluripotency of stem cells and the Hippo, TGF-beta and MAPK signaling pathways. Interestingly, we found significant enrichment of MT signaling pathways. According to the KEGG enrichment analysis results, six genes were enriched in three or more signaling pathways related to hair follicle

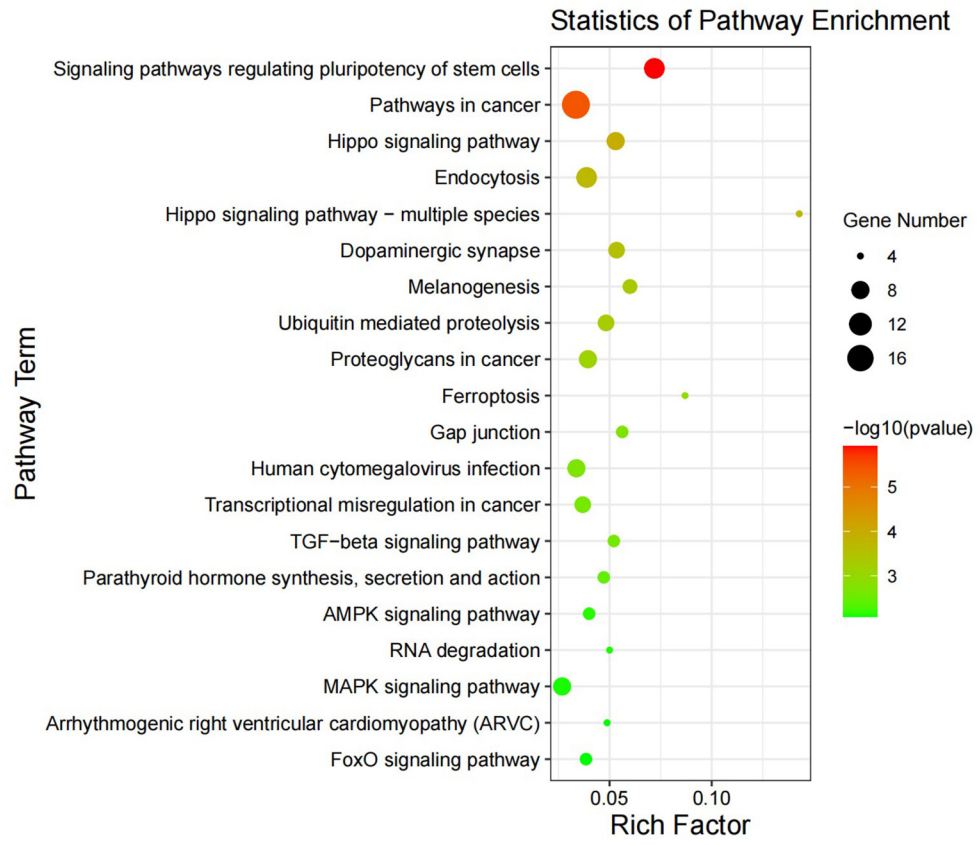


FIGURE 4
Differential gene KEGG enrichment diagram. The X-axis represents the ratio of enriched differential genes to the total number of genes in the term, and the Y-axis represents the enriched function/pathway. The colors represent the degree of enrichment significance; the closer the color is red, the more significant the results are. The circle size represents the number of genes enriched.

development: *PDGFRA*, *WNT5A*, *BMPR1A*, *BMPR2*, *PPP2R1A*, and *SMAD1* (Figure 5).

Combined analysis of hub genes and differentially expressed genes

We compared the 316 genes differentially upregulated in the control group and the implanted group in May with the important genes and hub genes in the PPI. *PDGFRA* was common to all three gene sets (Figure 6). This result suggests that *PDGFRA* is a key gene in the early growth of cashmere after implantation of MT.

The relative expression of core genes was determined by qRT–PCR

In this study, total RNA was extracted from skin samples of the control group and the implanted group at each of the 12 months. A NanoDrop 2000 UV spectrophotometer was used to determine whether the OD260/280 values were between 1.8 and 2.0. Four genes were randomly selected to evaluate their relative expression with respect to *GAPDH* as the internal reference gene to verify the accuracy of the sequencing results. The qRT–PCR results were basically consistent with the transcriptome data. The expression levels of *PDGFRA*, *WNT5A*, *BMPR2*, and *BMPR1A* in the control group showed a gradually increasing trend from May to September but first decreased and then increased in the implanted group. The *PDGFRA* level differed significantly ($P < 0.05$) between the control group and the implanted group in May (Figure 7). This finding is consistent with our analysis results: *PDGFRA* is a key gene in the promotion of premature cashmere production by MT implantation.

Discussion

The main objective of this study was to screen for and identify the genes related to early hair production after MT

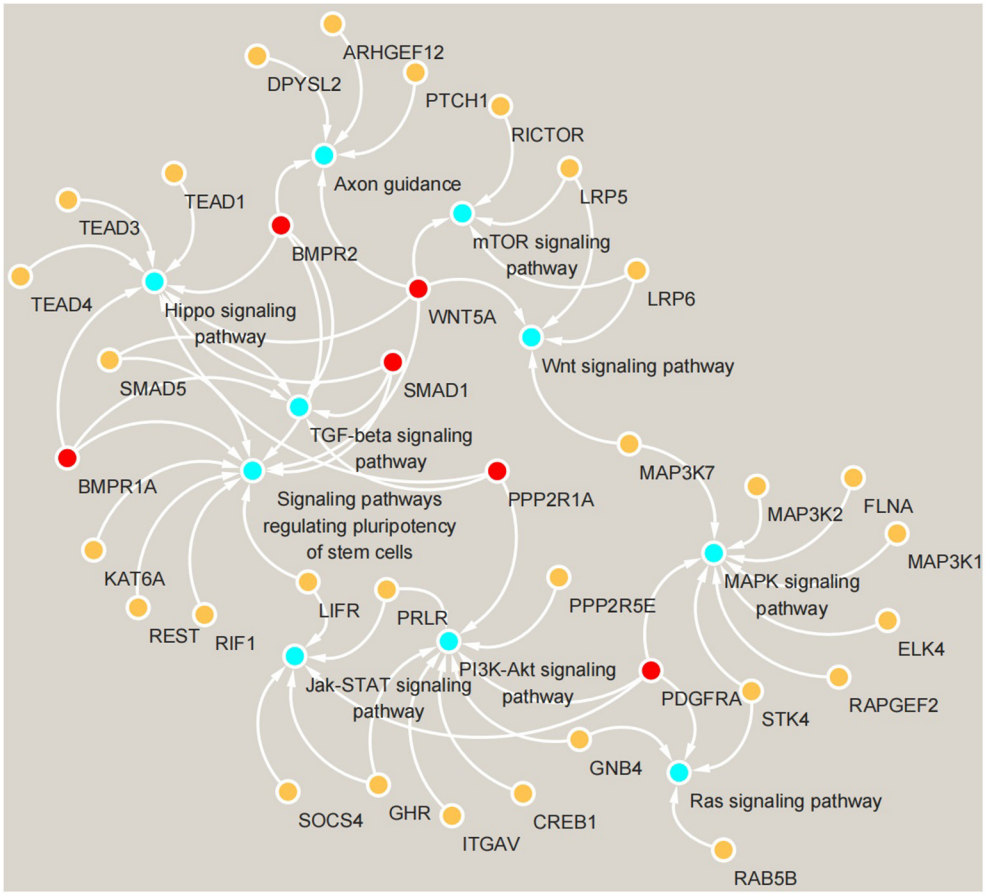


FIGURE 5
Gene enrichment map of KEGG signaling pathways. Blue circles represent signaling pathways, yellow circles represent genes enriched in pathways, and red circles represent genes enriched in more than three pathways simultaneously.

implantation in Inner Mongolia cashmere goats and to study the role of these genes in the growth and development of skin hair follicles in Inner Mongolia cashmere goats. According to the longitudinal analysis of sections of samples from the implanted group and control group that were collected in May, the hair follicles of the control group were located deeply in the epidermal layer, while the hair follicles of the implanted group were located near the epidermal layer, indicating that the hair follicles in the melatonin-implanted group emerged sooner than those in the control group. Through the analysis of cross-sections, it was found that the ratio of secondary hair follicles to primary hair follicles in the implanted group was significantly higher than that in the control group, indicating that the growth of secondary hair follicles in the implanted group was earlier than that in the control group. On the basis of transcriptome sequencing data, we identified 14 co-expression modules using WGCNA, selected the blue module as having an expression that aligned with the growth cycle of cashmere for analysis, and constructed a predicted protein interaction network. Key genes within the blue module were identified using the STRING database and CytoHubba. GO and KEGG analyses were performed with the differentially expressed genes. The upregulated differentially expressed genes were compared with hub genes and the genes enriched in 3 or more signaling pathways in the implanted group and the control group in May, providing important insights into the transcriptional mechanism of MT-mediated skin hair follicle development in Inner Mongolia cashmere goats.

Genes play a decisive role in the process of cashmere growth. Previous studies have found that implanting MT can promote the early growth of cashmere, but the specific genes and pathways involved in the early growth of cashmere remain to be studied. Therefore, it is highly important to study the genes related to cashmere growth regulation after MT implantation. The pathway enrichment analysis in this study identified six genes enriched in more than three pathways: *PDGFRA*, *WNT5A*, *PPP2R1A*, *BMPR2*, *BMPR1A*, and *SMAD1*. The signaling pathway regulating the pluripotency of stem

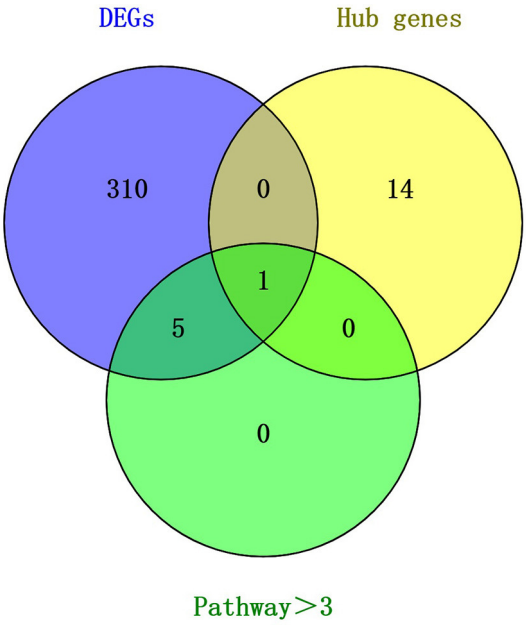


FIGURE 6
Venn diagram of key candidate genes.

cells was the pathway most significantly enriched with the differentially expressed genes. This result may indicate that stem cells play an important role in hair follicle development during the hair cycle, but no relevant reports have been made. Moreover, differentially expressed genes were also enriched in the Hippo, TGF-beta and MAPK signaling pathways, which are all related to cashmere growth (25–29).

In our previous studies, we found that the development of hair follicles is regulated by hair follicle stem cells, and through single-cell analyses, we found that hair follicle stem cells can affect the development cycle of cashmere. The epidermis regenerates in a steady state through balanced proliferation of basal cells and exfoliation of keratinized squamous cells on its surface. Hair follicles regenerate through a complex cyclical process. The cycle begins with the activation of stem cells, which is followed by the proliferation and differentiation of their progeny (30). *WNT5A,* which is critical to the initiation of hair growth, was enriched in the signaling pathway pluripotency of stem cells and the WNT signaling pathway (31–33). Wnt-secreted proteins can stimulate diverse signaling pathways in cells and play important roles in cell proliferation, differentiation and migration as growth regulators. Many studies have confirmed that the classical Wnt/β-catenin signaling pathway plays a key role in hair follicle growth (34). *WNT5A* can activate both classical and non-classical WNT signaling pathways and act through different WNT pathways after binding to different receptors in different cell types. The binding of *WNT5A* to the *FZD-4* receptor activates the β-catenin pathway to promote cell proliferation (35). Therefore, we hypothesized that *WNT5A* may affect the initiation of cashmere growth through activation of the β-catenin pathway.

A main role of the platelet-derived growth factor (PDGF) gene is to promote DNA synthesis in cells. PDGF participates in the regulation of hair follicle growth and development mainly through PDGF signaling (including through the *PDGFA* gene and PDGF receptors A and B). Studies have shown that PDGF signaling is related to hair follicle morphogenesis in early hair follicle development and that inhibition of the expression of PDGF or its receptor *PDGFRA* or *PDGFRB* in growing and developing hair follicles prevents hair follicle maturation (36). In this study, *PDGFRA* was identified as a key gene in four pathways, and many genes in the blue module are related to *PDGFRA*, such as *IGF1*, *ITGAV*, *FRS2*, and *CDK6*. Most of these genes are related to cell growth, and some researchers have found that the treatment of acute wounds in rats with *IGF1*-overexpressing fibroblasts can significantly improve the rate of wound healing (37). *ITGAV* has been proven to be a key gene for the initiation of goat cashmere growth (38). Cyclin-dependent kinase 6 (*CDK6*) is an important regulator of the cell cycle and plays a role to similar *CDK4* as a mediator of keratinocyte proliferation (39). The docking protein *FRS2* is involved in the transmission of extracellular signals from fibroblast growth factor (FGF) or nerve growth factor (NGF) receptors to the Ras/MAPK signaling pathway (40). The MAPK and PI3K-AKT signaling pathways have been confirmed to play an important role in the initial stage of cashmere development (25, 41). The Jak-stat signaling pathway may be involved in the transition from the resting to the growth phase in cashmere growth (42), and the Ras signaling pathway is also believed to be involved in the development and regeneration of hair follicles (43). Early research by our group through simple correlation analysis of the expression levels of cashmere growth-related genes showed that each signaling pathway was interconnected and interwoven into a network to regulate the periodicity of cashmere growth. MT implantation can promote the expression of *PDGFA* and its receptor, *PDGFRA*, as well as their binding. *PDGFA*, *PDGFRA,* and *NTRK3* play a synergistic role in the periodic growth of cashmere. During robust growth, the PDGF signaling pathway plays an important role, consistent with the results of this study.

At the beginning of the growth phase, follicular progenitor cells in the epidermis induce mesenchymal condensation to form the dermis, which then produces proliferating stromal cells. These epidermal cells further differentiate into hair stem cells; in addition, some of these stem cells form the basal layer of the epidermis. The BMP family is the largest subfamily of growth factors in the TGF-beta superfamily and controls hair follicle morphogenesis at many different stages (44). BMP signaling occurs through a complex of type I (BMPRI) and type II (BMPRII) BMP receptors. *BMPR1A* is one of the three type I BMP

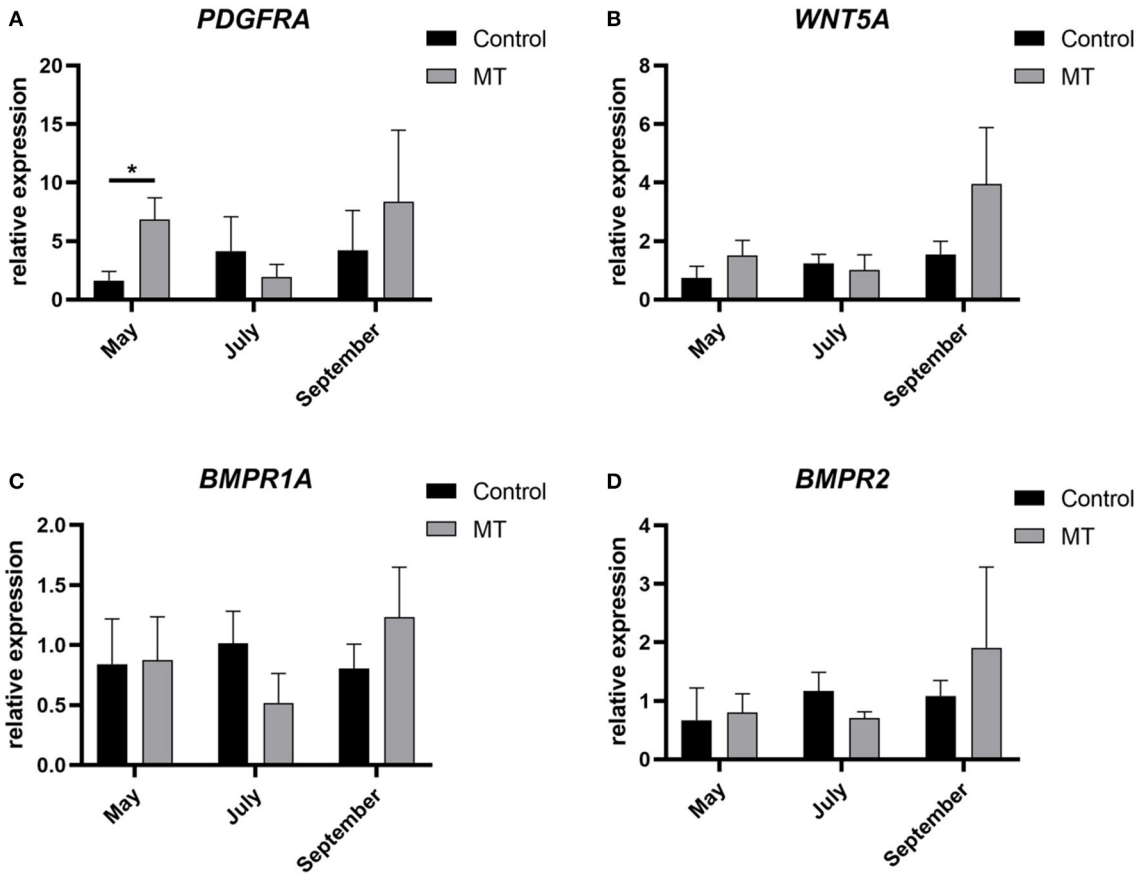


FIGURE 7
Sequencing results validated by fluorescent qRT-PCR. **(A–D)** The relative expression of four genes. The abscissa represents month, the ordinate represents relative expression, *Represents a significant difference between groups ($P < 0.05$), and the error bar represents the SD.

receptors. Deletion of the *BMPR1A* gene in mouse skin hair follicles resulted in reduced hair follicle differentiation and interfollicular epidermal cell growth in mouse fetal skin and in a significant reduction in the number of hair follicles in postnatal mice, which were hairless in the affected area (45). These results suggest that *BMPR1A* signaling is critical not only for hair follicle differentiation during development but also for epidermal cell proliferation or differentiation during the renewal of the adult cashmere cycle. In addition, TGF-beta signaling plays a key role in various aspects of hair follicle development and circulation (44), and some experimental studies have demonstrated that *BMPR2* is essential for normal hair development and maintenance. Decreased *BMPR2* expression results in premature end of the growth phase. This finding also confirms the positive effect of *BMPR2* on the initiation and maintenance of cashmere in this study.

Mammalian hair follicle development begins in the embryonic stage, and studies have shown that Wnt/β-catenin, TGF-beta/Smad and other signaling pathways are involved in the early initiation, differentiation and development of the hair follicle tissue structure (46). In skin hair follicle tissue, *SMAD1* mainly plays a role in promoting tissue differentiation and hair follicle formation; in contrast, *SMAD1* is rarely expressed in the base of mature hair follicles. Therefore, we speculated that *SMAD1* may be a key promoter gene of cashmere development. *PPP2R1A* is the scaffold subunit of protein phosphatase 2A (PP2A), one of the four major serine/threonine protein phosphatases. The mechanism of this gene in hair follicle development in cashmere goats has not been reported, but based on the signaling pathways and increased expression level of *PPP2R1A*, we speculate that *PPP2R1A* may be related to cell proliferation and differentiation, a possibility that will constitute our future research direction. In general, in this study, by using WGCNA, we identified 6 key genes that may mediate the effects of MT and lead to early cashmere production in cashmere goats, providing a foundation for subsequent research.

Conclusion

In this study, WGCNA was used to investigate the mechanism by which exogenous melatonin regulates the cycle of cashmere developmental. Histological sections of goat skin that were collected in May showed that the hair follicles of the control group were located deeply in the epidermal layer, while the hair follicles of the implanted group were located near the epidermal layer. The ratio of secondary hair follicles to primary hair follicles in the implanted group was significantly higher than that in the control group. A total of 14 co-expression modules were identified, and the blue module was found to be correlated with the cashmere growth cycle after MT implantation. Through analysis of the hub genes, the genes in the blue module and the genes differentially expressed between the implanted group and the control group in May, *PDGFRA* was identified as the key gene through which MT regulates cashmere growth, providing a theoretical basis for further improving cashmere yield and economic benefit.

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/, SRP145408.

Ethics statement

In this study, skin samples were collected in accordance with the International Guiding Principles for Biomedical Research involving animals, and the experiment was approved by the Special Committee on Scientific Research and Academic Ethics of Inner Mongolia Agricultural University, which is responsible for the approval of Biomedical Research Ethics of Inner Mongolia Agricultural University [Approval No. (2020) 056]. No specific permissions were required for these activities, and no endangered or protected species were involved (47).

Author contributions

ZhihL and YZhao contributed to the concept and design of the study. ZhicL, MZ, TC, ZhixW, CZhao, CZhan, QQ, XX, and ML collected the samples and extracted the RNA. QM performed the sections and staining. ZhicL, YX, and QQ conducted the data analysis. YZhan, RS, ZhiyW, RW, and JL provided technical support. ZhicL and ZhihL wrote the first draft of the manuscript. All the authors contributed to the revision of the manuscript and read and approved the submitted version.

Funding

This work was supported by the National Key R&D Program of China (2021YFD200901); the National Natural Science Foundation of China (32160772, 32060742, and 31860628); the Major Science and Technology Projects of Inner Mongolia Autonomous Region (2020ZD0004); the Key Technology Project of Inner Mongolia Autonomous Region (2020GG0030 and 2020GG0031).

Acknowledgments

The authors thank the staff of Inner Mongolia White Cashmere Goat Breeding Farm in Etoke Banner, Ordos City, Inner Mongolia for their help.

Conflict of interest

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

Publisher's note

References

1. Wang L, Zhang Y, Zhao M, Wang R, Su R, Li J. Snp discovery from transcriptome of cashmere goat skin. *Asian-Aust J Anim Sci.* (2015) 28:1235–43. doi: 10.5713/ajas.15.0172

2. Duan C, Xu J, Sun C, Jia Z, Zhang W. Effects of melatonin implantation on cashmere yield, fibre characteristics, duration of cashmere growth as well as growth and reproductive performance of inner mongolian cashmere goats. *J Anim Sci Biotechnol.* (2015) 6:22. doi: 10.1186/s40104-015-0023-2

3. Schmidt-Ullrich R, Paus R. Molecular principles of hair follicle induction and morphogenesis. *BioEssays News Rev Mol Cell Develop Biol.* (2005) 27:247–61. doi: 10.1002/bies.20184

4. Zhang QL Li JP, Chen Y, Chang Q, Li YM, Yao JY, et al. Growth and viability of Liaoning cashmere goat hair follicles during the annual hair follicle cycle. *Genet Mol Res GMR.* (2014) 13:4433–43. doi: 10.4238/2014.June.16.2

5. Liu J, Mu Q, Liu Z, Wang Y, Liu J, Wu Z, et al. Melatonin regulates the periodic growth of cashmere by upregulating the expression of Wnt10b and B-catenin in Inner Mongolia cashmere goats. *Front Genet.* (2021) 12:665834. doi: 10.3389/fgene.2021.665834

6. Fu S, Zhao H, Zheng Z, Li J, Zhang W. Melatonin regulating the expression of Mirnas involved in hair follicle cycle of cashmere goats skin. *Yi chuan = Hereditas.* (2014) 36:1235–42. doi: 10.3724/sp.J.1005.2014.1235

7. Foldes A, Hoskinson RM, Baker P, McDonald BJ, Maxwell CA, Restall BJ. Effect of Immunization against melatonin on seasonal fleece growth in feral goats. *J Pineal Res.* (1992) 13:85–94. doi: 10.1111/j.1600-079X.1992.tb00059.x

8. Geng R, Yuan C, Chen Y. Exploring differentially expressed genes by Rna-Seq in cashmere goat (Capra Hircus) skin during hair follicle development and cycling. *PLoS ONE.* (2013) 8:e62704. doi: 10.1371/journal.pone.0062704

9. Wang X, Cai B, Zhou J, Zhu H, Niu Y, Ma B, et al. Disruption of Fgf5 in cashmere goats using Crispr/Cas9 results in more secondary hair follicles and longer fibers. *PLoS ONE.* (2016) 11:e0164640. doi: 10.1371/journal.pone.0164640

10. Jin M, Wang J, Chu MX, Piao J, Piao JA, Zhao FQ. The study on biological function of keratin 26, a novel member of Liaoning cashmere goat keratin gene family. *PLoS ONE.* (2016) 11:e0168015. doi: 10.1371/journal.pone.0168015

11. Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, et al. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst Biol.* (2008) 2:95. doi: 10.1186/1752-0509-2-95

12. Wu Z, Hai E, Di Z, Ma R, Shang F, Wang Y, et al. Using Wgcna (weighted gene co-expression network analysis) to identify the hub genes of skin hair follicle development in fetus stage of Inner Mongolia cashmere goat. *PLoS ONE.* (2020) 15:e0243507. doi: 10.1371/journal.pone.0243507

13. Tian Z, He W, Tang J, Liao X, Yang Q, Wu Y, et al. Identification of important modules and biomarkers in breast cancer based on Wgcna. *Onco Targets Ther.* (2020) 13:6805–17. doi: 10.2147/OTT.S258439

14. Chen S, Yang D, Lei C, Li Y, Sun X, Chen M, et al. Identification of crucial genes in abdominal aortic aneurysm by Wgcna. *PeerJ.* (2019) 7:e7873. doi: 10.7717/peerj.7873

15. Nangraj AS, Selvaraj G, Kaliamurthi S, Kaushik AC, Cho WC, Wei DQ. Integrated Ppi- and Wgcna-retrieval of hub gene signatures shared between Barrett's esophagus and esophageal adenocarcinoma. *Front Pharmacol.* (2020) 11:881. doi: 10.3389/fphar.2020.00881

16. Liang W, Sun F, Zhao Y, Shan L, Lou H. Identification of susceptibility modules and genes for cardiovascular disease in diabetic patients using Wgcna analysis. *J Diabetes Res.* (2020) 2020:4178639. doi: 10.1155/2020/4178639

17. Langfelder P, Horvath S. Wgcna: An R package for weighted correlation network analysis. *BMC Bioinform.* (2008) 9:559. doi: 10.1186/1471-2105-9-559

18. Yang F, Liu Z, Zhao M, Mu Q, Che T, Xie Y, et al. Skin transcriptome reveals the periodic changes in genes underlying cashmere (ground hair) follicle transition in cashmere goats. *BMC Genom.* (2020) 21:392. doi: 10.1186/s12864-020-06779-5

19. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucl Acids Res.* (2021) 49:D605–d12. doi: 10.1093/nar/gkaa1074

20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303

21. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. Cytohubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* (2014) 8(Suppl 4):S11. doi: 10.1186/1752-0509-8-S4-S11

22. Bai Q, Liu H, Guo H, Lin H, Song X, Jin Y, et al. Identification of hub genes associated with development and microenvironment of hepatocellular carcinoma by weighted gene co-expression network analysis and differential gene expression analysis. *Front Genet.* (2020) 11:615308. doi: 10.3389/fgene.2020.615308

23. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucl Acids Res.* (2019) 47:W191–w8. doi: 10.1093/nar/gkz369

24. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, et al. Kobas-I: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucl Acids Res.* (2021) 49:W317–w25. doi: 10.1093/nar/gkab447

25. Pearson G, Robinson F, Beers Gibson T, Xu BE, Karandikar M, Berman K, et al. Mitogen-activated protein (map) kinase pathways: regulation and physiological functions. *Endocr Rev.* (2001) 22:153–83. doi: 10.1210/edrv.22.2.0428

26. Schlegelmilch K, Mohseni M, Kirak O, Pruszak J, Rodriguez JR, Zhou D, et al. Yap1 acts downstream of α-catenin to control epidermal proliferation. *Cell.* (2011) 144:782–95. doi: 10.1016/j.cell.2011.02.031

27. Glover JD, Wells KL, Matthäus F, Painter KJ, Ho W, Riddell J, et al. Hierarchical patterning modes orchestrate hair follicle morphogenesis. *PLoS Biol.* (2017) 15:e2002117. doi: 10.1371/journal.pbio.2002117

28. Wang J, Chen Z, Xiao Z, Weng Y, Yang M, Yang L, et al. Estrogen induces ido expression *via* Tgf-? in chorionic villi and decidua during early stages of pregnancy. *Int J Mol Med.* (2020) 46:1186–96. doi: 10.3892/ijmm.2020.4658

29. Lu Q, Gao Y, Fan Z, Xiao X, Chen Y, Si Y, et al. Amphiregulin promotes hair regeneration of skin-derived precursors *via* the Pi3k and Mapk pathways. *Cell Prolif.* (2021) 54:e13106. doi: 10.1111/cpr.13106

30. Plikus MV, Gay DL, Treffeisen E, Wang A, Supapannachart RJ, Cotsarelis G. Epithelial stem cells and implications for wound repair. *Semin Cell Dev Biol.* (2012) 23:946–53. doi: 10.1016/j.semcdb.2012.10.001

31. Andl T, Reddy ST, Gaddapara T, Millar SE. Wnt signals are required for the initiation of hair follicle development. *Dev Cell.* (2002) 2:643–53. doi: 10.1016/S1534-5807(02)00167-3

32. Feng Y, Gun S. Melatonin supplement induced the hair follicle development in offspring rex rabbits. *J Anim Physiol Anim Nutr.* (2021) 105:167–74. doi: 10.1111/jpn.13417

33. Zhang Y, Tomann P, Andl T, Gallant NM, Huelsken J, Jerchow B, et al. Reciprocal requirements for Eda/Edar/Nf-Kappab and Wnt/Beta-catenin signaling pathways in hair follicle induction. *Dev Cell.* (2009) 17:49–61. doi: 10.1016/j.devcel.2009.05.011

34. Slominski AT, Semak I, Fischer TW, Kim TK, Kleszczyński K, Hardeland R, et al. Metabolism of melatonin in the skin: why is it important? *Exp Dermatol.* (2017) 26:563–8. doi: 10.1111/exd.13208

35. Igota S, Tosa M, Murakami M, Egawa S, Shimizu H, Hyakusoku H, et al. Identification and characterization of Wnt signaling pathway in keloid pathogenesis. *Int J Med Sci.* (2013) 10:344–54. doi: 10.7150/ijms.5349

36. Ahn SY Pi LQ, Hwang ST, Lee WS. Effect of Igf-I on hair growth is related to the anti-apoptotic effect of Igf-I and up-regulation of Pdgf-a and Pdgf-B. *Ann Dermatol.* (2012) 24:26–31. doi: 10.5021/ad.2012.24.1.26

37. Fresno Vara JA, Casado E, de Castro J, Cejas P, Belda-Iniesta C, González-Barón M. Pi3k/Akt signalling pathway and cancer. *Cancer Treat Rev.* (2004) 30:193–204. doi: 10.1016/j.ctrv.2003.07.007

38. Bhat B, Yaseen M, Singh A, Ahmad SM, Ganai NA. Identification of potential key genes and pathways associated with the pashmina fiber initiation using Rna-Seq and integrated bioinformatics analysis. *Sci Rep.* (2021) 11:1766. doi: 10.1038/s41598-021-81471-6

39. Wang X, Sistrunk C, Rodriguez-Puebla ML. Unexpected reduction of skin tumorigenesis on expression of cyclin-dependent kinase 6 in mouse epidermis. *Am J Pathol.* (2011) 178:345–54. doi: 10.1016/j.ajpath.2010.11.032

40. Ong SH, Guy GR, Hadari YR, Laks S, Gotoh N, Schlessinger J, et al. Frs2 proteins recruit intracellular signaling pathways by binding to diverse targets on fibroblast growth factor and nerve growth factor receptors. *Mol Cell Biol.* (2000) 20:979–89. doi: 10.1128/MCB.20.3.979-989.2000

41. Chen Y, Fan Z, Wang X, Mo M, Zeng SB, Xu RH, et al. Pi3k/Akt signaling pathway is essential for *de novo* hair follicle regeneration. *Stem Cell Res Ther.* (2020) 11:144. doi: 10.1186/s13287-020-01650-6

42. Wang E, Harel S, Christiano AM. Jak-Stat signaling jump starts the hair cycle. *J Invest Dermatol.* (2016) 136:2131–2. doi: 10.1016/j.jid.2016.08.029

43. Doma E, Rupp C, Baccarini M. Egfr-Ras-Raf signaling in epidermal stem cells: roles in hair follicle development, regeneration, tissue remodeling and epidermal cancers. *Int J Mol Sci.* (2013) 14:19361–84. doi: 10.3390/ijms141019361

44. Li AG, Koster MI, Wang XJ. Roles of Tgfbeta signaling in epidermal/appendage development. *Cytokine Growth Factor Rev.* (2003) 14:99–111. doi: 10.1016/S1359-6101(03)00005-4

45. Yuhki M, Yamada M, Kawano M, Iwasato T, Itohara S, Yoshida H, et al. Bmpr1a signaling is necessary for hair follicle cycling and hair shaft differentiation in mice. *Development.* (2004) 131:1825–33. doi: 10.1242/dev.01079

46. Hu X, Zhang X, Liu Z, Li S, Zheng X, Nie Y, et al. Exploration of key regulators driving primary feather follicle induction in goose skin. *Gene.* (2020) 731:144338. doi: 10.1016/j.gene.2020.144338

47. Sibley BA, Etnier JL, Le Masurier GC. Effects of an acute bout of exercise on cognitive aspects of Stroop performance. *J Sport Exerc Psychol.* (2006) 28:285.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fgene.2022.1022681

OPEN ACCESS

EDITED BY
Zexi Cai,
Aarhus University, Denmark

REVIEWED BY
George R. Wiggans,
Council on Dairy Cattle Breeding, United States
María Muñoz,
INIA-CSIC, Spain

*CORRESPONDENCE
Marzieh Heidaritabar,
heidarit@ualberta.ca

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

RECEIVED 18 August 2022
ACCEPTED 22 September 2022
PUBLISHED 11 October 2022

CITATION
Heidaritabar M, Huisman A, Krivushin K, Stothard P, Dervishi E, Charagu P, Bink MCAM and Plastow GS (2022), Imputation to whole-genome sequence and its use in genome-wide association studies for pork colour traits in crossbred and purebred pigs. *Front. Genet.* 13:1022681. doi: 10.3389/fgene.2022.1022681

Imputation to whole-genome sequence and its use in genome-wide association studies for pork colour traits in crossbred and purebred pigs

Marzieh Heidaritabar[1]*, Abe Huisman[2], Kirill Krivushin[1], Paul Stothard[1], Elda Dervishi[1], Patrick Charagu[3], Marco C. A. M. Bink[2] and Graham S. Plastow[1]

[1]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada, [2]Hendrix Genetics Research, Boxmeer, Netherlands, [3]Hendrix Genetics, Business Unit Swine, Regina, SK, Canada

Imputed whole-genome sequence (WGS) has been proposed to improve genome-wide association studies (GWAS), since all causative mutations responsible for phenotypic variation are expected to be present in the data. This approach was applied on a large number of purebred (PB) and crossbred (CB) pigs for 18 pork color traits to evaluate the impact of using imputed WGS relative to medium-density marker panels. The traits included Minolta A*, B*, and L* for fat (FCOL), quadriceps femoris muscle (QFCOL), thawed loin muscle (TMCOL), fresh ham gluteus medius (GMCOL), ham iliopsoas muscle (ICOL), and longissimus dorsi muscle on the fresh loin (FMCOL). Sequence variants were imputed from a medium-density marker panel (61K for CBs and 50K for PBs) in all genotyped pigs using BeagleV5.0. We obtained high imputation accuracy (average of 0.97 for PBs and 0.91 for CBs). GWAS were conducted for three datasets: 954 CBs and 891 PBs, and the combined CBs and PBs. For most traits, no significant associations were detected, regardless of panel density or population type. However, quantitative trait loci (QTL) regions were only found for a few traits including TMCOL Minolta A* and GMCOL Minolta B* (CBs), FMCOL Minolta B*, FMCOL Minolta L*, and ICOL Minolta B* (PBs) and FMCOL Minolta A*, FMCOL Minolta B*, GMCOL Minolta B*, and ICOL Minolta B* (Combined dataset). More QTL regions were identified with WGS (n = 58) relative to medium-density marker panels (n = 22). Most of the QTL were linked to previously reported QTLs or candidate genes that have been previously reported to be associated with meat quality, pH and pork color; e.g., *VIL1, PRKAG3, TTLL4,* and *SLC11A1, USP37*. *CTDSP1* gene on SSC15 has not been previously associated with meat color traits in pigs. The findings suggest any added value of WGS was only for detecting novel QTL regions when the sample size is sufficiently large as with the Combined dataset in this study. The percentage of phenotypic variance explained by the most significant SNPs also increased with WGS compared with medium-density panels. The results provide additional insights into identification of a number of candidate regions and genes for pork color traits in different pig populations.

KEYWORDS

pork color traits, crossbred pigs, purebred pigs, imputed whole-genome sequence, GWAS, QTL regions

Introduction

Pork color is a key effective indicator for meat quality traits and freshness, since it has been shown that there is a moderate to high association between some of the pork color traits and other meat quality traits such as drip loss (e.g., genetic correlation of 0.55 ± 0.24 and 0.42 ± 0.19 between drip loss and Loin Minolta L* and Loin Minolta A*, respectively) and ultimate pH (genetic correlation of -0.37 ± 0.16) (Miar et al., 2014). Therefore, pork color is an important factor which influences consumer decisions for purchasing pork (Glitsch, 2000). Moreover, Miar et al. (2014) showed that pork color traits had moderate to high heritability, ranging from 0.10 ± 0.05 to 0.38 ± 0.06 (average = 0.25). This shows that in addition to the environmental factors, genetic factors control pork color. Hence, genetic improvement of pork color, which is economically important for the swine industry, is possible in pig breeding programs. Understanding the complex genetic mechanisms underlying pork color traits, which can be done by detection of new genomic regions associated with these traits, is a necessity for the genetic improvement of these traits. Genome-wide association studies (GWAS) using a part of the data[1] in the current study, have identified several regions associated with pork color traits (Zhang et al., 2015; Yang et al., 2017). Five genomic regions on *Sus scrofa* chromosomes (SSC) 1, 5, 9, 15, 16 and the X chromosome were identified (Zhang et al., 2015). The region on SSC15 spanning 133–134 Mb explained 3.51%–17.06% of genetic variance for five measurements of pH and color (Zhang et al., 2020). Yang et al. (2017) identified 20 genomic regions associated with 18 pork color traits. Three of the genomic regions (on 32–36 Mbp of SSC1 for quadriceps femoris muscle (QFCOL) Minolta A*, 130–134 Mbp of SSC15 for three traits (QFCOL Minolta A* and B*, thawed loin muscle (TMCOL) Minolta B*), and a region on SSC16) associated with three pork color traits identified by Zhang et al. (2015) were also detected by Yang et al. (2017).

To date, most GWAS have used medium-to high-density marker panels to detect the genomic regions associated with carcass and meat quality traits. Use of whole-genome sequence (WGS) is expected to improve identification of associated regions (in terms of both distinct and extended candidate regions and identifying novel genomic regions), because most of the causative variants are expected to be within WGS. The causative SNPs have low MAF (rare variants) and their variance is expected to be captured using WGS. According to simulations, using WGS data for GWAS, the precision of mapping for rare variants increased considerably, which supports the efficiency of WGS in detecting and fine-mapping of low frequency variants simultaneously (Wu et al., 2017). Identification of such variants can increase the utility of genomic selection (GS) for traits such as pork quality by increasing selection accuracy, particularly in multi-population or across population genetic evaluations as used in most commercial pig production which uses crossbreeding and ultimately accelerating genetic gain (Kizilkaya et al., 2014). A disadvantage of using WGS for genetic analyses is the cost of sequencing. Even though the costs of WGS are decreasing, it is still too expensive to sequence at sufficient coverage the thousands of animals required for accurately detecting the genomic regions associated with complex quantitative traits such as pork color traits. A promising alternative is to sequence influential founder animals with the highest genetic contribution to the target population (so-called "reference population") and to impute the sequence of the remaining animals from low density genotypes (so-called "target population") (Meuwissen and Goddard, 2010a; b). A cost-effective sequencing alternative to obtain large-scale genomic information is low-pass whole-genome sequence in which 1x coverage or less of a target genome is sequenced. Low-pass sequencing combined with imputation has been proposed as an alternative to genotyping arrays for improving both quantitative trait loci (QTL) detection through a GWAS (Li et al., 2021) and genomic prediction accuracy (Snelling et al., 2020).

Through imputation, based on WGS, the missing variants in the target population can be predicted by use of linkage and segregation analysis. Imputation accuracy is an important factor for more accurate detection of associated regions. Bouwman et al. (2018) assessed the accuracy of imputation from a 70K SNP panel to WGS, from a 660K SNP panel to WGS, and a two-step procedure from 70K to 660K to WGS, using three imputation programs including Beagle 4.1 (Browning et al., 2018), Minimac3 (Das et al., 2016), and FImpute (Sargolzaei et al., 2014). They showed that using a small reference set of 168 sequenced pigs, imputation from 660K was more accurate than imputation from 70K directly to WGS. Their two-step procedure (from 70K to 660K to WGS) resulted in the lowest imputation accuracy. They also showed that Beagle 4.1 outperformed Minimac3. In their study, FImpute performed less well compared with other imputation programs. A useful strategy to reduce imputation error rate is to filter SNPs based on their imputation accuracy prior to analysis.

The use of imputed WGS has been more common in GWAS for pig traits in recent years (Li et al., 2017; Yan et al., 2017; Yan

1 Only our CB animals with 61K single nucleotide polymorphisms (SNP) panel were used by Zhang et al. (2015) and Yang et al. (2017). Both authors used the same pork color traits as in the present study.

et al., 2018; Van den Berg et al., 2019; Wu et al., 2019; Yang et al., 2021). Van den Berg et al. (2019) showed that using the imputed WGS, the detected QTLs increased with increasing SNP density. They found that compared to 80K and 660K genotypes, using imputed WGS led to the identification of 48.9 and 64.4% more QTL regions, for Landrace and Large White pigs, respectively, and the most significant SNPs in the QTL regions explained a higher proportion of phenotypic variance. Wu et al. (2019) detected 113 and 18 SNPs associated with farrowing interval of different parities in two pig populations using imputed sequence variants. Also, Yan et al. (2017) identified a QTL associated with lumbar number in Sutai pigs using imputed WGS. Nevertheless, to the best of our knowledge, few studies have investigated using imputed WGS for GWAS for meat and carcass quality traits in both purebred and crossbred pigs.

We performed GWAS for 18 meat color traits including Minolta L*, A*, and B* for fat (FCOL), quadriceps femoris muscle (QFCOL), thawed loin muscle (TMCOL), fresh ham gluteus medius (GMCOL), ham iliopsoas muscle (ICOL), and longissimus dorsi muscle on the fresh loin (FMCOL). Analyses were conducted for two datasets: 954[2] crossbred pigs (CBs) and 891[3] purebred pigs (PBs). Sequence variants, called across the 60 sequenced pigs, were imputed from a medium-density marker panel (61K for CBs and 50K for PBs) in all genotyped pigs. We applied a single marker association analysis and accounted for polygenic effects through the genomic relationship matrix for each dataset. The main objectives of the study were therefore: 1) to assess the imputation accuracy from 61K CBs and 50K PBs to WGS using a small reference population of 60 sequenced pigs, and 2) to investigate whether the use of WGS detected more associated regions compared with lower density SNP panels. Furthermore, we performed GWAS on combined CBs and PBs to assess whether or not the power of GWAS increased with increasing population size. Finally, we identified potential candidate genes within the associated regions and described the biological roles of the most interesting regions through functional analyses.

Materials and methods

Data

Phenotypes

This study was performed using the data provided by Hendrix Genetics (Hypor Inc., *Regina*, SK, Canada). Phenotypes of 18 meat color traits were available for 1,037 commercial crossbred pigs (524 female and 513 male CBs, mostly from three-way cross between Duroc boars and Landrace-Yorkshire sows, and 76 were from F1 hybrid sows (Landrace-Yorkshire)). Also, phenotypes of 15 meat color traits were available for 891 purebred Duroc females. The list of the 18 meat color traits and their abbreviations are given in Table 1. Number of individuals in the pedigree were 4,420 and 5,260 for CBs and PBs, respectively. The combined PB and CB pedigree was made by defining the genetic groups in ASReml program V4.0 (Gilmour et al., 2015), as the animals from PBs and CBs were considered to belong to different genetic groups. Thus, the combined pedigree comprised 6,419 individuals including the genetic groups. The details on how the pork color phenotypes were measured in the six locations of the pork have been described in Yang et al. (2017).

Genotypes

Of the 1,037 crossbred individuals that had phenotypic records, 941-954 individuals (depending on the trait) had both phenotypes and genotypes with a custom 61K (61,565 SNPs)[4] Illumina SNP panel (Table 2). Genotyping of CBs was performed by Delta Genomics (Edmonton, AB, Canada) using Illumina PorcineSNP60 V2 Genotyping Beadchip according to the Illumina Infinium Assay (Illumina, Inc., San Diego, CA, United States). Of the 891 purebred Duroc females that had phenotypic records, 873-891 individuals (depending on the trait) had both phenotypes and genotypes with a custom 50K (50,703 SNPs) Illumina SNP panel (Table 2). Genotyping of purebred pigs was performed by Neogen Corporation - GeneSeek operations (Lincoln, Nebraska, NE, United States). Based on the "proportion of genetic diversity" approach (Druet et al., 2014), 60 Duroc boars were identified as key ancestors of the PB population and DNA of these boars was used for sequencing. Moreover, for 17 of the 891 purebred Duroc sows, genotypes from the 660K SNP panel including 659,692 SNPs were available. We used this set of individuals to assess potential increase in imputation accuracy when using a two-step procedure. The two-step procedure was from 50K to 660K to WGS, while in the one-step approach the imputation was conducted from 50K to WGS directly.

Collection of deoxyribonucleic acid samples, deoxyribonucleic acid extraction, library preparation and next-generation sequencing

Genomic DNA extraction from blood and tissue was carried out using the Qiagen DNeasy extraction protocol (Qiagen,

2 The number of phenotypes vary per trait, ranging from 941 to 954 (See *Methods*).

3 The number of phenotypes vary per trait, ranging from 873 to 891.

4 When we mentioned a medium-density SNP panel throughout the manuscript, we meant 61K and 50K SNP panels.

TABLE 1 List of pork color traits and their abbreviations.

Number	Trait abbreviation	Trait description
1	FCOLA	Fat Minolta A*
2	FCOLB	Fat Minolta B*
3	FCOLL	Fat Minolta L*
4	QFCOLA	Quadriceps femoris muscle Minolta A*
5	QFCOLB	Quadriceps femoris muscle Minolta B*
6	QFCOLL	Quadriceps femoris muscle Minolta L*
7	FMCOLA	Fresh marbling color A* - longissimus dorsi
8	FMCOLB	Fresh marbling color B* - longissimus dorsi
9	FMCOLL	Fresh marbling color L* - longissimus dorsi
10	TMCOLA	Thawed loin muscle Minolta A*
11	TMCOLB	Thawed loin muscle Minolta B*
12	TMCOLL	Thawed loin muscle Minolta L*
13	GMCOLA	Ham gluteus medius Minolta A*
14	GMCOLB	Ham gluteus medius Minolta B*
15	GMCOLL	Ham gluteus medius Minolta L*
16	ICOLA	Ham iliopsoas Minolta A*
17	ICOLB	Ham iliopsoas Minolta B*
18	ICOLL	Ham iliopsoas Minolta L*

TABLE 2 The descriptive statistics for 18 pork color traits: number of animals per trait (N), means, SD, minimum (Min.), and maximum (Max.) values for different datasets (CB, PB, and Combined dataset).

	CB					PB					Combined dataset				
Trait	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
FCOLA	941[a]	3.71	1.13	0.70	7.90	873	2.84	1.34	-0.40	7.30	1844	3.38	1.52	−0.40	19.20
FCOLB	953	18.31	1.48	11.70	24.40	885	10.84	1.75	6.00	17.20	1844	14.70	4.08	3.80	24.40
FCOLL	953	75.29	1.63	66.60	79.80	891	78.96	2.40	64.00	84.60	1844	77.06	2.74	64.00	84.60
QFCOLA	953	4.82	1.60	0.70	11.30	881	2.39	1.37	-1.00	6.70	1844	3.68	1.96	-1.00	13.40
QFCOLB	953	13.61	1.57	9.60	18.70	885	8.27	1.35	4.70	12.10	1844	11.04	3.04	4.70	18.70
QFCOLL	953	49.42	3.46	39.10	62.10	881	53.23	3.37	42.10	65.50	1844	51.27	4.04	36.50	68.90
FMCOLA	953	6.07	1.47	2.00	11.48	891	4.58	1.12	1.08	8.75	1844	5.35	1.51	1.08	11.48
FMCOLB	953	14.91	1.69	10.38	21.90	891	9.38	1.24	5.95	13.90	1844	12.24	3.14	5.95	21.90
FMCOLL	953	48.46	2.64	39.88	60.50	891	48.15	2.53	41.43	55.73	1844	48.31	2.59	39.88	60.50
TMCOLA	950	7.65	1.19	3.39	11.39	-	-	-	-	-	-	-	-	-	-
TMCOLB	950	2.70	1.29	-1.54	7.48	-	-	-	-	-	-	-	-	-	-
TMCOLL	950	44.26	3.11	31.99	55.88	-	-	-	-	-	-	-	-	-	-
GMCOLA	953	6.74	1.20	2.40	10.70	891	5.46	1.27	1.20	9.60	1844	6.12	1.39	1.20	10.70
GMCOLB	953	13.63	1.11	9.60	17.30	891	8.91	1.13	5.40	12.70	1844	11.35	2.61	5.40	17.30
GMCOLL	953	45.31	2.45	38.00	54.20	891	47.50	2.65	39.30	57.20	1844	46.37	2.77	38.00	57.20
ICOLA	953	19.30	1.73	12.00	24.10	891	15.97	2.17	8.80	23.00	1844	17.69	2.59	1.60	24.10
ICOLB	953	13.61	1.57	9.60	18.70	891	11.33	1.50	5.50	15.80	1844	12.51	1.91	5.50	18.70
ICOLL	953	42.54	2.86	35.10	51.80	891	44.04	3.08	34.70	55.60	1844	43.26	3.06	34.70	55.60

[a]The total number of PBs, with both phenotypes and genotypes were 891. However, for these traits, there were extreme phenotypic records which were removed in the analyses to check if the GWAS, results would improve. Due to little changes in GWAS, results for PBs, those removed individuals were added to the analyses of combined CBs, and PBs. CB, crossbred; PB, purebred; N, number of animals; SD, standard deviation; Min, minimum; Max, maximum.

Mississauga, ON) by Delta genomics. Extracted DNA was quantified using the Qubit dsDNA HS Assay (Life Technologies, Burlington, ON). 100ng to 1ug of gDNA was sheared using the Covaris S2 focused sonicator (Covaris Inc.) to achieve a fragment size ranging from 300 to 400bp. Sheared DNA fragments were used for library preparation according to respective library preparation protocol that were compatible with Illumina next generation sequencing platform. Quality check and library preparations were done by NEOGEN Canada (Edmonton, AB, Canada). Sequencing was done by McGill University and Génome Québec Innovation Centre (Montréal, Québec, Canada). Libraries were normalized and pooled and then denatured in 0.05N NaOH and neutralized using HT1 buffer. ExAMP was added to the mix following the manufacturer's instructions. The pool was loaded at 200pM on a Illumina cBot and the flowcell was run on a HiSeq X for 2 × 151 cycles (paired-end mode). A phiX library was used as a control and mixed with libraries at 1% level. The Illumina HiSeq Control Software was HCS HD 3.4.0.38, and the real-time analysis program was RTA v. 2.7.7. Program bcl2fastq2 v2.20 was then used to de-multiplex samples and generate fastq reads.

Sequence depth, read trimming, alignment, and variant calling

Sequence reads trimming and adapter clipping was performed using Trimmomatic algorithm 0.38 (Bolger et al., 2014). The average sequence coverage was computed using *depth* in VCFTOOLS (Danecek et al., 2011) and was 21.75 across the 60 sequenced animals (Supplementary Table S1). Sequence reads alignment was conducted using the current pig reference genome (*Sus scrofa* 11.1 (https://uswest.ensembl.org/Sus_scrofa/Info/Index), www.ensembl.org/biomart/martview) with BWA *mem* (BWA 0.7.17) using the default parameters (Li and Durbin, 2009). The alignment SAM files were converted to BAM format using Samtools-0.1.19 (Li et al., 2009). Next, BAM files were sorted and indexed by Samtools 1.8 (Li et al., 2009). Potential PCR duplicates were removed by tool *MarkDuplicates* from Picard v2.18.2 (http://broadinstitute.github.io/picard/). Variants (SNPs and insertion-deletions (INDELs)) were called using GenomeAnalysisToolKit-3.8-1-0 (GATK) (McKenna et al., 2010). Tool *HaplotypeCaller* was used for variant calling. Default parameter settings of *HaplotypeCaller* were used for variant calling, except for the following parameters: minimum base quality required to consider a base for calling equal to 20 and the minimum phred-scaled confidence threshold for variant calling equal to 20. Base quality recalibration was performed according to GATK best practices guidelines using tools *BaseRecalibrator* and *PrintReads* (McKenna et al., 2010; van der Auwera et al., 2013). Finally, BAM files were pooled for variant calling. In the 60 Duroc males, the total numbers of SNPs and INDELs called were more than 19 and more than five million, respectively.

Quality control of called sequenced variants

During variant calling, the variants were filtered using parameters recommended by GATK Best Practices (DePristo et al., 2011). Some other filters were applied to choose sequencing variants for GWAS analyses. Due to the complexity of imputation for INDELS, we only used SNPs as variants in this study. The following filters were applied to SNPs before subsequent analyses. A SNP was excluded with: the strand bias p-value < 0.01 calculated with Fischer's exact test, two or more alternative alleles, a MAF <0.025, missing observation of the alternative allele on either the forward or reverse reads, being located within 4 bp of each other, being located within 5 bp of an INDEL, a mapping quality (MQ) score of <40, a phred score <20, a read depth (DP) of less than 10% of median or more than median plus 3 standard deviation of read depth, a quality depth (QD) < 5. We also removed sex chromosomes. After filtering, 11, 946, 148 SNPs on autosomes (SSC1 to SSC18) remained for the 60 animals across the whole-genome (Table 3).

Quality control of 50K, 61K, and 660K SNP panel

Quality control of the 50K (for 891 PBs), 61K (for 954 CBs) and 660K (for 17 PBs) were as follows: SNPs were excluded if they were duplicated, if they had a MAF <0.01. Furthermore, SNPs with genotype call rate <0.95 and SNPs with unknown map positions were removed. The quality control of genotypes was done for each trait separately, because the number of animals with both genotypes and phenotypes differ among the pork color traits. The numbers of SNPs after these exclusions are indicated in Table 4.

Imputation to whole-genome sequence

Beagle V5.0 (Browning et al., 2018) was used for imputation of 61K genotypes of CBs, 50K genotypes of PBs, and 660K genotypes of 17 purebred Duroc sows to the WGS (60 sequenced pigs). Default parameter settings of Beagle V5.0 were used, except for number of iterations for genotype phasing (default value was 12, but we used 25), and for effective population size (default value was 1,000,000 which is appropriate for a large population such as the human population, but we used 100 for our pig populations which helps with accurate

TABLE 3 Total number of SNPs, chromosome length, and average imputation accuracy (allelic DR^2) per chromosome after filtrations, and before and after imputation filtration criteria (allelic DR^2) in crossbreds (CBs) and purebreds (PBs).

		CB				PB			
		Before filtering on allelic DR^2		After filtering allelic DR^2 > 0.8 on CBs		Before filtering on allelic DR^2		After filtering allelic DR^2 > 0.8 on PBs	
Chromosome	Length (Mb)	Total number of SNPs	Mean allelic DR^2	Total number of SNPs	Mean allelic DR^2	Total number of SNPs	Mean allelic DR^2	Total number of SNPs	Mean allelic DR^2
SSC1	274	945,428	0.83	641,461	0.92	945,428	0.91	831,268	0.98
SSC2	152	791,977	0.81	509,047	0.92	791,977	0.88	668,474	0.97
SSC3	133	660,517	0.80	402,300	0.91	660,517	0.90	572,923	0.97
SSC4	131	740,467	0.84	517,636	0.92	740,467	0.92	669,685	0.97
SSC5	105	544,278	0.77	285,211	0.90	544,278	0.89	466,844	0.96
SSC6	171	824,770	0.79	475,917	0.91	824,770	0.89	696,109	0.96
SSC7	122	679,812	0.81	432,486	0.91	679,812	0.91	596,197	0.97
SSC8	139	866,293	0.81	534,659	0.91	866,293	0.92	774,027	0.97
SSC9	140	728,874	0.79	435,857	0.91	728,874	0.90	626,371	0.97
SSC10	69	579,468	0.79	341,971	0.90	579,468	0.88	483,451	0.95
SSC11	79	529,008	0.79	311,109	0.91	529,008	0.90	465,485	0.96
SSC12	62	429,775	0.79	244,130	0.90	429,775	0.88	359,175	0.95
SSC13	208	872,065	0.82	564,990	0.91	872,065	0.92	781,689	0.97
SSC14	142	755,463	0.81	480,616	0.92	755,463	0.91	660,825	0.97
SSC15	140	708,954	0.82	463,330	0.91	708,954	0.92	636,278	0.97
SSC16	80	523,910	0.80	320,423	0.91	523,910	0.91	465,914	0.96
SSC17	63	459,174	0.77	254,861	0.91	459,174	0.89	384,099	0.96
SSC18	56	305,915	0.79	177,266	0.92	305,915	0.91	270,808	0.97
Total/ Average	126	11,946,148	0.80	7,393,270	0.91	11,946,148	0.90	10,409,622	0.97

CB, crossbred; PB, purebred; SNP, single nucleotide polymorphism; Mb, megabyte; SSC, *sus scrofa*.

imputation of small populations (Browning et al., 2018)). Pedigree information was not used for imputation.

Evaluation of imputation accuracy is needed particularly for SNPs with low minor allele frequency (MAF) which are abundant in WGS. Evaluation of imputation accuracy was done in two ways. The first measure of imputation accuracy per SNP was obtained from the allelic DR^2 generated by Beagle, which is defined as the squared correlation between the expected dose (i.e., P (AB) + 2*P(BB)) and the true dose (Browning et al., 2018). Second, we were interested in imputation accuracy per pig (animal-specific imputation accuracy). True and imputed genotypes are needed to evaluate animal-specific imputation accuracy. Of the 60 sequenced pigs, 61K genotypes were available for 55 individuals which were used for assessing the animal-specific imputation accuracy using leave-one-out cross validation. Imputation accuracy was defined as the correlation between true and the most likely imputed genotypes. The leave-one-out cross validation analyses were performed using both Beagle V5.0 and FImpute (Sargolzaei et al., 2014) to compare the performance of the two programs. Due to large computation time, animal-specific imputation accuracy was assessed with the data for SSC18 only. For FImpute, the default values on all parameters were used, except for the error rate threshold to find progeny-parent mismatches, shrink factor for sliding windows, and amount of overlap for sliding windows. The values used for progeny-parent mismatches, shrink factor, and amount of overlap for sliding windows were 0.03, 0.15, and 0.65, respectively.

To assess whether a two-step imputation strategy would improve imputation accuracy compared with a one-step imputation strategy, particularly for low MAF SNPs (Kreiner-Møller et al., 2014; Lent et al., 2016; Bouwman et al., 2018), we performed imputation, using Beagle V5.0 only, from 50K SNP panel to WGS with 60 Duroc boars (one-step imputation strategy) and from 50K SNP panel to 660K SNP panel to WGS with 60 Duroc boars (two-step imputation strategy).

Quality control of imputed genotypes

Imputed genotypes were filtered based on the imputation reliability (allelic DR^2) produced by Beagle (Table 4). The chosen cut-off threshold for filtrations of allelic DR^2 was 0.8.

TABLE 4 Number of individuals and SNPs used for GWAS after quality control for different datasets (61K genotypes of CBs, 50K genotypes of PBs, and Combined dataset).

	CB			PB			Combined dataset		
Trait	61K	N	Imputed WGS	50K	N	Imputed WGS	61K + 50K	N	Imputed WGS
FCOLA	44,098	941[a]	7,376,594	35,775	873[a]	10,094,644	29,349	1,844	10,331,074
FCOLB	44,068	953	7,376,594	35,799	885[a]	10,090,482	29,349	1,844	10,331,074
FCOLL	44,068	953	7,377,298	35,782	879[a]	10,096,911	29,349	1,844	10,331,074
QFCOLA	44,070	953	7,376,594	35,801	882[a]	10,095,407	29,349	1,844	10,331,074
QFCOLB	44,068	953	7,376,594	35,809	885[a]	10,097,161	29,349	1,844	10,331,074
QFCOLL	44,068	953	7,376,594	35,809	881[a]	10,097,982	29,349	1,844	10,331,074
FMCOLA	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
FMCOLB	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
FMCOLL	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
TMCOLA	44,103	950	7,377,387	-	-	-	-	-	-
TMCOLB	44,103	950	7,377,387	-	-	-	-	-	-
TMCOLL	44,103	950	7,377,387	-	-	-	-	-	-
GMCOLA	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
GMCOLB	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
GMCOLL	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
ICOLA	44,070	953	7,376,594	35,809	891	10,099,416	29,349	1,844	10,331,074
ICOLB	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
ICOLL	44,070	953	7,376,594	35,809	891	10,099,578	29,349	1,844	10,331,074
Min	44,068	950	7,376,594	35,775	891	10,090,482	29,349	1,844	10,331,074
Max	44,103	954	7,377,387	35,809	891	10,099,578	29,349	1,844	10,331,074

[a]The total number of PBs, with both phenotypes and genotypes were 891. However, for these traits, there were extreme phenotypic records which were removed in the analyses to check if the GWAS, results would improve. Due to little changes in GWAS, results for PBs, those removed individuals were added to the analyses of combined CBs, and PBs. CB, crossbred; PB, purebred; WGS, whole-genome sequence; N, number of animals.

The reason for adapting a cut-off threshold of 0.8 was to achieve a balance between the average imputation reliability and the number of excluded SNPs. Consequently, of the 11,946,148 SNPs used for imputation, after exclusion of SNPs with imputation reliability less than 0.8, 7,393,270 and 10, 409, 622 SNPs remained for further analyses for CBs and PBs, respectively (Table 3).

Variance component estimates

Variance components, additive genetic variance (σ^2_A) and residual variance (σ^2_E), were estimated *via* the restricted maximum likelihood (REML) using ASReml program V4.0 (Gilmour et al., 2015) using a best linear unbiased prediction (BLUP) animal model as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \boldsymbol{Z}_a\mathbf{a} + \mathbf{e} \quad (1)$$

where **y** is the vector of phenotypic records, 1 is a vector of ones, **μ** is overall mean of phenotypic records, b is a vector of fixed class effects (the significant fixed effects for each trait is given in Table 5), **X** is a design matrix corresponding to the fixed effects, **a** is a vector of breeding values considered as random effects, $\boldsymbol{Z}_a$ is an incidence matrix that related phenotypic records to breeding values, and **e** is a vector of random residual effects. It is assumed that $\mathbf{a} \sim N(0, \mathbf{A}\sigma^2_a)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$ where σ^2_a and σ^2_e are the additive genetic and residual variances, respectively, and A is the numerator relationship matrix based on pedigree. Moreover, a narrow-sense heritability (h^2) was calculated as the division of the additive genetic variance by the total phenotypic variance as shown in Table 6. Standard errors of the variance components were also estimated by ASReml.

When the CBs and PBs were combined for variance component estimations, the heterogeneous genetic and residual variances were fitted in the model. Since CBs contained both males and females individuals, while PBs contained only female individuals, first an animal model was fitted to check the difference between the residual variances in CB and PBs as well as the difference between the residual variances between the male and female individuals. For all traits, the residual variances were different between the two populations as well as between the two sexes. Then, the first model was expanded to check if there was a difference between the genetic variances between the two populations (CBs *versus* PBs).

TABLE 5 Significance of the fixed effects (sex, slaughter date, room, pen, birth year-month, and population) included in the mixed model for different datasets (CB, PB, and Combined dataset) for the pork color traits.

	CB					PB				Combined dataset					
Trait	Sex	Slaughter date	Room	Pen	Birth year-month	Slaughter date	Room	Pen	Birth year-month	Sex	Slaughter date	Room	Pen	Birth year-month	population
FCOLA	**	***	NS	**	NS	*	NS	NS	*	***	***	NS	NS	NS	**
FCOLB	***	***	NS	NS	NS	***	NS	NS	***	***	***	NS	NS	NS	***
FCOLL	***	NS	NS	***	***	NS	*	NS	***	*	***	NS	NS	NS	***
QFCOLA	NS	NS	NS	NS	***	NS	NS	NS	***	NS	***	NS	NS	NS	***
QFCOLB	NS	***	NS	NS	***	NS	***	**	***	NS	***	NS	NS	NS	***
QFCOLL	NS	**	NS	NS	NS	NS	NS	*	NS	NS	***	NS	*	NS	***
FMCOLA	***	NS	**	NS	***	NS	NS	NS	***	***	***	NS	NS	NS	***
FMCOLB	*	NS	NS	NS	***	***	NS	NS	***	**	***	NS	NS	NS	***
FMCOLL	NS	***	NS	NS	*	NS	NS	NS	***	NS	***	NS	NS	**	NS
TMCOLA	***	***	NS	NS	***	-	-	-	-	-	-	-	-	-	-
TMCOLB	*	***	NS	NS	***	-	-	-	-	-	-	-	-	-	-
TMCOLL	NS	***	NS	NS	***	-	-	-	-	-	-	-	-	-	-
GMCOLA	NS	***	*	NS	NS	***	NS	NS	***	***	***	**	NS	NS	*
GMCOLB	NS	***	NS	NS	***	***	NS	NS	NS	*	***	NS	NS	NS	***
GMCOLL	NS	***	NS	*	*	***	NS	NS	NS	*	***	NS	NS	NS	***
ICOLA	NS	***	NS	**	***	***	NS	NS	***	*	***	NS	NS	NS	***
ICOLB	NS	***	NS	NS	NS	***	NS	NS	NS	NS	***	NS	NS	NS	***
ICOLL	NS	***	NS	NS	NS	NS	***	NS	***	***	***	*	NS	NS	***

CB, crossbred; PB, purebred; NS: non-significant. ***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

For parameter estimation, the data size presented in Table 2 was used, which ranges from 941 (FCOLA) to 954 (QFCOLB) for CBs, from 873 (FCOLA) to 891 for most of the traits. For Combined dataset, the total number of individuals was 1,844 for all traits. Relevant fixed effects fitted in the mixed model analysis for the 18 color traits are in Table 5.

Genome-wide association analyses

The model used for GWAS was a single-marker mixed linear association model (MLMA, mixed linear model based association analysis) implemented in GCTA version 1.92.1beta6 (Yang et al., 2011; Yang et al., 2014). The statistical model was as follows:

$$\hat{\mathbf{y}} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{g} + \mathbf{e} \quad (2)$$

Where $\hat{\mathbf{y}}$ was the vector of phenotypic records corrected for fixed effects (only significant fixed effects was used for correcting each trait, See Table 5). **u** was the additive effect (fixed effect) of the candidate SNP to be tested for association, **Z** was a vector containing the SNP genotype indicator variable coded as 0 (AA), 1 (AB), and 2 (BB). **g** was a vector of random polygenetic effects, and **e** was a vector of random residual effects. It was assumed that $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_g^2 and σ_e^2 were the genetic and residual variances, respectively. **G** was the genomic relationship matrix based on genotypes, constructed using GCTA software tool (Yang et al., 2011). GWAS was done using both medium-density panels and WGS data.

Significance testing

The significance threshold of SNP effects was assessed by using a false discovery rate (FDR) of 0.1 (Benjamini and Hochberg, 1995). Such threshold is needed to reduce the number of unacceptable false positives due to multiple testing. To account for population structure, the GWAS *p*-values for each trait were corrected for their corresponding genomic inflation factor (here called lambda) (Yang et al., 2011). Lambda was used for evaluating the bias. Lambda values for each data panel (medium-density and WGS) were computed as the median of the observed chi squared test statistics divided by the expected median of the corresponding chi squared distribution assuming 1 degree of freedom. *p*-values were used to compute the chi square test statistics. Moreover, quantile-quantile (qq) plot for each trait was used to evaluate the inflation of *p*-values by comparing the genome wide distribution of -log10 of the *p*-values with the

TABLE 6 Variance component estimates (additive and residual variances), and estimates of total heritability for 18 pork color traits for different datasets (CB, PB, and Combined dataset).

	CB			PB			Combined dataset					
Trait	σ^2_A (se)	σ^2_E (se)	h^2 (se)	σ^2_A (se)	σ^2_E (se)	h^2 (se)	σ^2_A (se)		σ^2_E (se)		h^2 (se)	
							CB	PB	CB	PB	CB	PB
FCOLA	0.37 (0.12)	1.07 (0.10)	0.26 (0.08)	0.33 (0.18)	2.39 (0.19)	0.12 (0.06)	0.33 (0.11)	0.32 (0.17)	1.03 (0.11)	2.15 (0.22)	0.24 (0.08)	0.13 (0.07)
FCOLB	0.41 (0.10)	0.64 (0.07)	0.39 (0.08)	0.48 (0.18)	1.61 (0.16)	0.23 (0.08)	0.39 (0.10)	0.39 (0.15)	0.57 (0.08)	1.46 (0.17)	0.40 (0.09)	0.21 (0.08)
FCOLL	0.54 (0.30)	1.52 (0.19)	0.26 (0.13)	0.56 (0.35)	4.84 (0.37)	0.10 (0.06)	0.43 (0.15)	0.66 (0.36)	1.4 (0.15)	4.23 (0.42)	0.23 (0.08)	0.13 (0.07)
QFCOLA	1.00 (0.23)	1.31 (0.16)	0.43 (0.08)	0.72 (0.22)	1.55 (0.18)	0.32 (0.09)	0.85 (0.21)	0.77 (0.22)	1.37 (0.17)	1.51 (0.26)	0.38 (0.08)	0.34 (0.09)
QFCOLB	0.46 (0.16)	1.55 (0.13)	0.23 (0.07)	0.12 (0.10)	1.59 (0.12)	0.07 (0.06)	0.47 (0.15)	0.10 (0.09)	1.56 (0.16)	1.56 (0.21)	0.23 (0.07)	0.06 (0.06)
QFCOLL	4.35 (1.04)	7.68 (0.79)	0.36 (0.08)	1.19 (0.81)	12.09 (0.91)	0.09 (0.06)	5.07 (1.13)	1.30 (0.83)	8.30 (0.98)	14.32 (1.34)	0.38 (0.07)	0.08 (0.05)
FMCOLA	0.86 (0.19)	0.94 (0.13)	0.48 (0.09)	0.46 (0.13)	0.77 (0.10)	0.38 (0.09)	0.67 (0.17)	0.44 (0.12)	1.06 (0.14)	0.81 (0.17)	0.39 (0.08)	0.35 (0.10)
FMCOLB	0.52 (0.15)	1.14 (0.11)	0.31 (0.08)	0.20 (0.09)	0.92 (0.09)	0.18 (0.08)	0.46 (0.15)	0.16 (0.08)	1.18 (0.13)	1.07 (0.16)	0.28 (0.08)	0.13 (0.07)
FMCOLL	2.73 (0.63)	2.70 (0.41)	0.50 (0.09)	1.06 (0.45)	4.30 (0.41)	0.20 (0.08)	3.65 (0.75)	1.08 (0.46)	2.47 (0.52)	4.35 (0.61)	0.60 (0.09)	0.20 (0.08)
TMCOLA	0.62 (0.13)	0.48 (0.08)	0.57 (0.09)	-	-	-	-	-	-	-	-	-
TMCOLB	0.34 (0.09)	0.72 (0.07)	0.32 (0.08)	-	-	-	-	-	-	-	-	-
TMCOLL	1.89 (0.53)	4.28 (0.41)	0.31 (0.08)	-	-	-	-	-	-	-	-	-
GMCOLA	0.63 (0.14)	0.72 (0.09)	0.47 (0.08)	0.67 (0.17)	0.88 (0.13)	0.43 (0.10)	0.61 (0.14)	0.58 (0.16)	0.77 (0.11)	0.99 (0.17)	0.44 (0.08)	0.37 (0.10)
GMCOLB	0.20 (0.07)	0.78 (0.06)	0.20 (0.06)	0.29 (0.09)	0.69 (0.08)	0.29 (0.09)	0.17 (0.06)	0.29 (0.09)	0.85 (0.07)	0.86 (0.12)	0.17 (0.06)	0.25 (0.08)
GMCOLL	1.53 (0.44)	3.98 (0.36)	0.28 (0.07)	2.00 (0.62)	4.43 (0.51)	0.31 (0.09)	1.47 (0.42)	2.08 (0.63)	4.16 (0.42)	4.75 (0.71)	0.26 (0.07)	0.31 (0.09)
ICOLA	0.79 (0.21)	1.77 (0.16)	0.31 (0.07)	0.95 (0.38)	3.40 (0.34)	0.22 (0.08)	0.57 (0.18)	1.07 (0.39)	1.71 (0.17)	3.01 (0.41)	0.25 (0.07)	0.26 (0.09)
ICOLB	0.44 (0.15)	1.55 (0.13)	0.22 (0.07)	0.19 (0.12)	1.62 (0.12)	0.10 (0.06)	0.45 (0.15)	0.26 (0.14)	1.57 (0.15)	1.62 (0.22)	0.22 (0.07)	0.14 (0.07)
ICOLL	2.61 (0.62)	3.53 (0.43)	0.43 (0.08)	1.29 (0.61)	6.81 (0.60)	0.16 (0.07)	2.37 (0.60)	1.11 (0.57)	3.31 (0.47)	6.16 (0.76)	0.42 (0.09)	0.15 (0.08)

CB, crossbred; PB, purebred; σ^2_A, additive genetic variance; σ^2_E, residual variance; h^2, narrow-sense heritability; se: standard error.

expected median of the corresponding chi squared distribution assuming a degree of freedom of one.

Linkage disequilibrium decay

LD decay pattern between pairwise SNPs (imputed sequence) was evaluated for both CBs and PBs. The pairwise LD values (r^2, defined as the correlation between alleles of two SNPs harbored at different loci (Hill & Robertson, 1968) between SNP pairs were computed for SNPs located within 2000 Kb windows and shorter (Figure 1). Due to large computation time, LD analyses were only done for SSC1.

Quantitative trait loci definition

For all traits, we defined the quantitative trait loci (QTL) regions according to the definition described by van den Berg et al. (2019) as follows. First, the SNPs on each chromosome were

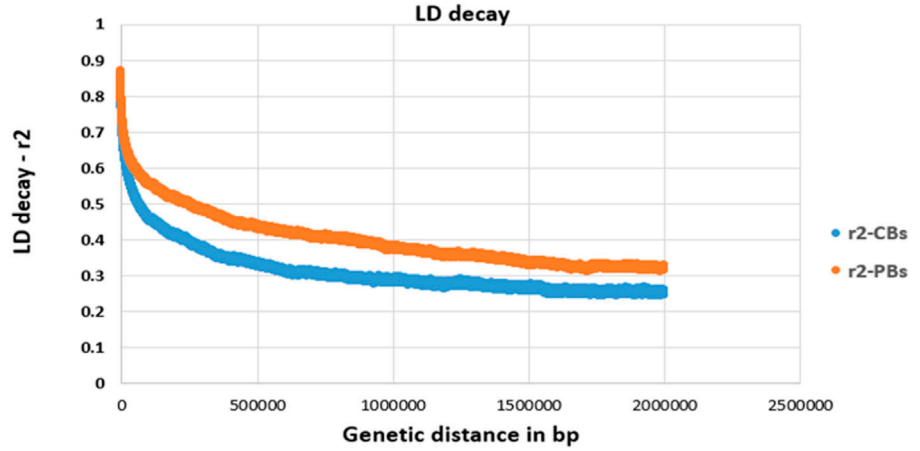


FIGURE 1
Linkage disequilibrium (LD, r^2) decay for SSC1 of CBs and PBs as a function of inter-SNP distance. Physical (genetic) distance is measured in base pair (bp).

ranked based on their -log10 *p*-values. Secondly, starting with the SNPs with the largest -log10 *p*-value, all significant SNPs that exceeded the FDR of 0.1 and surrounding SNPs within a 0.5 Mb region to the left and right of the SNP were assigned to that QTL region. These two steps were repeated until all significant SNPs were assigned to a QTL region. A distance of 0.5 Mb was chosen as the average LD of commercial pig lines decreases to less than 0.3 (Figure 1) when the SNPs are more than 0.5 Mb apart.

Variance explained by significant variants

The percentage of phenotypic variance explained by each SNP was estimated as: $\frac{2*p*q*a^2}{phenotypic\ variance}$, where *p* and *q* are the allele frequencies of major and minor alleles, and *a* is the estimated allele substitution effect. It should be noted that for the Combined dataset, the average of phenotype variance of crossbreds and purebreds was used for computation of variance explained.

Post-genome-wide association studies analyses

After GWAS, candidate gene identification and functional annotation for the significant SNPs were obtained using Ensemble annotation of *Sus scrofa* 11.1 (https://www.ensembl.org/info/data/biomart/). Genomic regions associated with the pork color traits were identified using a 1 Mb window (up- and down-stream of significant peak). The ClueGo plug-in (Bindea et al., 2009) and Cytoscape program (Shannon et al., 2003) were used to group and visualize the genes according to the biological processes in which they are involved in. The ClueGO plug-in uses both Gene Ontology (GO) terms and KEGG/BioCarta pathways to develop a GO/pathway network. Furthermore, ClueGO calculates enrichment and depletion tests for groups of genes based on the hypergeometric distribution and corrects the *p*-values for multiple testing. The *Sus scrofa* database (http://ftp.ensembl.org/pub/current_fasta/sus_scrofa/dna/) was used in pathway and biological processes investigation. We selected the 5th to the 10th levels of the GO hierarchy and a kappa score of 0.4 (Bindea et al., 2009). When no biological functions or pathways were found, these parameters were relaxed to be less stringent.

Results

Total number of pigs used for GWAS, and the descriptive statistics for 18 pork color traits including the minimum, maximum, mean and standard deviation of traits for different datasets (CB, PB and combined CBs and PBs[5]) are in Table 2. Because of the quality control during and after variant calling on WGS, not all SNPs on the 61K, 660K, and 50K SNP panels were present in the WGS, i.e., for the CBs, 26,585 SNPs of the 61K SNPs and 430,404 SNPs of the 660K SNPs were present, and for the PBs, 34,733 SNPs of the 50K SNPs were present in the WGS.

5 Through the manuscript, we call the combination of CBs and PBs as Combined dataset.

Population structure

Supplementary Figure S1 demonstrates (Supplementary Figure S1) population structure among the CBs (n = 954) and PBs (n = 891) populations, which was computed in Plink using the principal component analysis (PCA) procedure. The common SNPs between 61K and 50K (~30K) were used for plotting. The blue color shows the PB animals and the red color shows CB pigs. The CB individuals are dispersed across the plot.

Minor allele frequency distribution

The distribution of MAF from the 61K and 50K SNP panels were uniform, whereas the distribution of MAF from WGS was U-shaped with a substantial proportion of SNPs with small MAF values (approximately 19% of SNPs had a MAF lower than 0.025) (Supplementary Figure S2A). MAF distribution of sequence SNPs used for downstream analyses, after excluding the MAF <0.025, is given in Supplementary Figure S2B. Average MAF across the 28 autosomes before excluding MAF <0.025 was 0.28. After filtration of MAF with 0.025 cut-off threshold, the average MAF was 0.33.

Evaluation of accuracy of imputation

The average allelic DR^2 from the 61K and 50K SNP panels to sequence imputation before any filtration was 0.80 and 0.90 across all chromosomes, for CBs and PBs, respectively (Table 3). After filtration of allelic $DR^2 < 0.8$, the average allelic DR^2 from the 61K and 50K SNP panels to sequence imputation across all chromosomes was 0.91 for CBs and 0.97 for PBs (Table 3). The number of SNPs before and after allelic DR^2 filtration is given in Table 3. Beagle DR^2 varied between the CBs and PBs and also among the 18 chromosomes. For CBs, the smallest and largest Beagle DR^2 were obtained for SSC5 (0.77) and SSC4 (0.84), respectively. For PBs, the smallest Beagle DR^2 were obtained for SSC2 and SSC12 (0.88) and the largest Beagle DR^2 were obtained for SSC4, 8, 13, and 15 (0.92). Across all chromosomes, the average allelic DR^2 was larger for PBs than CBs.

The distribution of allelic DR^2 against MAF for CBs and PBs are shown in Figure 2. As expected, the imputation accuracy was lower for SNPs with lower MAF, and increased with MAF. The most pronounced increase in imputation accuracy was for MAF from the 0.01 to 0.10 for CBs and from 0.01 to 0.05 for PBs (Figure 2). For MAF larger than 0.10 for CBs and 0.05 for PBs, Beagle allelic DR^2 reached a plateau at about 0.15 for both CBs and PBs. When we performed filtration on Beagle allelic DR^2, most SNPs with a very low MAF (<0.01) were removed. Also, the average imputation accuracy was higher for PBs compared with CBs, which is most likely due to the higher genetic relationships between the sequenced pigs (60 Duroc males) and the PBs (Duroc females) compared with CBs. Moreover, CBs receive alleles from two other purebred parental lines and these lines are not represented in the reference panel for imputation.

The average animal-specific imputation reliability across the 55 sequenced PBs (only 55 individuals were both genotyped and sequenced) for SSC18 was 0.94 using Beagle V5.0 and 0.91 using FImpute (Supplementary Figure S3). Since the imputation accuracies produced by Beagle V5.0 were larger than FImpute for all analyses, be it only slightly, we used the imputed data from Beagle V5.0 in all subsequent analyses.

Two-step imputation accuracy

For all chromosomes, the mean imputation accuracy (Beagle allelic DR^2) was higher (0.90) for one-step imputation approach compared with the two-step imputation approach (0.85) (Supplementary Figure S4). After filtering Beagle allelic $DR^2 < 0.8$, the mean imputation for the two-step approach was slightly larger than those obtained from the one-step procedure (Supplementary Figure S4). Figure 3 compares the imputation accuracy (Beagle allelic $DR^2 > 0.8$) in one-step (50K to WGS) and two-step imputation (50K to 660K to WGS) procedures, which are plotted against MAF. As shown, the imputation accuracy of low MAF SNPs (MAF <0.02) remains challenging. The average allelic DR^2 across the genome was 0.996 for one-step approach and 0.985 for two-step approach. Due to very small difference in imputation accuracies between the two approaches, we performed the GWAS analyses only for the imputed variants from the one-step method.

Variance component estimates

Variance components and heritability estimates obtained from different datasets (CBs, PBs, and Combined dataset) for each color trait are in Table 6. Generally, the heritability estimates were low to high across the 18 meat color traits, and ranged from 0.20 ± 0.06 (GMCOLB) to 0.57 ± 0.09 (TMCOLA) for CBs, from 0.07 ± 0.06 (QFCOLB) to 0.43 (0.10) (GMCOLA) for PBs. When the Combined dataset was used, since the heterogeneous genetic and residual variances were fitted in the model, the heritability for CBs and PBs were estimated by the model separately and the heritabilities ranged from 0.17 ± 0.06 for GMCOLB to 0.60 ± 0.09 for FMCOLL in CBs and from 0.08 ± 0.05 for QFCOLL to 0.37 (0.10) GMCOLA in PBs.

Genome-wide association studies for pork color traits

Putative family stratifications were accounted for the GWAS analyses by incorporating the full genomic covariance among animals. Lambda ranged from 0.77 for ICOLB in Combined

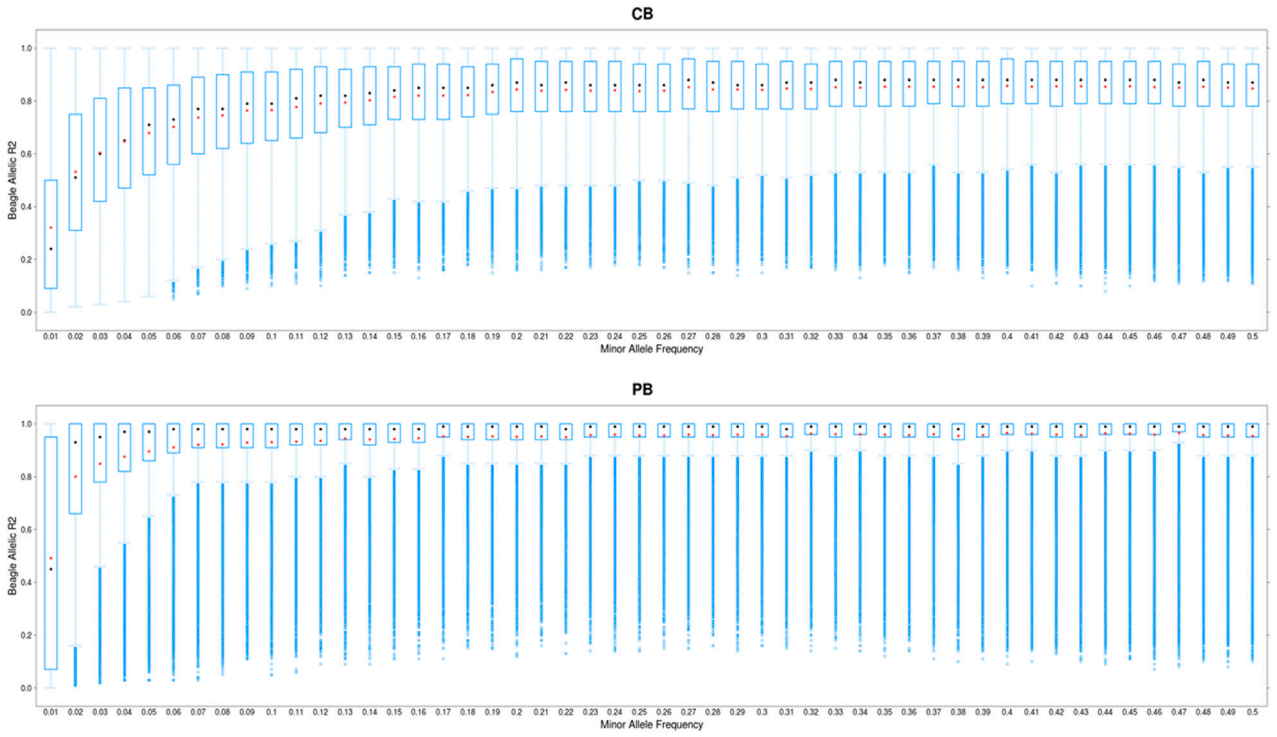


FIGURE 2
Boxplot showing the imputation accuracy (allelic DR^2) to whole-genome sequence (WGS) *versus* minor allele frequency (MAF) for CBs and PBs. The *x*-axis represents different classes of MAF (ranging from 0.01 to 0.5, with the steps of 0.01), and *y*-axis shows the imputation accuracies. The red and black dots are the mean and median of imputation accuracies across individuals in each MAF class.

dataset (61K CBs plus 50K PBs) to 1.00 for FMCOLB in sequenced PBs, and the mean lambda across all traits was 0.91 (results not shown), suggesting that any potential bias and any major effect of population stratification was taken into account in the GWAS analyses. For all traits for which the QTL regions were found and for all datasets (CBs, PBs, and Combined dataset), lambda increased slightly as SNP density increased (Table 7). This shows that the inflation of *p*-values is lower when WGS data was used compared with the medium-density SNP panel.

For most color traits (16 color traits for CBs, 12 color traits for PBs, and 11 color traits for Combined dataset), zero associated regions were detected at FDR of 0.1. Total number of significant SNPs, number of QTL regions, FDR threshold, and genomic inflation factor values are given in Table 7. Generally, more QTL were detected for traits of PBs than those of CBs at FDR = 0.1. However, when we used a more relaxed FDR threshold >0.1 and up to 0.4, suggestive significant SNPs were detected for some traits in different datasets, i.e., for FCOLA (PBs), FCOLL (PBs), QFCOLA (CBs, PBs, and Combined dataset), FMOCLA (CBs and PBs), FMOCLB (CBs), FMCOLL (Combined dataset), GMCOLA (PBs and Combined dataset), GMCOLB (PBs), GMCOLL (CBs), and ICOLB (PBs) (results not shown). For CBs, significant SNPs were identified for TMCOLA and GMCOLB (Table 7; Figure 4), and for PBs, the associated SNPs were found only for FMCOLB, FMCOLL, and ICOLB (Table 7; Figure 5). For the Combined dataset, we found the associated variants for more traits including FMCOLA, FMOCLB, GMCOLB, and ICOLB (Table 7; Figure 6).

For all traits and using the medium-density panels, 22 QTL regions containing 71 significant SNPs at a genome-wide FDR of 0.1 were detected, whereas 58 QTL regions comprising 16,261 significant SNPs were detected at the same significance level using WGS data (Table 7; Figures 4–6). The twenty two regions detected by medium-density panels overlaped with those detected by WGS. The number of QTL regions (2 using 61K and 2 using WGS) and significant SNPs (3 using 61K and 579 using WGS) were lowest for CBs using both SNP panel densities (61K and WGS), while the number of QTL regions (11 using 50K and 37 using WGS) and significant SNPs (11,352 using WGS) were highest when the Combined dataset was used for GWAS for both panel densities, except for the number of significant SNPs detected by PBs using 50K data which was highest (41 SNPs) compared with CBs (3 SNPs) and combined data (27 SNPs) (Table 7).

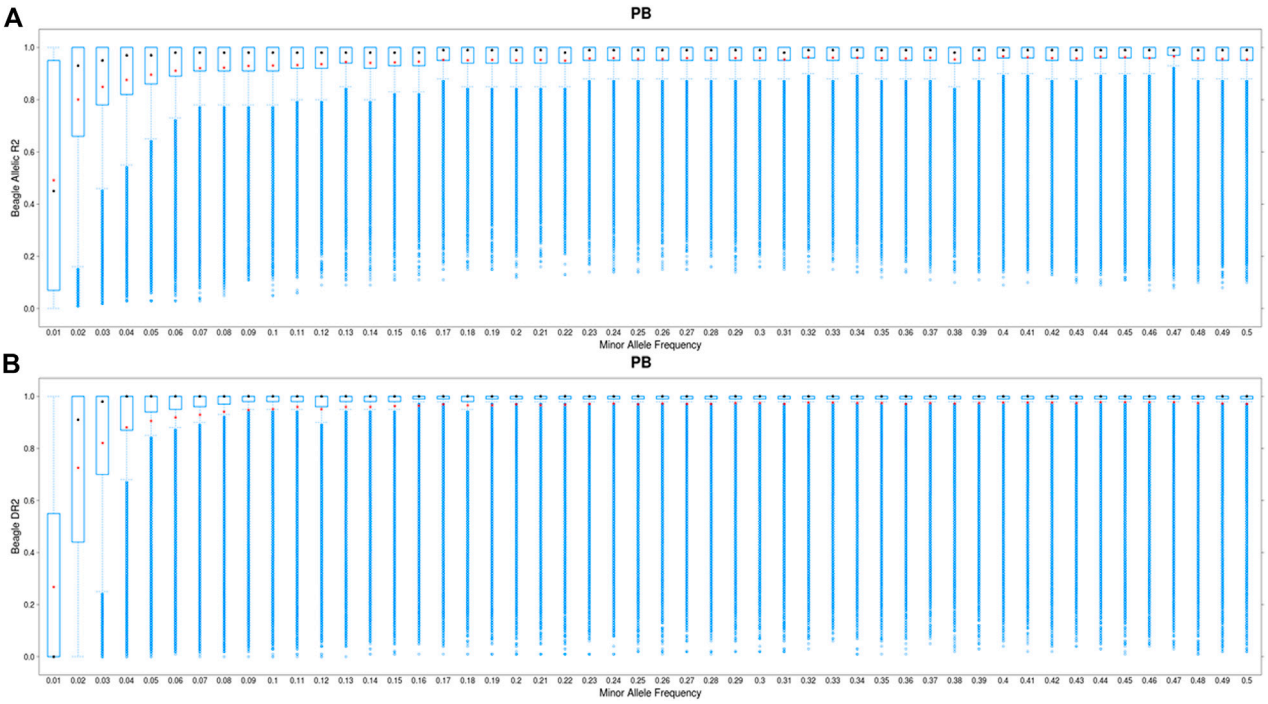


FIGURE 3
Boxplot showing the imputation accuracy (allelic DR^2) to whole-genome sequence (WGS) *versus* minor allele frequency (MAF) for PBs using a one-step imputation procedure from 50K to WGS **(A)** and a two-step imputation procedure, from 50K to 660K to WGS **(B)**. The *x*-axis represents different classes of MAF (ranging from 0.01 to 0.5, with the steps of 0.01), and *y*-axis shows the imputation accuracies. The red and black dots are the mean and median of imputation accuracies across individuals in each MAF class. Average DR^2 across the genome is 0.966 and 0.985 for the one-step and two-step imputation procedure, respectively.

Generally, the number of QTL regions increased with increasing panel density mainly for PBs and the Combined dataset, and did not change for CBs. For instance, for GMCOLB, the number of detected QTL region was only 1, regardless of what SNP density (61K or WGS) were used. For PBs, the additional QTL regions were located on SSC2 at 142.79–144.77 Mb and on SSC8 at 34 Mb, for FMCOLB, and on SSC10 at 38.13–38.29 Mb for FMCOLL (Table 8). Of all the new detected QTLs by WGS in PBs, the strongest new significant QTL was identified on SSC10 for FMCOLL (Figure 5). For the Combined dataset, the novel QTL regions identified by WGS compared with the medium-density SNP panel are given in Table 8 (also see Figure 6). Of all the new detected QTLs by WGS, the strongest new significant QTLs were identified on SSC1 for FMCOLA. Moreover, the total number of associated SNPs increased by increasing SNP density from 61K or 50K to WGS (Table 7; Figures 4–6). For example, it increased from 3 to 579 for CBs, from 41 to 4,330 for PBs, and from 27 to 11,352 for the Combined dataset.

For all datasets (CBs, PBs, and Combined dataset) and for all SNP panel densities (61K, 50K, and WGS), the majority of the significant SNPs were on SSC15 (Figures 4–6). For WGS, most of the significant SNPs were on SSC15 (93.17%), following by SSC5 (2.44%), and SSC2 (2.43%). The genomic location of the peak on SSC15 (across the traits) was between 119.57 and 122.50 Mb and between 119.56 and 123.56 for medium-density SNP panel and WGS, respectively. The position of the majority of SNPs within this window was the same between the medium-density and WGS data. For medium-density SNP panels, of the 71 significant SNPs, almost all of the significant SNPs were on SSC15 (~88%), except for 9 significant SNPs. Those 9 SNPs were: five SNPs detected by PBs for FMCOLL on SSC2 at 147.22–150.43 Mb, 1 SNP detected by the combined data for FMCOLA on SSC1 at 164.72 Mb, 1 SNP detected by the combined data for GMCOLB on SSC2 at 144.95 Mb, and finally 2 SNPs detected by PBs for TMCOLA on SSC5 at 9.41–9.44 Mb. Based on these results, using the medium-density SNP panels, only a few new QTL regions and SNPs were detected by Combined dataset compared with BPs and CBs. However, using WGS data, many more new QTL regions and SNPs were detected by only Combined dataset, and not detected by PBs and CBs, suggesting that increasing both the sample size and SNP density together improves identification of associated genomic regions.

Besides the increase of the number of QTL regions with WGS, the percentage of the phenotypic variance explained by the most significant SNPs also increased by WGS compared with medium-

TABLE 7 Descriptive statistics of results of the GWAS for the pork colors with detected associated regions in at least one of the datasets at FDR >0.1 (CBs, PBs, Combined dataset) using different SNP densities and imputed whole-genome sequence (WGS).

CB								
	61K				WGS			
Trait	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor
TMCOLA	2	1	5.29	0.88	396	1	5.27	0.86
GMCOLB	1	1	6.09	0.90	183	1	5.68	0.91
Total number of QTL/significant SNPs	3	2	-	-	579	2	-	-
PB								
	50K				WGS			
Trait	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor
FMCOLB	6	1	4.99	0.98	1,363	5	4.86	1.00
FMCOLL	16	6	4.36	0.91	2,034	12	4.70	0.94
ICOLB	19	2	4.27	0.96	933	2	4.98	0.98
Total number of QTL/significant SNPs	41	9	-	-	4,330	19	-	-
Combined Dataset								
	61K + 50K				WGS			
Trait	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor	Number of significant SNPs	Number of QTL regions	Threshold	Genomic inflation factor
FMCOLA	6	4	4.98	0.90	2,773	9	4.57	0.95
FMCOLB	5	2	4.90	0.93	2,645	12	4.60	0.96
GMCOLB	8	3	4.61	0.87	2,593	9	4.60	0.97
ICOLB	8	2	4.57	0.77	3,341	7	4.49	0.85
Total number of QTL/significant SNPs	27	11	-	-	11,352	37	-	-
Total number of QTL/significant SNPs for all datasets (CBs, PBs, Combined dataset)	71	22	-	-	16,261	58	-	-

CB, crossbred; PB, purebred; SNP, single nucleotide polymorphism; WGS, whole-genome sequence; QTL, quantitative trait loci.

density panels (Figure 7). Figure 7 shows the distribution of the percentage of phenotypic variance explained by the most significant SNPs identified using both WGS and medium-density panels (50K) for three pork color traits (FMCOLB, FMCOLL, and ICOLB) in PBs. For these three traits, the number of SNPs that explained more than two percent of phenotypic variance increased from 5 to 299 for FMCOLB and from 0 to 321 for FMCOLL, and from 8 to 263 for ICOLB, when WGS was used for GWAS compared with using 50K

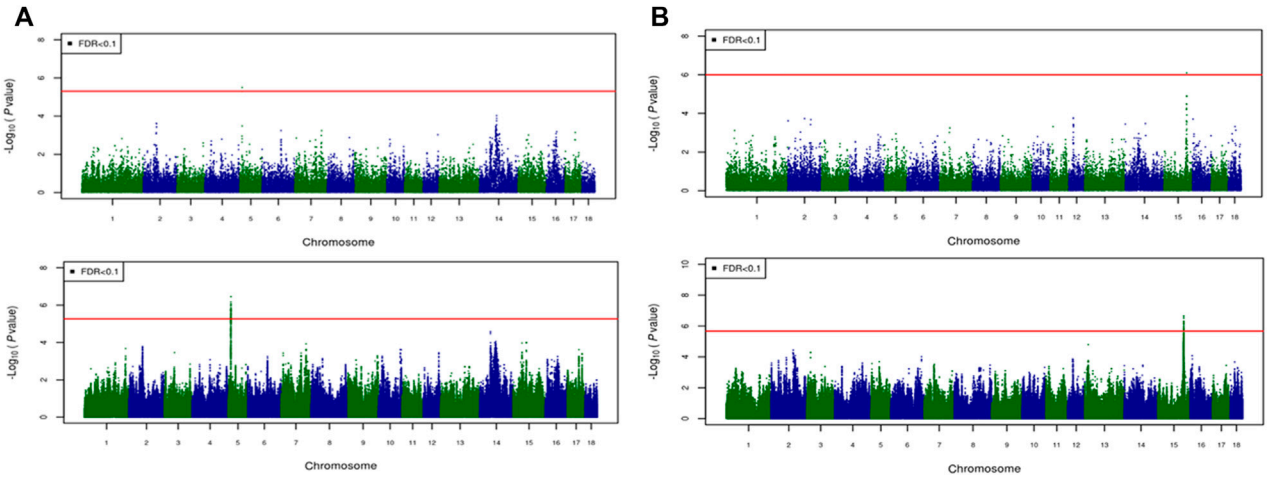


FIGURE 4
Associated regions detected by GWAS for crossbred pigs. Manhattan plots for: **(A)** TMCOLA and **(B)** GMCOLB using a 61K medium-density panel (top Figure) and WGS (bottom Figure). The -log10 *p*-values of single-SNP association along the entire genome are plotted against the genomic position of SNPs along the 18 autosome chromosomes. The SNPs associated with the corresponding traits exceeded the significance threshold at false discovery rate (FDR) of 0.1, having significant effects.

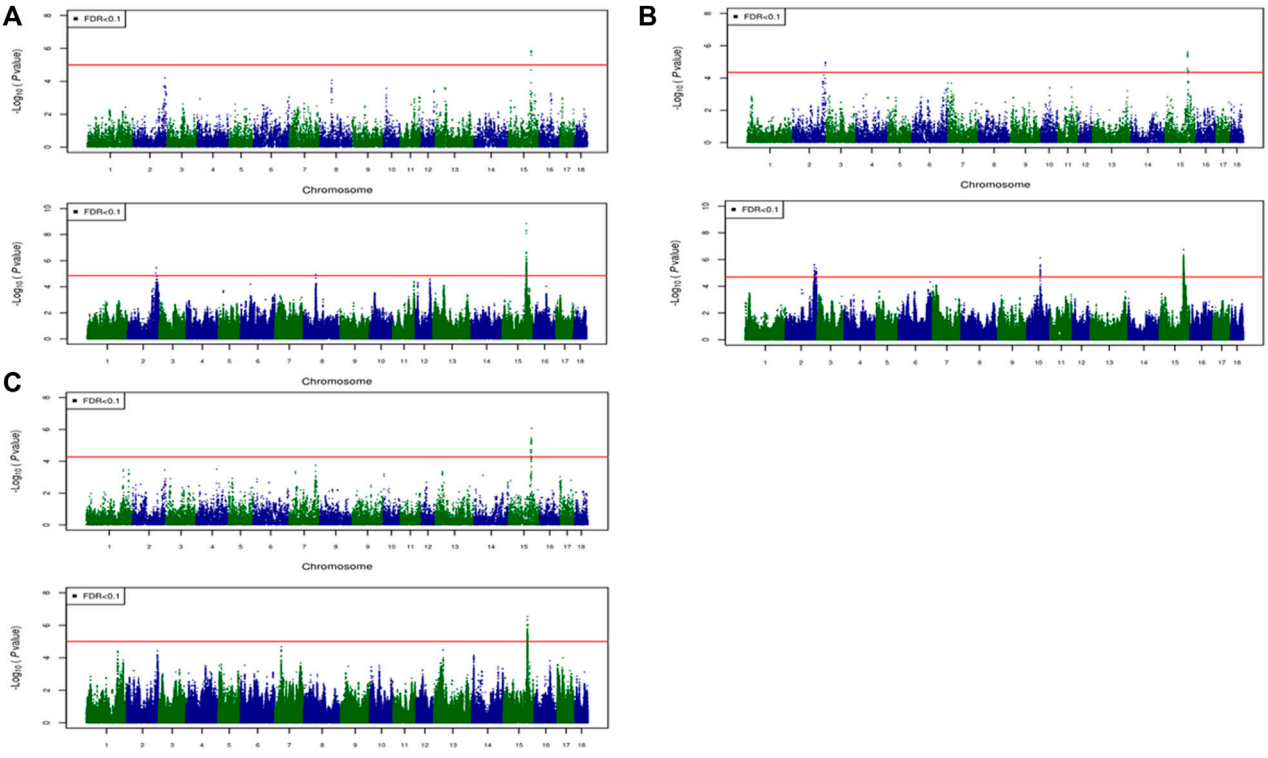


FIGURE 5
Associated regions detected by GWAS for purebred pigs. Manhattan plots for: **(A)** FMCOLB, **(B)** FMCOLL, and **(C)** ICOLB using a 50K medium-density panel (top Figure) and WGS (bottom Figure). The -log10 *p*-values of single-SNP association along the entire genome are plotted against the genomic position of SNPs along the 18 autosome chromosomes. The SNPs associated with the corresponding traits exceeded the significance threshold at false discovery rate (FDR) of 0.1, having significant effects.

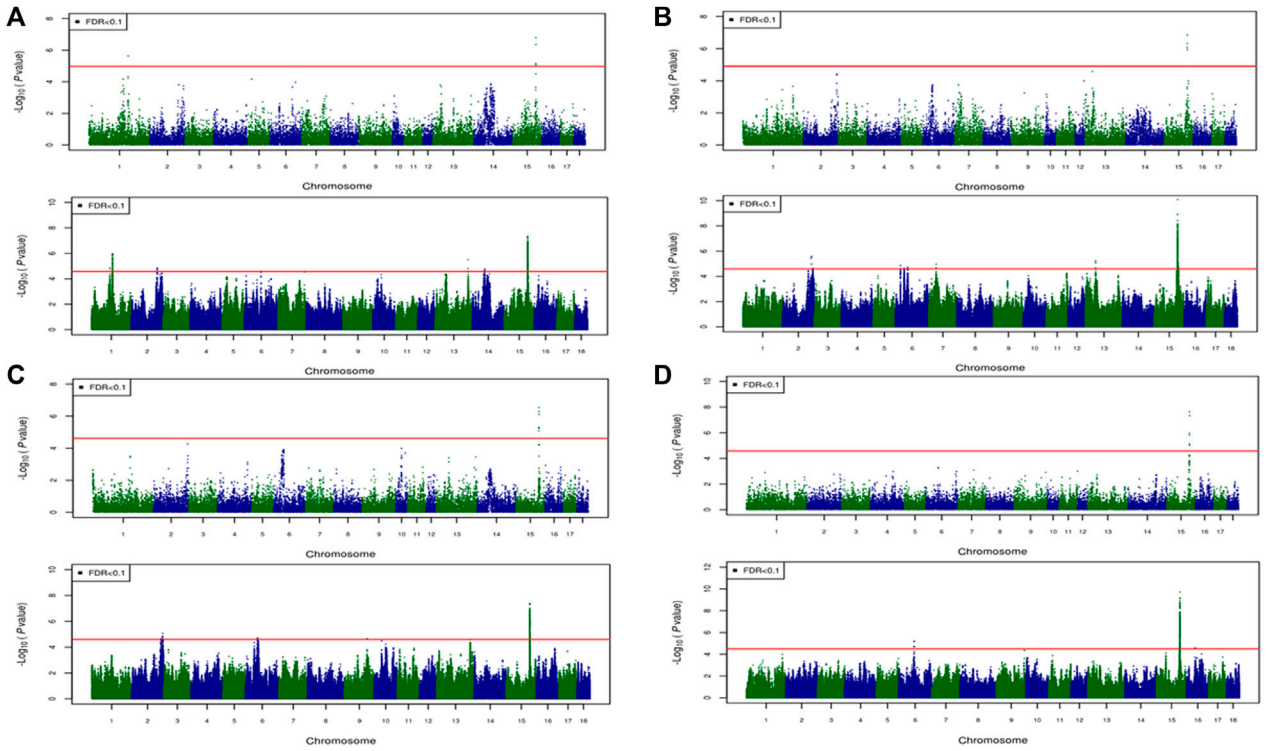


FIGURE 6
Associated regions detected by GWAS for combined crossbred and purebred pigs. Manhattan plots for: **(A)** FMCOLA **(B)** FMCOLB **(C)** GMCOLB, and **(D)** ICOLB using a combined (61K + 50K) medium-density panel (top Figure) and WGS (bottom Figure). The -log10 *p*-values of single-SNP association along the entire genome are plotted against the genomic position of SNPs along the 18 autosome chromosomes. The SNPs associated with the corresponding traits exceeded the significance threshold at false discovery rate (FDR) of 0.1, having significant effects.

TABLE 8 Novel genomic regions detected by WGS in PB pigs and in Combined dataset (combined CBs and PBs).

Trait	Number of QTL regions and their genomic region in mega base pairs (Mb)
PB	
FMCOLB	2 QTL regions on SSC2 (142.79 and 144.77–144.77). 1 QTL region on SSC8 (34 Mb)
FMCOLL	1 QTL region on SSC10 (38.12–38.28)
Combined dataset	
FMCOLA	1 QTL region on SSC2 (134.56–134.74). 1 QTL region on SSC13 (194.82–194.82). 1 QTL region on SSC14 (44.85–45.02)
FMCOLB	3 QTL regions on SSC2 (142.79, 144.77–144.81, and 148.00). 3 QTL regions on SSC6 (19.22–19.26, 57.29–57.30, and 32.18–32.22). 1 QTL region on SSC7 (23.73–23.85). 1 QTL region on SSC13 (24.73–24.74)
GMCOLB	2 QTL regions on SSC2 (144.89–144.95 and 149.09–149.12). 2 QTL regions on SSC6 (57.25 and 52.57–52.62). 1 QTL region on SSC9 (119.68–119.69). 1 QTL region on SSC13 (195.48)
ICOLB	1 QTL region on SSC6 (73.97–74.05). 1 QTL region on SSC16 (22.84)

PB, purebred; SNP, QTL, quantitative trait loci; Mb, megabyte; SSC, *sus scrofa*.

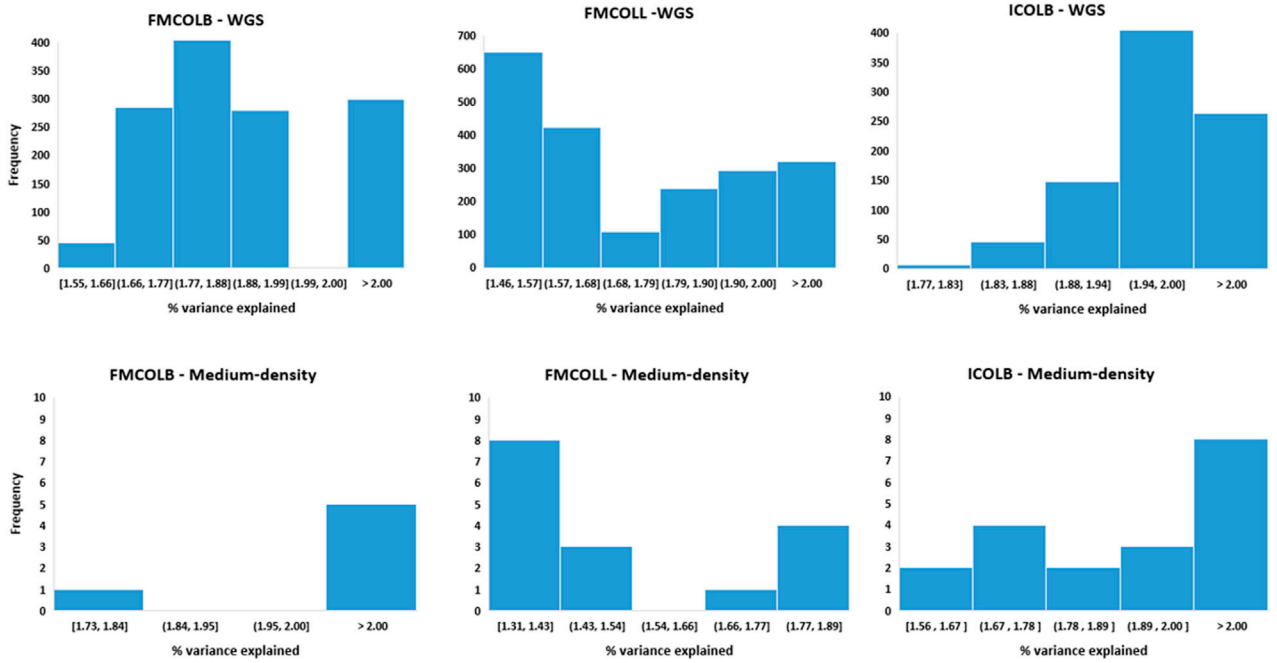


FIGURE 7
Distribution of the percentage of phenotypic variance explained by the most significant SNPs identified using different SNP densities (WGS (top row) and medium-density panels (bottom row)) for pork color traits (FMCOLB, FMCOLL, and ICOLB) in PBs.

for GWAS. The threshold 2% was chosen, because the maximum percentage of variance explained by the significant SNPs that exceed the FDR of 0.1 were ~3% and we therefore chose an arbitrary threshold lower than 3%.

Candidate genes identified by functional analysis

Candidate genes (and their functions) located on significant regions and/or nearby regions identified by WGS for meat color traits in different populations (crossbreds (CBs), purebreds (PBs), combined CBs and PBs) are given in Table 9. Since most of the significant SNPs for most traits are located on SSC15, for simplicity only the results of the functional analyses on SNPs detected on SSC15 are explained. For all traits, the significant SNPs span a region from 119.56 to 123.56 Mb on SSC15. The genes on this region are in Table 9. Many of these genes such as *PRKAG3* have been previously reported by Zhang et al. (2015) to be associated with pork pH and color. Some of the genes located on this region including *CNOT9, PRKAG3, CDK5R2, VIL1, TTLL4, CTDSP1, SLC11A1, ZFAND2B, USP37, RNF25, STK36, FEV, WNT6, IHH, WNT10A, NHEJ1, TMBIM1* are involved in the regulation of protein phosphorylation processes, proteolysis, intracellular transduction, and negative regulation of cell communication. In addition, *CATIP* and *ARPC2*, *VIL1* and *BCS1L* are involved in actin filament organization. There is evidence in the literature that meat color stability is inversely related to the phosphorylation of sarcoplasmic proteins (Mato et al., 2019; Li et al., 2020). An example of visualized gene network for the genes on SSC15 of ICOLB (region: 119.5–122.5 Mb) which shows the involved biological process is given in Supplementary Figure S5.

Discussion

In this study, we first assessed the imputation accuracy to WGS for two pig populations; CBs and PBs, using a small reference population of 60 sequenced PB key ancestors (Duroc males). Then, using the imputed WGS, we investigated whether the use of WGS data in a GWAS for pork color traits will improve the identification of the associated regions with respect to the extended QTL regions and/or detection of novel QTL regions in a sequenced-based GWAS relative to a medium-density SNP panel. The superiority of WGS over SNP panels is because of the existence of causal

TABLE 9 Candidate genes located on significant regions and/or nearby regions identified by whole-genome sequence (WGS) for meat color traits in different populations (crossbreds (CBs), purebreds (PBs), combined CBs and PBs).

	Trait	Chromosome[a]	Physical position (Mb)[b]	Candidate genes (gene functions)
CB	TMCOLA	SSC5	115.3–117.4	*FBX O 7*, *TIMP3*, *PWP1*, *RAC2*, *EIF3D*, *KCTD17* (Negative regulation of protein phosphorylation), cellular metal ion homeostasis (*KCTD17, PVALB*), Apolipoprotein L3-like, *RBFOX2, FBX O 7*
	GMCOLB	SSC15	120.6–120.9	*CATIP, ARPC2, VIL1* (Actin filament organization), *CNOT9*, *PLCD4, PRKAG3*, *DNAJB2, ZFAND2B, CNPPD1, INHA, CDK5R2, STK16, TTLL4, USP37, CTDSP1*, *SLC11A1*
PB	FMCOLB	SSC2	142.7–144.7	*ARHGAP26* (MAPK cascade and protein transport)
		SSC15	120.1–120.9	*ARPC2*, *CATIP*, *BCS1L*, *VIL1* (Actin filament polymerization), *PRKAG3, TTLL4, CTDSP1, USP37, SLC11A1* (Protein modification processes and protein phosphorylation)
	FMCOLL	SSC15	120.1–122.5	The same genes as previously described for FMCOLB.
	ICOLB	SSC15	119.5–122.5	*CNOT9,PRKAG3,CDK5R2,VIL1,TTLL4, CTDSP1, SLC11A1, ZFAND2B, USP37, RNF25, STK36, FEV, WNT6, IHH, WNT10A,NHEJ1, MBIM1* (Regulation of protein phosphorylation processes, proteolysis, intracellular transduction, and negative regulation of cell communication), *CATIP* and *ARPC2*, *VIL1*, *BCS1L* (actin filament organization)
Combined dataset	FMCOLA	SSC1	163.9–166.8	*MEGF11, U2, DIS3L, TIPIN, SCARNA14, MAP2K1, SNAPC5, RPL4, SNORD18, ZWILCH, LCTL, SMAD3, SMAD6, AAGAB, IQCH, C15orf61, MAP2K5, SKOR1, U6, PIAS1, CALML4, CLN6, FEM1B, ITGA11,* and *COR O 2B* (positive regulation of proteolysis, negative regulation of cell cycle, and regulation of transforming growth factor beta receptor signaling pathway)
		SSC15	120.1–120.9	The same genes as previously described for FMCOLB (purebreds)
	FMCOLB	SSC15	120-123.5	The same genes as previously described for FMCOLB (purebreds)
	GMCOLB	SSC2	144.7–150.9	nuclear receptor subfamily 3 group C member 1 (*NR3C1*), phosphodiesterase 6A (*PDE6A*), serine peptidase inhibitor, Kazal type 6 (*SPINK6*), *ARHGAP26*
		SSC6	52.57–59.13	Zinc finger protein 836-like, zinc finger protein 347 gene, NLR family pyrin domain containing 7, *PRK2*, *STRN4*
		SSC15	119.98–120.92	*PNKD, CNOT9, PLCD4, TMBIM1*, zinc finger protein 142 and *SLC11A1, TNS1, RUFY4, ARPC2, GPBAR1, AAMP, CATIP, CTDSP1, VIL1, USP37, BCS1L, RNF25, STK36, TTLL4, CYP27A1, PRKAG3, WNT6* and *WNT10A*
	ICOLB	SSC6	73.93–74.04	kazrin, periplakin interacting protein (*KAZN*) gene
		SSC15	119.55–120.92	The same genes as previously described for GMCOLB (Combined dataset)
		SSC16	22.59	*WDR70*

[b]This is the physical position in Mb and their nearby regions where the candidate regions were found (See Materials and Methods).
[a]If a significant region was not reported, no genes were found in that region. CB, crossbred; PB, purebred; Mb, megabyte; SSC, *sus scrofa*.

variants (rare variants responsible for phenotype variation) and rare variants with low LD with the SNPs on a medium-density panel (which most have moderate MAF), as the variance explained by these causal and rare variants can be better captured by WGS. Moreover, due to the relatively small size of our CB and PB populations (lower than 1,000 individuals per population), the Combined dataset was used in a GWAS, to assess whether enlarging the sample size will improve the potential advantage of WGS and enhance the power of detecting QTLs. Our imputation results showed a relatively high imputation accuracy obtained by Beagle V5.0 for both PBs (0.97) and CBs (0.91) after filtering the less accurate imputed genotypes (<0.8). Of the 18 pork colors, using different datasets, the genetic associations were identified only for a few traits (Table 7; Figures 4–6), and we did not detect any associated regions for most traits, regardless of panel density and dataset. WGS detected additional novel genomic regions for a few traits and with larger sample size (Combined dataset) (Table 8), the added value of WGS was more for detecting novel regions compared with SNP panel arrays. In the following sections, first, the factors influencing imputation accuracies are discussed, and then, the impact of using WGS data on GWAS results are discussed in detail.

Factors influencing imputation accuracy

Several factors influence the accuracy of imputation. These include the size of the reference population, the level of genetic relationship between the reference and validation population

(Hickey et al., 2011; Heidaritabar et al., 2015), MAF of the SNPs to be imputed (Heidaritabar et al., 2015; Bouwman et al., 2018), the program used for imputation (Bouwman et al., 2018; Bolormaa et al., 2019), and the density of validation population (Heidaritabar et al., 2015). The biggest challenge when imputing to WGS data is the imputation of the rare variants with low frequency. Figure S2 shows that of approximately 12 million called SNPs on SSC1 to SSC18, about 28% have a frequency less than 0.05 (Supplementary Figure S2). Due to the existence of this large proportion of rare SNPs, it is crucial to impute these variants as accurately as possible. To achieve the highest possible imputation accuracy for rare SNPs, several things can be done including careful selection of the reference individuals, appropriate imputation programs (Calus et al., 2014; Bouwman et al., 2018), and sequencing a sufficient number of animals, (Calus et al., 2014). The 60 Duroc males we chose for sequencing were key ancestors and jointly captured the maximum proportion of genetic variation present among the PBs. This is most likely reason that we achieved relatively high average imputation accuracies (average across all chromosomes and across all MAF) for both CBs (0.80) and PBs (0.90) (Table 3). Moreover, for low MAF SNPs (≤0.05), the average imputation accuracy ranged from 0.35 (when MAF was 0.01) to 0.65 (when MAF was 0.05) in CBs, and ranged from 0.5 to 0.9 in PBs, when MAF was 0.01 and 0.05 respectively (Figure 2). Even though the panel density of PBs is lower than CBs (50K *versus* 61K), PBs imputation accuracies are higher, which is likely due to the larger genetic relationships between the 60 reference sequenced pigs and the female PBs in the validation, as both population are Duroc and results in sharing more and longer haplotypes between the two populations (Hickey et al., 2011), while the CB population include the three-way cross between Duroc boars and Landrace-Yorkshire sows, and therefore, there is lower genetic relationship between the 60 Duroc boars and the CB population. Several studies have investigated the imputation of low MAF SNPs when imputing to the WGS in different species such as dairy cattle (van Binsbergen., 2017), beef cattle (Froberg Brøndum et al., 2014), pigs (Yan et al., 2017; Bouwman et al., 2018; Ros-Freixedes et al., 2019), sheep (Bolormaa et al., 2019), and found a poor imputation accuracy for low MAF SNPs. For example, Ros-Freixedes et al. (2019) reported imputation accuracy of 0.79 for MAF between 0.005 and 0.028 (n = 2,111), and 0.93 for MAF above 0.028 (n = 25,968) with simulated data, and for accuracy ranging from 0.51 (n = 11,312) for MAF <0.001 to 0.93 (n = 89,701) for MAF ≥0.028 in pigs. Even though Ros-Freixedes et al. (2019) used a much larger reference population compared to the 60 individuals in our study, our imputation accuracy from PBs for low MAF SNPs are similar to the values reported by them. Also, Bouwman et al. (2018) used three different imputation programs, and found imputation accuracy ranging from 0.5 to ~0.83 for SNPs with MAF lower than 0.05, when 168 sequenced pigs were used for imputation. Our imputation accuracy for low MAF SNPs from CBs are within the range reported by Bouwman et al. (2018) (0.35–0.65). Of note is that our measure of imputation accuracy is allelic DR^2, which is reliability, whereas the measure reported by Bouwman et al. (2018) and Ros-Freixedes et al. (2019) is the correlation between the true genotypes and imputed dosages. Meaning that with conversion of the allelic DR^2 to correlations, our imputation accuracy becomes even higher (r = 0.59 to 0.81 for CBs and r = 0.71 to 0.94 for PBs). This suggests that the overall performance of Beagle V5.0 for imputation of low MAF SNPs was good, even with a small reference population size and small genetic relationship between the CBs and PBs. However, to be more certain about the performance of Beagle V5.0 compared with other imputation programs, we compared imputation accuracies from Beagle V5.0 and FImpute in a leave-one-out cross validation approach (Supplementary Figure S3). The average animal-specific imputation accuracy across 55 pigs was slightly higher for Beagle (0.94) than FImpute (0.91).

Increasing the size of the reference population was more beneficial for imputing rare SNPs compared with more common SNPs for both imputation to the WGS in cattle (van Binsbergen et al., 2014), and imputation from low-to medium-density SNP panel (60K) in layer chickens (Heidaritabar et al., 2015). This is because with a larger reference population, the probability that multiple copies of alleles are present for correct haplotype construction increases and this in turn increases the quality of imputation of low-frequency SNPs. For dairy cattle, it was proposed to sequence not more than 500 individuals, as more than this number only slightly improved the accuracy of imputation accuracy. However, it is generally hard to determine exactly how many more sequenced individuals are required as the reference, and which level of genetic relationship to the validation population is required for minimizing the imputation error rate (Meuwissen et al., 2013). Based on our results, it seems that the low number of sequenced animals, when carefully selected, is only a limiting factor for imputation of low MAF SNPs, as we still obtained reasonable imputation reliabilities for high MAF SNPs. In our analyses, we excluded many of those low MAF SNPs (~20%) with low accuracy of imputation (Supplementary Figure S2), meaning that some of the causative mutations contributing to the genetic variation of a complex trait may have been removed during the filtration of MAF. If enlarging the reference population is not possible due to high costs of sequencing, an alternative to retaining the low MAF SNPs (potential causative mutations) is to use dosage scores instead of genotypes for downstream analyses such as GWAS, or genomic predictions. Van den Berg et al. (2019) compared the GWAS results of using genotypes with those of dosage scores and found an improvement of QTL detection (56.7 and 26.9% additional QTL regions for their two studied lines), because dosage scores coded as any real value between 0 and 2 accounted for uncertainty of imputation, and therefore all SNPs were used in their analysis. They also found that the most significant SNPs in the QTL regions explained more of the

TABLE 10 Percentage of phenotypic variance explained by the most significant SNP on SSC15 for the pork color traits at different panel densities and different populations.

Trait	Population	Panel density	Physical position (Mb)	Percentage of phenotypic variance explained
GMCOLB	CB	WGS	120.72	1.82
GMCOLB	CB	61K	120.71	1.67
FMCOLB	PB	WGS	120.42	3.05
FMCOLB	PB	50K	120.80	2.17
FMCOLL	PB	WGS	120.86	2.37
FMCOLL	PB	50K	120.80	1.89
ICOLB	PB	WGS	120.67	2.59
ICOLB	PB	50K	120.86	2.37
FMCOLA	Combined CB and PB	WGS	120.19	1.16
FMCOLA	Combined CB and PB	Medium-density	120.80	1.05
FMCOLB	Combined CB and PB	WGS	120.42	2.06
FMCOLB	Combined CB and PB	Medium-density	120.80	1.29
GMCOLB	Combined CB and PB	WGS	120.66	1.27
GMCOLB	Combined CB and PB	Medium-density	120.21	1.05
ICOLB	Combined CB and PB	WGS	120.67	1.57
ICOLB	Combined CB and PB	Medium-density	120.70	1.15

CB, crossbred; PB, purebred; Mb, megabyte; WGS, whole-genome sequence. Combined CB and PB means combining crossbred and purebred populations, Physical position (Mb) means genomic position in Megabyte,

phenotypic variance when using dosage scores compared to using genotypes (Van den Berg et al., 2019).

Genome-wide association studies using purebred pigs, crossbred pigs and Combined dataset

We did a GWAS for 18 pork color traits in CBs, PBs, and combined data using both SNP panel arrays and WGS and investigated whether the WGS can improve the power of GWAS compared to the medium-density SNP panels. Of the 18 pork colors, using different datasets, we did not detect any associated regions for most traits, regardless of panel density (see Results). The QTL regions were identified (with FDR of 0.1) only for a few traits including TMCOLA and GMCOLB (CBs), FMCOLB, FMCOLL, and ICOLB (PBs) and FMCOLA, FMCOLB, GMCOLB, and ICOLB (Combined dataset). Generally, we identified more QTL regions with WGS (n = 58) compared with medium-density SNP panels (n = 22). Most of the identified QTL regions with all genotype densities were also reported in other GWAS studies that used the same color traits (Zhang et al., 2015; Yang et al., 2017). The most significant QTL region reported by Zhang et al. (2015) was located on SSC15 spanning 133–134 Mb which explained 3.51%–17.06% of genetic variance for five measurements of pH and some color traits (Minolta color A* and B* for fresh ham and color B* measured on thawed loin muscle). This region is very close to previously reported gene *PRKAG3* controlling both meat pH and color in pigs. Our results are consistent with results of Zhang et al. (2015) and Yang et al. (2017), as this region[6] on SSC15 was identified by both densities and the three datasets. In the present study, for both WGS and medium-density panels and for most traits, most of the significant SNPs were on SSC15 at 119.57 and 122.50 Mb for WGS and at 119.56 and 123.56 for medium-density SNP panel (see Results). The percentage of phenotypic variance explained by the most significant SNP on SSC15 for different pork color traits and different density panels are shown in Table 10. It should be noted that in the present study, the percentage of variance explained is not cumulative, because variants were tested one at a time (See model 2) in *Materials and Methods*). Thus, the estimated SNP effects of surrounding variants were not independent due to LD. For all traits where the genomic region on SSC15 was significant, the variance explained by the most significant SNP was higher for WGS compared with medium density panels. The added value of WGS for improving the power of GWAS (with respect to the number of identified QTL) have been shown in several species including dairy cattle (Daetwyler et al., 2014; van den Berg et al., 2019), beef cattle (Zhang et al., 2015; Wang et al., 2020), pig (Yan et al.,

6 Zhang et al. (2015) and Yang et al. (2014) used the Sscrofa 10.2, while we used Sscrofa 11.1. The region 133–134 Mb on SSC15 on Sscrofa 10.2 is the same region on 120–121 Mb on Sscrofa 11.1.

2017), human (The 1000 Genomes Project Consortium. 2010; Höglund et al., 2019), and tomato (Van Binsbergen et al., 2014). A general speculation for more power of GWAS in denser genome coverage with (WGS) is the presence of causative SNPs and SNPs with higher LD within the data, which improves the power for identification of SNPs with small effects.

When the combined dataset was used for GWAS, many more QTL regions (11 for medium-density panels and 37 for WGS) were identified, suggesting that the added value of WGS was more for detecting novel regions compared with medium-density SNP panels in larger samples. This could be because with the larger sample size, the effect of causative mutations on polygenic quantitative traits might be estimated more accurately. Also, for the Combined dataset, we filtered the imputed genotypes based on the allelic DR^2, meaning that some of the imputed SNPs excluded in CBs analyses (3,016,352 SNPs) due to imputation accuracy less than 0.8 were included in the Combined dataset GWAS analysis, and yet the power of GWAS improved. This shows that the imputation error rate is not really a limiting factor for GWAS. Similar results are shown by Van Binsbergen et al. (2014) where they found that despite their relatively low imputation accuracy (average correlation of 0.34 between true genotypes and allele dosage) in tomato WGS data, the power of a GWAS can still be improved. They reported that more significant SNPs (>65 SNPs in 9 regions) were found in the GWAS using the imputed WGS compared to using the low-density SNP arrays (no significant SNPs). They argued that as long as the squared imputation accuracy (allelic DR^2 in our study) is higher than the expected LD between the SNPs on the lower density panel (50K and 61K in our study) and the SNPs in the WGS data, imputation is advantageous, as more information is still added by imputation (Van Binsbergen et al., 2014). Average LD between the imputed sequenced SNPs located within 2 Mb windows and shorter (on SSC1) was 0.31 and 0.40 for CBs and PBs, respectively, which is lower than the average squared imputation accuracy, which is 0.91 and 0.97 for the corresponding populations (see Table 3). This may explain why the imputed sequence data improved the QTL detection through a GWAS. Moreover, van den Berg et al. (2019) found that although their imputation from 80K to 660K to WGS in pig populations resulted in poor imputation accuracy (Beagle allelic DR^2 in their study ranged from 0.39 to 0.49 and from 0.83 to 0.93, before and after variant filtrations), they still found that using imputed WGS instead of a lower density SNP panel increased the number of detected QTL (48.9 and 64.4% more for their different lines) and the estimated proportion of phenotypic variance explained by these QTL (van den Berg et al., 2019). Also, Heidaritabar et al. (2015) found that the average allelic DR^2 (before quality control) from the 60K SNP panel to WGS imputation in layers was 0.64, but they still observed an increase of prediction accuracy of 1% using WGS compared with 60K for number of eggs. All these results show that most likely the accuracy of the imputed genotypes is not a limiting factor for GWAS and genomic predictions.

Functional analyses

We detected several candidate genes for the color traits in CBs, PBs and Combined dataset. For most color traits, a region spanning from 119.5 to 123.5 Mb on SSC15 was consistently detected. Some of the genes located on this region including: *ciliogenesis-associated TTC17-interacting protein* (*CATIP*), *villin-1 (VIL1), protein kinase AMP-activated non-catalytic subunit gamma 3 (PRKAG3), tubulin tyrosine ligase like 4 (TTLL4), ubiquitin specific peptidase 37 (USP37), CTD small phosphatase 1 (CTDSP1)* and *solute carrier family 11 member 1* (*SLC11A1*) were consistently detected for all color traits reported here, hence they were considered the best candidates' genes in the QTL region for the color traits. Genes such as *VIL1, PRKAG3, TTLL4,* and *SLC11A1, USP37* have been previously reported to be associated with meat quality, pH and color (Ciobanu et al., 2001; Uimari and Sironen, 2014; Zhang et al., 2015; Verardo et al., 2017). Although *CTDSP1* gene has not been previously associated with meat color traits in pigs, it has been found to be associated with meat color Minolta L* traits in Nellore cattle (Marin-Garzon et al., 2021). The genes reported in this study are involved in actin filament organization, regulation of protein phosphorylation processes, proteolysis, and intracellular transduction. There is evidence in the literature that meat color stability is inversely related to the phosphorylation of sarcoplasmic proteins such as myoglobin (Mato et al., 2019; Li et al., 2021). Meat color is determined by myoglobin concentration as well as the relative content of oxymyoglobin, deoxymyoglobin and metmyoglobin (Zhang et al., 2015; Li et al., 2021). Studies have shown that myoglobin phosphorylation may lead to changes in its secondary structure, therefore reducing myoglobin stability and increasing its autoxidation rate, which further accelerated the accumulation of metmyoglobin (Zhang et al., 2015). Further exploration of these genes and protein phosphorylation pathway will improve our understanding on genetic factors affecting meat quality hence leading to strategies to improve color in pork.

Conclusion

Use of purebred and crossbred populations genotyped by medium-density panels resulted in relatively high imputation accuracy (0.97 for purebreds and 0.91 for crossbreds after variants quality control) to WGS. Additional QTL regions were detected when using the WGS data compared with a medium-density SNP panels. The performance of WGS relative to the medium-density panels is best when the sample size is the largest (combining cross- and purebreds), suggesting that sample size is a limiting factor to capitalize on the added value of WGS in a GWAS.

Data availability statement

The sequence data was generated on commercial Duroc pigs owned by Hendrix Genetics, Hypor. Data may be available from authors upon reasonable request and authorization from the company.

Ethics statement

The data for the current study was reviewed by the University of Alberta Animal Care and Use Committee and considered as Category A, meaning that there was no animal manipulation. The pigs were produced as part of commercial pig breeding and pork operations and cared for according to the Canadian Quality Assurance Program, which take animal health and well-being into consideration in line with the Canadian Council on Animal Care guidelines. Therefore, no formal ethics approval was needed.

Author contributions

MH, AH, MCAMB, PC and GSP conceived and designed the experiments. MH analyzed the data. KK and PS performed alignment and variant calling. ED helped with performing the post-GWAS analyses. MH wrote the manuscript. MH, MCAMB, AH, KK, PS, PC, ED, and GSP discussed and improved the manuscript. All authors read and approved the final manuscript.

Funding

This project was financially supported by a Collaborative Research and Development Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), Grant # CRDPJ 485526—2015 and Hendrix Genetics.

Acknowledgments

We would like to thank Hendrix Genetics for providing the data.

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.2022.1022681/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Multidimensional scaling result for assessing the structure of the population.

SUPPLEMENTARY FIGURE S2
(A) Minor allele frequency (MAF) distribution in the 61K SNP panel of CBs and 50K SNP panels of PBs, and sequence Duroc purebred males. **(B)** Minor allele frequency (MAF) distribution of sequence data (11,946,148 SNPs) after MAF filtration with cut-off threshold = 0.025.

SUPPLEMENTARY FIGURE S3
Animal-specific imputation accuracy for the 55 sequenced animals, using leave-one-out cross-validation approach.

SUPPLEMENTARY FIGURE S4
Comparison of imputation accuracy (Beagle allelic DR2) in one-step (50k to WGS) and two-step imputation approach before any filtration **(A)** and after filtering Beagle allelic DR2 > 0.8 **(B)**.

SUPPLEMENTARY FIGURE S5
Gene network constructed based on the candidate and/or nearest genes to the significant SNPs on SSC15 for ICOLB in PBs. Functionally grouped network with terms as nodes are linked based on their kappa score level (0.4).

References

The 1000 Genomes Project ConsortiumAbecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi:10.1038/nature09534

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 15, 1091–1093. doi:10.1093/bioinformatics/btp101

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Bolormaa, S., Chamberlain, A. J., Khansefd, M., Stothard, P., Swan, A. A., Mason, B., et al. (2019). Accuracy of imputation to whole-genome sequence in sheep. *Genet. Sel. Evol.* 51, 1. doi:10.1186/s12711-018-0443-5

Bouwman, A. C., van Son, M., Harlizius, B., and Zumbach, B. (2018). Imputation accuracy of whole-genome sequence in pigs." in *World congress on genetics applied to Livestock production (WCGALP)*.

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

Calus, M. P., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F., and Mulder, H. A. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. *Animal* 8, 1743–1753. doi:10.1017/S1751731114001803

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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi:10.1038/ng.3656

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

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* 112, 39–47. doi:10.1038/hdy.2013.13

Froberg Brøndum, R., 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, 728. doi:10.1186/1471-2164-15-728

Gilmour, A., Gogel, B., Cullis, B., Welham, S. J., and Thompson, R. (2015). *ASReml user guide release 4.1 structural specification*. Hemel Hempstead. VSN International Ltd.

Glitsch, K. (2000). Consumer perceptions of fresh meat quality: Cross-national comparison. *Br. Food J.* 102 (3), 177–194. doi:10.1108/00070700010332278

Heidaritabar, M., Calus, M. P. L., Vereijken, A., Groenen, M. A. M., and Bastiaansen, J. W. M. (2015). Accuracy of imputation using the most common sires as reference population in layer chickens. *BMC Genet.* 16, 101. doi:10.1186/s12863-015-0253-5

Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., and van der Werf, J. H. J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.* 43, 12. doi:10.1186/1297-9686-43-12

Höglund, J., Rafati, N., Rask-Andersen, M., Enroth, S., Karlsson, T., Ek, W. E., et al. (2019). Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Sci. Rep.* 9, 16844. doi:10.1038/s41598-019-53111-7:

Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2014). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88, 544–551. doi:10.2527/jas.2009-2064

Kreiner-Møller, E., Medina-Gomez, C., Uitterlinden, A. G., Rivadeneira, F., and Estrada, K. (2014). Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur. J. Hum. Genet.* 23, 395–400. doi:10.1038/ejhg.2014.91

Lent, S., Deng, X., Cupples, L. A., Lunetta, K. L., Liu, C. T., and Zhou, Y. (2016). Imputing rare variants in families using a two-stage approach. *BMC Proc.* 10, 209–214. doi:10.1186/s12919-016-0032-y

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, J. H., Mazur, C. A., Berisa, T., and Pickrell, J. K. (2021). Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res.* 31, 529–537. doi:10.1101/gr.266486.120

Li, M., Li, X., Xin, J., Li, Z., Li, G., Zhang, Y., et al. (2017). Effects of protein phosphorylation on color stability of ground meat. *Food Chem.* 219, 304–310. doi:10.1016/j.foodchem.2016.09.151

Li, Y., Li, B., Yang, M., Han, H., Chen, T., Wei, Q., et al. (2020). Genome-wide association study and fine mapping reveals candidate genes for birth weight of Yorkshire and Landrace pigs. *Front. Genet.* 11, 183. doi:10.3389/fgene.2020.00183

Mato, A., Rodríguez-Vázquez, R., López-Pedrouso, M., Bravo, S., Franco, D., and Zapata, C. (2019). The first evidence of global meat phosphoproteome changes in response to pre-slaughter stress. *BMC Genomics* 20, 590. doi:10.1186/s12864-019-5943-3

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

Meuwissen, T., and Goddard, M. (2010a). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185, 623–631. doi:10.1534/genetics.110.116590

Meuwissen, T., and Goddard, M. (2010b). The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. *Genetics* 185, 1441–1449. doi:10.1534/genetics.110.113936

Meuwissen, T., Hayes, B., and Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci.* 1, 221–237. doi:10.1146/annurev-animal-031412-103705

Miar, Y., Plastow, G. S., Moore, S. S., Manafiazar, G., Charagu, P., Kemp, R. A., et al. (2014). Genetic and phenotypic parameters for carcass and meat quality traits in commercial crossbred pigs. *J. Anim. Sci.* 92, 2869–2884. doi:10.2527/jas.2014-7685

Ros-Freixedes, R., Whalen, A., Chen, C., Gorjanc, G., Herring, W. O., Mileham, A. J., et al. (2019). Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genet. Sel. Evol.* 52, 17. doi:10.1186/s12711-020-00536-8

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15, 478. doi:10.1186/1471-2164-15-478

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Snelling, W. M., Hoff, J. L., Li, J. H., Kuehn, L. A., Keel, B. N., Lindholm-Perry, A. K., et al. (2020). Assessment of imputation from low-pass sequencing to predict merit of beef steers. *Genes.* 11, 1312. doi:10.3390/genes11111312

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

van Binsbergen, R. (2017). *Prospects of whole-genome sequence data in animal and plant breeding*. Wageningen: Wageningen University. PhD thesis.

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, 2. doi:10.1186/s12711-019-0445-y

van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., LevyMoonshine, A., et al. (2013). From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 11, 1–11. doi:10.1002/0471250953.bi1110s43

Wang, Y., Zhang, F., Mukiibi, R., Chen, L., Vinsky, M., Plastow, G., et al. (2020). Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: Carcass merit traits. *BMC Genomics* 21, 38. doi:10.1186/s12864-019-6273-1

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, Q., 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

Wu, Y., Zheng, Z., Visscher, P. M., and Yang, J. (2017). Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* 18, 86–10. doi:10.1186/s13059-017-1216-0

Yan, G., Guo, T., Xiao, S., Zhang, F., Xin, W., Huang, T., et al. (2018). Imputation-based whole-genome sequence association study reveals constant and novel loci for hematological traits in a large-scale swine F_2 resource population. *Front. Genet.* 9, 401. doi:10.3389/fgene.2018.00401

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

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* 46, 100–106. doi:10.1038/ng.2876

Yang, R., Guo, X., Zhu, D., Tan, C., Bian, C., 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, giab048–14. doi:10.1093/gigascience/giab048

Yang, T., Wang, Z., Miar, Y., Bruce, H., Zhang, C., and Plastow, G. (2017). A genome-wide association study of meat colour in commercial crossbred pigs. *Can. J. Anim. Sci.* 97, 4. doi:10.1139/cjas-2016-0248

Zhang, C., Wang, Z., Bruce, H., Kemp, R. A., Charagu, P., Miar, Y., et al. (2015). Genome-wide association studies (GWAS) identify a QTL close to PRKAG3 affecting meat pH and colour in crossbred commercial pigs. *BMC Genet.* 16, 33. doi:10.1186/s12863-015-0192-1

Zhang, F., Wang, Y., Mukiibi, R., Chen, L., Vinsky, M., lastow, G., et al. (2020). Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: Feed efficiency and component traits. *BMC Genomics* 21, 36. doi:10.1186/s12864-019-6362-1

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 25 October 2022
DOI 10.3389/fgene.2022.1025004

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Tao Luo,
Nanchang University, China
Woo-Sung Kwon,
Kyungpook National University, South Korea

*CORRESPONDENCE
Arumugam Kumaresan,
ogkumaresan@gmail.com

†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 22 August 2022
ACCEPTED 03 October 2022
PUBLISHED 25 October 2022

CITATION
Ebenezer Samuel King JP, Sinha MK, Kumaresan A, Nag P, Das Gupta M, Arul Prakash M, Talluri TR and Datta TK (2022), Cryopreservation process alters the expression of genes involved in pathways associated with the fertility of bull spermatozoa.
Front. Genet. 13:1025004.
doi: 10.3389/fgene.2022.1025004

Cryopreservation process alters the expression of genes involved in pathways associated with the fertility of bull spermatozoa

John Peter Ebenezer Samuel King[1†], Manish Kumar Sinha[1†], Arumugam Kumaresan[1*], Pradeep Nag[1], Mohua Das Gupta[1], Mani Arul Prakash[1], Thirumala Rao Talluri[1] and Tirtha Kumar Datta[2]

[1]Theriogenology Laboratory, Veterinary Gynaecology and Obstetrics, Southern Regional Station of ICAR-National Dairy Research Institute, Bengaluru, Karnataka, [2]ICAR-Central Institute for Research on Buffaloes, Hisar, Haryana

In bovines, cryopreserved semen is used for artificial insemination; however, the fertility of cryopreserved semen is far lower than that of fresh semen. Although cryopreservation alters sperm phenotypic characteristics, its effect on sperm molecular health is not thoroughly understood. The present study applied next-generation sequencing to investigate the effect of cryopreservation on the sperm transcriptomic composition of bull spermatozoa. While freshly ejaculated bull spermatozoa showed 14,280 transcripts, cryopreserved spermatozoa showed only 12,375 transcripts. Comparative analysis revealed that 241 genes were upregulated, 662 genes were downregulated, and 215 genes showed neutral expression in cryopreserved spermatozoa compared to fresh spermatozoa. Gene ontology analysis indicated that the dysregulated transcripts were involved in nucleic acid binding, transcription-specific activity, and protein kinase binding involving protein autophosphorylation, ventricular septum morphogenesis, and organ development. Moreover, the dysregulated genes in cryopreserved spermatozoa were involved in pathways associated with glycogen metabolism, MAPK signalling, embryonic organ morphogenesis, ectodermal placode formation, and regulation of protein auto-phosphorylation. These findings suggest that the cryopreservation process induced alterations in the abundance of sperm transcripts related to potential fertility-associated functions and pathways, which might partly explain the reduced fertility observed with cryopreserved bull spermatozoa.

KEYWORDS

cryopreservation, spermatozoa, pathways, transcripts, fertility, bovine

Introduction

Artificial insemination using cryopreserved semen is widely and routinely practiced for the genetic improvement of livestock. Ultralow freezing allows the preservation of semen for several years, which can be used later for artificial breeding (Barbas and Mascarenhas, 2009). Although the cryopreservation of sperm is an important process in assisted reproduction technologies, the fertility of cryopreserved spermatozoa is reportedly not as good as fresh spermatozoa (Watson, 1995; Bailey et al., 2000; Kadirvel et al., 2009). Several reasons are attributed to the decreased fertility of cryopreserved semen, primarily damage that occurs in spermatozoa during cryopreservation (cryodamage), which alters its fertilising potential (Cormier et al., 1997; Watson, 2000; Yeste, 2016). The process of cryopreservation results in the death of almost 50% of spermatozoa, while the remaining sperm population shows altered functional competencies (Kumaresan et al., 2011; Kumaresan et al., 2012; Singh et al., 2016; Kumaresan et al., 2017; Saraf et al., 2019; Rather et al., 2020; Vignesh et al., 2020; Nag et al., 2021), which might be linked to the reduced fertility of cryopreserved semen. However, the fertility of cryopreserved semen subjected to sperm selection methods (for the selection of viable, active, and phenotypically normal spermatozoa) was also not as good as that of fresh semen, although some improvement in fertility was observed (Said and Land, 2011; Marzano et al., 2020). Therefore, besides cryopreservation-induced sperm structural and functional alterations, other inherent factors in spermatozoa might also be altered during cryopreservation as cryopreserved spermatozoa with normal phenotypic characteristics also show altered fertilising potential compared to fresh spermatozoa (Kadirvel et al., 2009; Elango et al., 2021). Therefore, the assessment of the molecular alterations induced by the process of cryopreservation is essential for understanding the decreased fertility associated with cryopreserved semen.

Sperm was previously believed to deliver only the paternal DNA to the oocyte after fertilisation. However, several recent studies have demonstrated the role of sperm transcripts in male fertility (Paul et al., 2020; Prakash et al., 2021; Saraf et al., 2021) and subsequent embryonic development (Zhang et al., 2018). After fertilisation, spermatozoal mRNAs are transferred to the oocyte and play important roles in embryonic development, morphogenesis, and implantation (Ostermeier et al., 2004). Bull spermatozoa harbours a repertoire of transcripts (Prakash et al., 2021) that vary with semen quality and fertility (Karuthadurai et al., 2022). Earlier studies reported different transcriptomic profiles between epididymal and ejaculated spermatozoa, and also among different seasons (AL-Sahaf, 2012), indicating the influence of the micro-environment on sperm transcripts. Any changes in the transcript content of the sperm affect sperm properties such as motility, DNA intactness, and acrosome integrity (Bissonnette et al., 2009).

Although structural damage to sperm during cryopreservation has been studied extensively, the molecular alterations induced by the process of cryopreservation are not well documented. We hypothesised that the transcriptomic composition of sperm could be influenced by the process of cryopreservation, which might be associated with the reduced fertility of cryopreserved semen. Therefore, in the present study, we assessed the global transcriptomic composition of spermatozoa immediately after ejaculation and after cryopreservation using next-generation sequencing. We compared the transcriptomic profile of fresh spermatozoa to that of cryopreserved spermatozoa and identified common and dysregulated transcripts. Using functional annotation of these genes, we report the alterations in sperm transcripts and important pathways associated with fertility, which were induced by the cryopreservation process.

Materials and methods

Ethical approval statement

The current study was carried out at the Theriogenology Laboratory of the Southern Regional Station of ICAR-National Dairy Research Institute, Bengaluru, Karnataka. All the experiments and procedures performed in this study were approved by the Animal Ethical Committee of the institute (CPCSEA/IAEC/LA/SRS-ICAR-NDRI-2019/No.04).

Experimental bulls and sample preparation

This investigation was conducted on Holstein-Friesian crossbred bulls (n = 6; age 4–6 years). All experimental bulls had passed breeding soundness evaluations and were routinely used for artificial breeding. Ejaculates were collected using an artificial vagina; after preliminary evaluation, only ejaculates with minimum sperm concentrations of 600 million/ml, +3.0 mass activity (0–5 scale), ≥70% progressive individual motility, and <20% sperm abnormalities were utilised for further processing. The ejaculates were divided into two aliquots; one aliquot was used fresh while the other aliquot was subjected to cryopreservation as per the standard procedure. Briefly, the ejaculates were diluted using pre-warmed (34°C) Tris-egg yolk glycerol extender (20% egg yolk and 7% glycerol fractions) and then further processed for cryopreservation. The diluted semen was then filled and sealed in 0.25 ml mini straws (20×10^6 sperm per dose) using an automatic filling and sealing machine. The straws were then equilibrated in a cold handling cabinet (IMV Technologies, France) for 4 h at 4°C. Post-equilibration, the doses were loaded into a programmable Biofreezer (Digitcool, IMV Technologies, France) for cryopreservation as per standard protocol. After reaching −140°C, the straws were directly plunged into liquid nitrogen. After cryopreservation, the frozen semen was thawed at 37°C for 30 s

and used for further processing. Ejaculates from three bulls were pooled to obtain one representative sample. Therefore, we obtained two representative samples from six bulls for each condition (fresh and cryopreserved), which were individually subjected to transcriptomic analysis.

RNA extraction and synthesis of cDNA

Discontinuous Percoll gradients (90–45%) were used to fractionate pure sperm by eliminating epithelial cells and seminal plasma. The method described by Parthipan et al. (2015), with minor modifications was used for the extraction of total sperm RNA from fresh and cryopreserved samples using TRIzol (Ambion, Thermo Fisher Scientific, United States). The RNA was quantified using Nanodrop (ND-1000, Thermo-scientific, United States). For cDNA library preparation, RNA samples with 260/280 ratios of 1.85–2.0 were used. An initial concentration of 50–100 ng of RNA was used for cDNA synthesis using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, United States). The synthesised cDNA was stored at −20°C until further use.

Transcriptomics library preparation

Total RNA (1 μg) enriched for mRNA using the NEB Magnetic mRNA Isolation Kit (Illumina, United States). RNA library preparation was performed using the NEB Ultra II RNA library prep kit (Illumina, United States) and sequencing (paired-end technology) was performed on an Illumina NextSeq 500 instrument (Illumina, United States. The enriched mRNA was fragmented to 200 bp using fragmentation buffer. Complementary RNA sequence hybridisation was performed by adding random hexamer primers. Reverse transcriptase enzyme and dNTPs were used for synthesising the first strand of cDNA from fragments. DNA polymerase I and RNase H were used to convert the single-strand cDNA into double-stranded cDNA, which was purified using 1.8x AMPure beads. Adaptor-ligated cDNA was purified using AMPure beads and was enriched with specific primers for sequencing on the Illumina platforms.

RNA sequencing and data analysis

The Galaxy online server tool was used to analyse the sequences. The quality of the raw generated data was checked using the Fast QC program with a Phred quality score cut-off of Q30. The Cutadapt tool was used to remove adaptor sequences from the FASTQ files. HISAT2 (version2.1) was used to align the sequences to the reference genome. The samples were aligned to the reference genome of *Bos taurus* (version UCD 3.1.94). Cufflinks (version—2.2.1.2) was used to identify and estimate the abundance of the transcripts. After normalisation, the transcript expression levels were calculated as FPKM (fragments per kilobase of transcript per million mapped reads). Full-length transcript analysis was performed using the depth of coverage program version-0.0.2 in GATK.

Functional annotation and Gene Ontology analysis

Functional annotation and gene ontology (GO) analysis were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Bioinformatics Resources (v6.8) (https://david.ncifcrf.gov/). Molecular Function (MF), Biological Process (BP), cellular components (CC), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed to identify the genes with variations. For network analysis, genes related to sperm quality were selected from MF, BP, CC, and KEGG. ClueGo (Version 2.5.4) plugins in the open-source Cytoscape software (version 3.7.1) (Cluego.org) were used to identify the interactions of novel genes and their associated pathways.

Statistical analysis

DAVID was used to perform functional enrichment analysis of the Gene Ontology categories and KEGG pathway elements. The EASE score was calculated using the Fisher's exact *p*-value based on the input list of genes and genes participating in a certain pathway, resulting in a strongly enriched word. A *p*-value threshold of 0.01 was used to determine the EASE score cut-off. After enrichment analysis, terms with *p*-values <0.01 were identified as enriched terms. Thus, the lower the *p*-value, the more enriched the phrase. At least two genes were required for the examination of connection. The corrected *p*-values obtained after adjustments using the Benjamini and Bonferroni tests were defined as the false discovery rate (FDR). The processed reads were mapped to the reference genome for analysis of the raw transcriptomic data. The number of reads mapped to the exonic region indicated the gene expression. The read counts were used by DESeq to determine the differential expression based on the number of mapped reads. DESeq was used to create size factors (using the size factors function) and fit the data with a negative binomial distribution (using the nbinom test function) to compare the control (fresh) and treated (cryopreserved) groups and return log2-fold changes and significant *p*-values.

Results

Differentially regulated transcripts

The primary analysis of fresh and cryopreserved spermatozoa resulted in 14,280 and 12,375 transcripts, respectively (Supplementary File S1 and Supplementary Figure S1) indicating

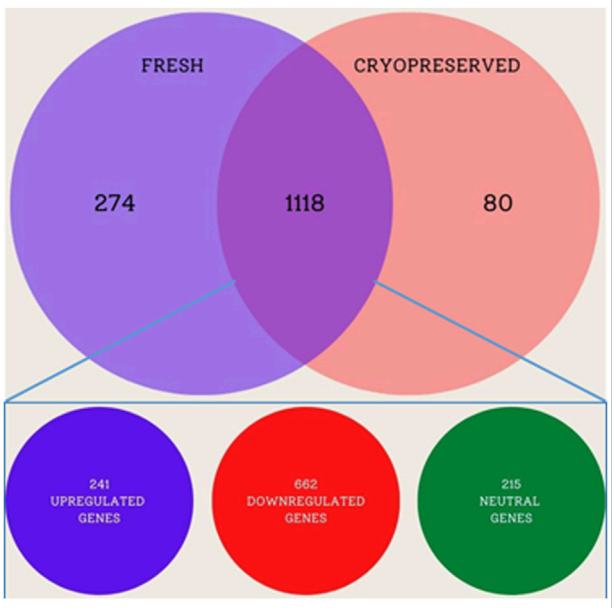


FIGURE 1
Venn diagram of the transcriptomic compositions of fresh and cryopreserved bull spermatozoa.

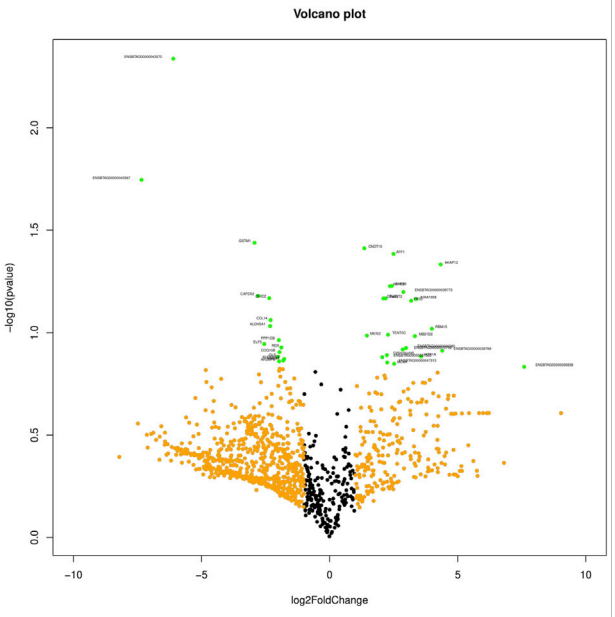


FIGURE 2
Volcano plot representing the gene log-fold changes and *p*-value. Green dots: significantly differentially expressed genes.

that the process of cryopreservation altered the sperm transcriptomic profile. After normalisation of the read counts, a total of 1118 genes were common between fresh and cryopreserved spermatozoa. In cryopreserved spermatozoa, 241 genes were upregulated, 662 genes were downregulated, and 215 genes were neutral (neither up- regulated nor downregulated) in expression compared to fresh spermatozoa (Figure 1). The genes that were upregulated or downregulated based on log2 fold-change values are shown in a volcano plot (Figure 2). A heat map of the top 20 upregulated and top 20 downregulated genes in the cryopreserved sample (Figure 3) showed the genes with highly dysregulated expression. *C11H9orf50*, *ARL2BP*, and *HSPB1* were the most downregulated, while *RPRD2*, *ENSBTAG00000030892*, and *ENSBTAG00000030838* were the most upregulated in cryopreserved spermatozoa.

Gene Ontology

The dysregulated genes were selected for Gene Ontology (GO) analysis based on a log2 fold-change cut-off of ±1. These dysregulated genes were then subjected to DAVID analysis to annotate GO terms including molecular function (MF), biological process (BP), and cellular components (CC). A total of 54 upregulated and 411 downregulated genes were annotated for GO analysis. From these, the top MF, BP, and CC for upregulated and downregulated genes (Figures 4A,B) (Supplementary File S2) were selected based on the counts of genes involved in each GO term. The important upregulated genes in the cryopreserved spermatozoa were involved in specific DNA binding, nucleic acid binding, transcription-specific activity, protein kinase binding, core promotor sequence-specific DNA binding, and inward rectifier potassium channel activity. These were mostly localised in the nucleus, cytosol, and potassium channel complex and were involved in biological processes such as the regulation of transcription from the RNA polymerase II promotor, the regulation of neuronal differentiation, the positive regulation of protein autophosphorylation, ventricular septum morphogenesis, organ development, Rap protein signal transduction, neuroblast proliferation, and negative regulation of myeloid cell differentiation. The downregulated genes were involved in poly(A) RNA binding, structural constituents of ribosomes, DNA binding, protein binding, RNA binding, ubiquitin protein ligase binding, and nucleotide binding.

KEGG pathway and network analysis

KEGG pathway enrichment analysis of the downregulated genes (Supplementary File S3) in cryopreserved spermatozoa revealed the involvement of genes in important pathways related to ribosomes (50 genes), Huntington's disease (17 genes), thermogenesis (16 genes), spliceosomes (16 genes), and endocytosis (16 genes). KEGG pathway enrichment analysis of the upregulated genes (Supplementary File S3) in cryopreserved spermatozoa revealed the involvement of genes in pathways related

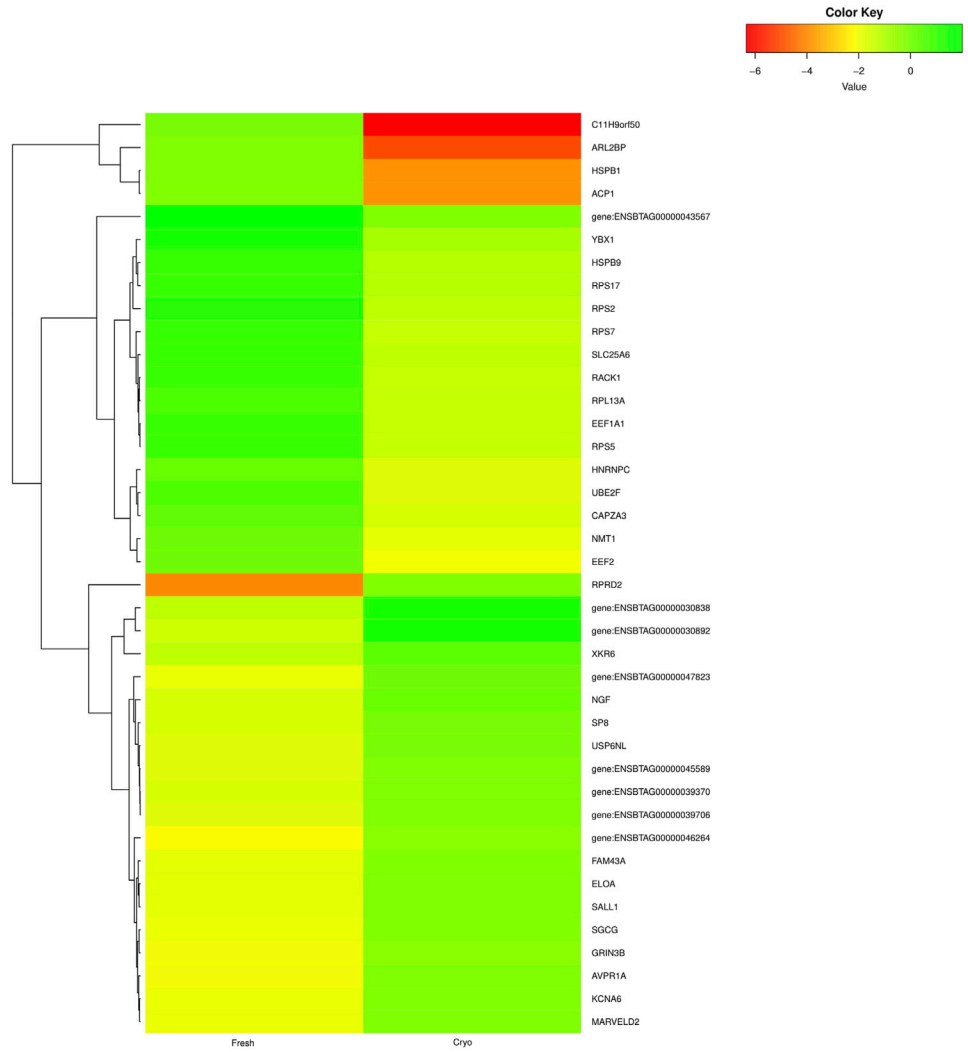


FIGURE 3
Heat map representing the top 20 upregulated (green) and top 20 downregulated (red and yellow) genes in cryopreserved spermatozoa compared to fresh spermatozoa.

to the neurotrophin signalling pathway (4 genes), transcriptional misregulation in cancer (4 genes), and neuroactive ligand-receptor interaction (5 genes). A network analysis of GO terms and KEGG pathways for upregulated and downregulated genes related to fertility was also performed. The network analysis of downregulated (Figure 5) genes in cryopreserved spermatozoa showed their involvement in pathways such as glycogen metabolism, the MAPK signalling pathway, fluid shear and stress atherosclerosis, cytoskeletal protein binding, the oestrogen signalling pathway, amphetamine addiction, transcription factor binding, Huntington's disease, ubiquitin protein ligase binding, translation factor activity RNA binding, mRNA binding, transcription regulation region sequence-specific DNA binding, RNA polymerase II core promoter sequence-specific DNA binding, oxidoreductase activity, and glycolysis/gluconeogenesis. The network analysis of upregulated (Figure 6) genes in cryopreserved spermatozoa showed their involvement in pathways such as embryonic organ morphogenesis, ectodermal placode formation, regulation of protein auto-phosphorylation, and death receptor binding (Supplementary File S4).

Discussion

The process of cryopreservation alters sperm phenome and functions, which affect the quality and fertility of spermatozoa. Recent studies have shown that sperm transcripts are associated with sperm quality and fertility, as well as embryonic

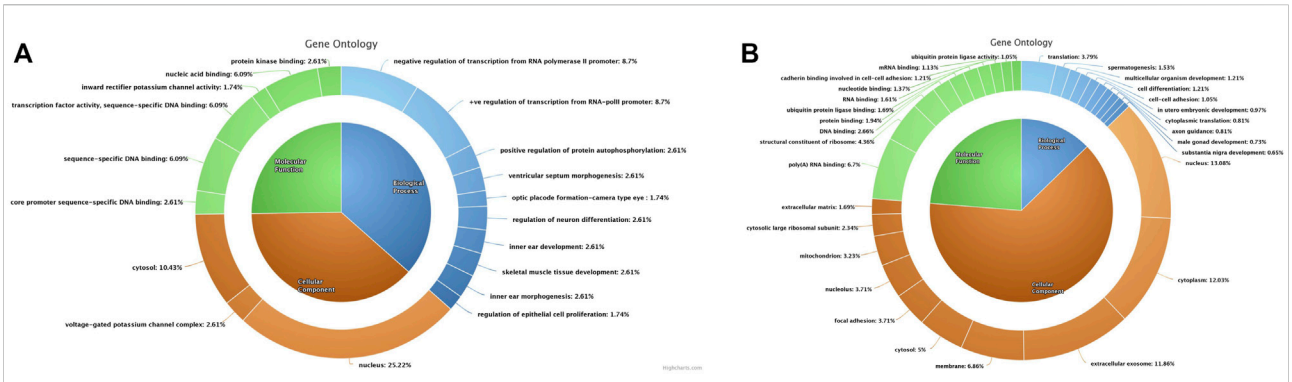


FIGURE 4
Gene Ontology chart of differentially expressed genes between fresh and cryopreserved spermatozoa. All three components (molecular function, biological process, and cellular functions) are represented in the donut chart. **(A)** Upregulated genes. **(B)** Downregulated genes.

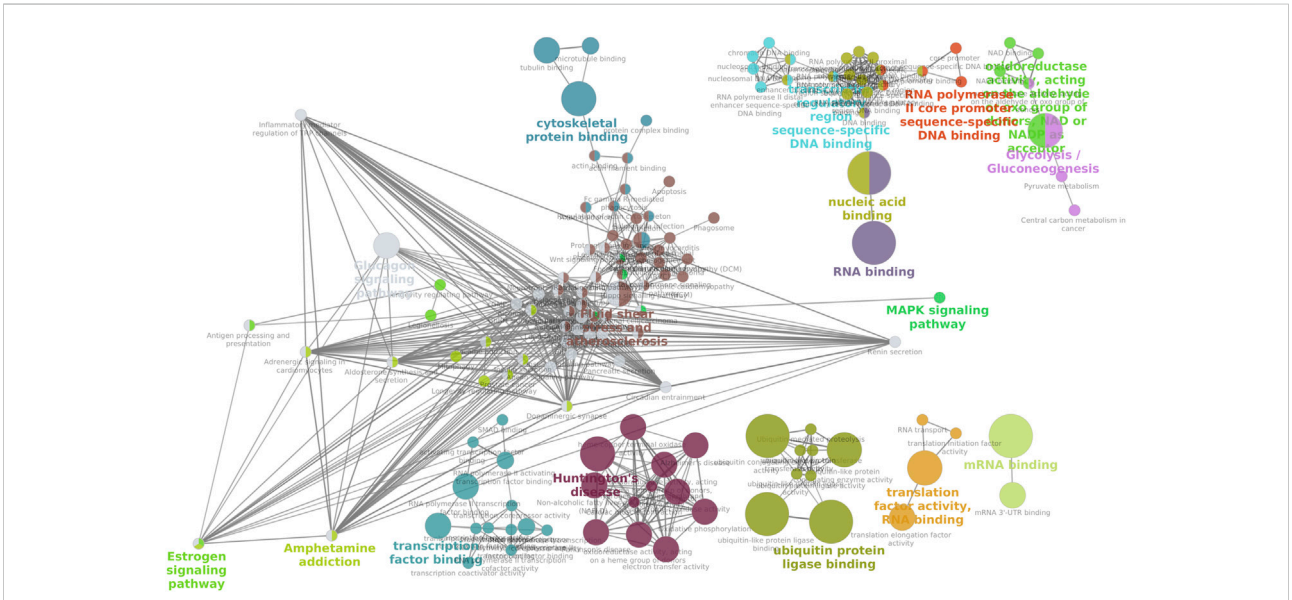


FIGURE 5
Network analysis of all downregulated genes in the cryopreserved samples. MAPK signalling pathway, fluid shear and stress atherosclerosis, cytoskeletal protein binding, and the oestrogen signalling pathway are mapped.

development (Zhang et al., 2018; Paul et al., 2020; Prakash et al., 2021). Since the process of cryopreservation alters sperm structure, this process may also affect the transcriptomic composition of sperm. The results of the present study demonstrated the significant alterations in the transcriptomic composition of bull sperm by the process of cryopreservation, which might explain the reduced fertility of cryopreserved semen.

The mature spermatozoal RNA undergoes certain cellular modifications during spermiogenesis. The required mRNAs are then packed in spermatozoa before it reaches the transcriptionally nascent stage in the epididymis (Grunewald et al., 2005; Vijayalakshmy et al., 2018; Paul et al., 2020) for delivery to the oocyte after fertilisation. We observed highly downregulated expression of *C11H9orf50*, *ARL2BP*, and *HSPB1* and highly upregulated expression of *RPRD2*, *ENSBTAG00000030892*, and *ENSBTAG00000030838* in cryopreserved spermatozoa. *ARL2BP* is required for ciliary microtubule structure. A mouse *ARL2BP* knockout showed structural impairments such as abnormal head and misassembled tail of sperm (Moye et al., 2019). *HSBP1* is

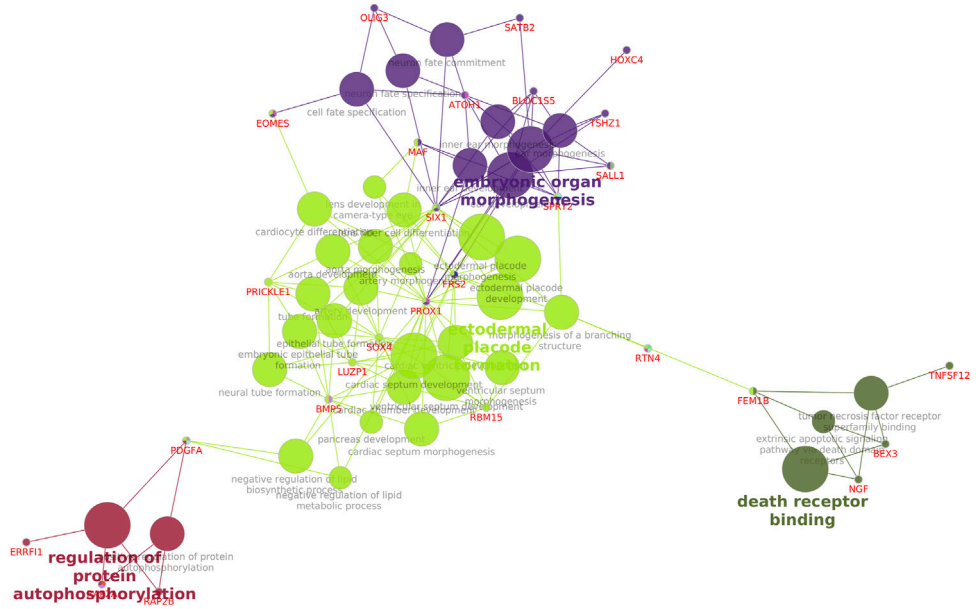


FIGURE 6
Network analysis of all upregulated genes in the cryopreserved samples. Embryonic organ morphogenesis, ectodermal placode formation, regulation of protein auto-phosphorylation, and death receptor binding are mapped.

responsible for organising the muscle cytoskeleton. *HSBP1* (also known as *HSP27*) is important for muscle formation. A mouse *HSBP1* knockout showed destructured myofibrils and increased gaps between myofibrils (Kammoun et al., 2016). *RPRD2* is an RNA polymerase II interacting protein that co-purifies with RPRD1A, RPRD1B, and RPRD2 and contains serine and proline-rich regions (Ni et al., 2011). The proline-rich domains function as docking sites for signalling protein modules and play important roles in protein-protein interactions (Chin et al., 1997; Elias et al., 2020). The overexpression of this gene might alter transcription, thus hampering protein interactions.

The results of the network analysis showed downregulation of cGMP/PKG signalling, which is responsible for sperm capacitation through Ca^{2+} and tyrosine phosphorylation in the presence of C-type natriuretic peptide (CNP) (Wu et al., 2019). CNP, which is localised in the acrosomal region of the sperm head and tail, has a dose-dependent role in acrosome reaction and motility (Xia et al., 2016). CNP might be affected by the process of cryopreservation, as fast freezing increases the ice-liquid interface with protein molecules, thereby increasing protein damage (Cao et al., 2003), which in turn affects sperm motility and acrosome reaction. The MAPK signalling pathway is involved in sperm development and function through spermatogenesis and fertilising potential. In mature spermatozoa, the MAPK signalling pathway plays an important role in acrosome reaction, hyperactivation, and motility (Almog and Naor, 2010; Ebenezer Samuel King et al., 2022). We also observed the downregulation of cytoskeletal protein binding pathways in the cryopreserved sample. These pathways are linked to tubulin binding. Actin, along with the cytoskeletal protein tubulin, is involved in the regulation of sperm motility, capacitation, and acrosome reaction (Salvolini et al., 2013). In normal spermatogenesis, oestrogen signalling through ESR1 is essential. Human spermatozoa contain functional aromatase, which is expressed even after ejaculation. Aromatase, in the presence of oestrogen receptor, has functions including sperm mobility and fertilising ability (Carreau et al., 2011; Cooke and Walker, 2022). The ubiquitin ligase complex, which is required for caspase regulation, was also downregulated in cryopreserved spermatozoa. Caspase is an important protease required for apoptosis. The activation of caspase during sperm differentiation is regulated by the ubiquitin ligase complex (Arama et al., 2007). mRNA binding proteins are needed for the differentiation of spermatids into spermatozoa by modulating the expression of specific mRNAs (Wishart and Dixon, 2002). All the pathways important for spermatozoa function, especially motility, capacitation, and acrosome reaction, were downregulated in the cryopreserved spermatozoa in the present study.

Genes involved in the embryo organ morphogenesis pathway were upregulated in cryopreserved spermatozoa. This pathway is required for tissue and organ development to perform special functions during the embryonic stage. Among these, *SPRY2* is required for the normal development of genitalia; *TSHZ1* is required for the development of the soft palate, middle ear, and axial skeleton; and *SATB2* codes for the jaw development (Britanova et al., 2006; Coré et al., 2007; Ching et al., 2014). The overexpression of these genes might lead to abnormal embryo development. *PRICKLE 1*

is involved in ectodermal placode formation and also plays a role in nervous system development and nerve cell movement. A polymorphism in *PRICKLE 1* reportedly affected acrosome integrity and DNA quality in cryopreserved spermatozoa (Mańkowska et al., 2020). The regulation of protein autophosphorylation, one of the upregulated pathways in cryopreserved spermatozoa, is involved in the positive regulation of protein autophosphorylation. *RAP2B* in protein autophosphorylation codes for miR-205; miR-205 overexpression inhibits the PI3K/AKT signalling pathway (Cui et al., 2020). Earlier studies also showed that cryopreservation suppressed the expression of important pathways and genes in sperm, causing DNA fragmentation and morphological deformation (Hossen et al., 2021). Moreover, sperm stored in liquid nitrogen for longer times showed significantly increased abnormalities (Malik et al., 2015). A study in zebrafish showed that cryopreservation was the major cause of genetic and epigenetic changes in germ cells (Riesco and Robles, 2013). Our findings demonstrated the alteration of the transcriptional abundances of genes involved in important pathways in sperm function and fertility by the process of cryopreservation. These findings open new avenues for targeted studies on individual pathways and the development of ameliorative measures to minimise the effect of cryopreservation on the molecular health of sperm.

Conclusion

In conclusion, the transcriptional abundances of many genes were altered in cryopreserved spermatozoa, leading to changes in pathways important for sperm function and fertility. Earlier studies also indicated that the cryopreservation process altered the sperm expression levels of genes crucial for fertilisation and early embryo development (Valcarce et al., 2013), consistent with our observations. The cryopreservation process induced alterations in the abundance of sperm transcripts and potential fertility-associated pathways, including glycogen metabolism, cGMP/PKG signalling, MAPK signalling, and the ubiquitin ligase complex, which could be a possible explanation for the reduced fertility observed in cryopreserved bull spermatozoa.

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) are as shown below: https://www.ncbi.nlm.nih.gov/, PRJNA847399; https://www.ncbi.nlm.nih.gov/, PRJNA516089.

Ethics statement

The animal study was reviewed and approved by the Animal Ethical Committee of the institute.

Author contributions

JE—investigation, methodology, visualisation, and writing–original draft. MS: investigation, methodology, visualisation, writing - original draft. AK—conceptualisation, writing—review and editing, resources, supervision. PN–methodology, review, editing, and referencing. MD—formal investigation. MA—formal investigation. TT—review and editing. TD—review and editing, resources.

Funding

The present work was funded by a Bill & Melinda Gates Foundation project entitled “Molecular markers for improving reproduction in cattle and buffaloes” (grant number OPP1154401).

Acknowledgments

We thank the Director of the ICAR-National Dairy Research Institute, Karnal, and the Head, SRS of ICAR-NDRI, Bengaluru, for providing the facilities to conduct this study. The authors acknowledge and thank the Kerala Livestock Development Board in Kerala for providing the semen samples.

Conflict of interest

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

Publisher’s note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1025004/full#supplementary-material

References

Al-Sahaf, M. M. (2012). Monthly changes in testes and epididymis measurements with some semen characteristics of tail epididymis for Iraqi buffalo: Mohammed Mehdi AL-Sahaf and Najlaa Sami Ibrahim. *Iraqi J. Vet. Med.* 36, 204–208. doi:10.30539/iraqijvm.v36i2.497

Almog, T., and Naor, Z. (2010). The role of Mitogen activated protein kinase (MAPK) in sperm functions. *Mol. Cell. Endocrinol.* 314, 239–243. doi:10.1016/j.mce.2009.05.009

Arama, E., Bader, M., Rieckhof, G. E., and Steller, H. (2007). A ubiquitin ligase complex regulates caspase activation during sperm differentiation in Drosophila. *PLoS Biol.* 5, e251. doi:10.1371/journal.pbio.0050251

Bailey, J. L., Bilodeau, J. F., and Cormier, N. (2000). Semen cryopreservation in domestic animals: A damaging and capacitating phenomenon. *J. Androl.* 21, 1–7.

Barbas, J. P., and Mascarenhas, R. D. (2009). Cryopreservation of domestic animal sperm cells. *Cell Tissue Bank.* 10, 49–62. doi:10.1007/s10561-008-9081-4

Bissonnette, N., Lévesque-Sergerie, J-P., Thibault, C., and Boissonneault, G. (2009). Spermatozoal transcriptome profiling for bull sperm motility: A potential tool to evaluate semen quality. *Reproduction* 138, 65–80. doi:10.1530/REP-08-0503

Britanova, O., Depew, M. J., Schwark, M., Thomas, B. L., Miletich, I., Sharpe, P., et al. (2006). Satb2 haploinsufficiency phenocopies 2q32-q33 deletions, whereas loss suggests a fundamental role in the coordination of jaw development. *Am. J. Hum. Genet.* 79, 668–678. doi:10.1086/508214

Cao, E., Chen, Y., Cui, Z., and Foster, P. R. (2003). Effect of freezing and thawing rates on denaturation of proteins in aqueous solutions. *Biotechnol. Bioeng.* 82, 684–690. doi:10.1002/bit.10612

Carreau, S., Bois, C., Zanatta, L., Silva, F. R. M. B., Bouraima-Lelong, H., and Delalande, C. (2011). Estrogen signaling in testicular cells. *Life Sci.* 89, 584–587. doi:10.1016/j.lfs.2011.06.004

Chin, K-C., Li, G. G-X., and Ting, J. P-Y. (1997). Importance of acidic, proline/serine/threonine-rich, and GTP binding regions in the major histocompatibility complex class II transactivator: Generation of transdominant-negative mutants. *Proc. Natl. Acad. Sci. U. S. A.* 94, 2501–2506. doi:10.1073/pnas.94.6.2501

Ching, S. T., Cunha, G. R., Baskin, L. S., Basson, M. A., and Klein, O. D. (2014). Coordinated activity of Spry1 and Spry2 is required for normal development of the external genitalia. *Dev. Biol.* 386, 1–11. doi:10.1016/j.ydbio.2013.12.014

Cooke, P. S., and Walker, W. H. (2022). Nonclassical androgen and estrogen signaling is essential for normal spermatogenesis. *Semin. Cell Dev. Biol.* 121, 71–81. doi:10.1016/j.semcdb.2021.05.032

Coré, N., Caubit, X., Metchat, A., Boned, A., Djabali, M., and Fasano, L. (2007). Tshz1 is required for axial skeleton, soft palate and middle ear development in mice. *Dev. Biol.* 308, 407–420. doi:10.1016/j.ydbio.2007.05.038

Cormier, N., Sirard, M. A., and Bailey, J. L. (1997). Premature capacitation of bovine spermatozoa is initiated by cryopreservation. *J. Androl.* 18, 461–468.

Cui, Y., Chen, R., Ma, L., Yang, W., Chen, M., Zhang, Y., et al. (2020). miR-205 expression elevated with EDS treatment and induced leydig cell apoptosis by targeting RAP2B via the PI3K/AKT signaling pathway. *Front. Cell Dev. Biol.* 8, 448. doi:10.3389/fcell.2020.00448

Ebenezer Samuel King, J. P., Kumaresan, A., Talluri, T. R., Sinha, M. K., Raval, K., Nag, P., et al. (2022). Genom-wide analysis identifies single nucleotide polymorphism variations and altered pathways associated with poor semen quality in breeding bulls. *Reprod. Domest. Anim.* doi:10.1111/rda.14185

Elango, K., Kumaresan, A., Ashokan, M., Karuthadurai, T., Nag, P., Bhaskar, M., et al. (2021). Dynamics of mitochondrial membrane potential and DNA damage during cryopreservation of cattle and buffalo bull spermatozoa. *Indian J. Anim. Sci.* 91, 9–14.

Elias, R. D., Ma, W., Ghirlando, R., Schwieters, C. D., Reddy, V. S., and Deshmukh, L. (2020). Proline-rich domain of human ALIX contains multiple TSG101-UEV interaction sites and forms phosphorylation-mediated reversible amyloids. *Proc. Natl. Acad. Sci. U. S. A.* 117, 24274–24284. doi:10.1073/pnas.2010635117

Grunewald, S., Paasch, U., Glander, H-J., and Anderegg, U. (2005). Mature human spermatozoa do not transcribe novel RNA. *Andrologia* 37, 69–71. doi:10.1111/j.1439-0272.2005.00656.x

Hossen, S., Sukhan, Z. P., Cho, Y., and Kho, K. H. (2021). Effects of cryopreservation on gene expression and post thaw sperm quality of pacific abalone, Haliotis discus hannai. *Front. Mar. Sci.* 8, 652390. doi:10.3389/fmars.2021.652390

Kadirvel, G., Kumar, S., Kumaresan, A., and Kathiravan, P. (2009). Capacitation status of fresh and frozen-thawed buffalo spermatozoa in relation to cholesterol level, membrane fluidity and intracellular calcium. *Anim. Reprod. Sci.* 116, 244–253. doi:10.1016/j.anireprosci.2009.02.003

Kammoun, M., Picard, B., Astruc, T., Gagaoua, M., Aubert, D., Bonnet, M., et al. (2016). The invalidation of HspB1 gene in mouse alters the ultrastructural phenotype of muscles. *PLoS ONE* 11, e0158644. doi:10.1371/journal.pone.0158644

Karuthadurai, T., Das, D. N., Kumaresan, A., Sinha, M. K., Kamaraj, E., Nag, P., et al. (2022). Sperm transcripts associated with odorant binding and olfactory transduction pathways are altered in breeding bulls producing poor-quality semen. *Front. Vet. Sci.* 9, 799386. doi:10.3389/fvets.2022.799386

Kumaresan, A., Johannisson, A., Al-Essawe, E. M., and Morrell, J. M. (2017). Sperm viability, reactive oxygen species, and DNA fragmentation index combined can discriminate between above- and below-average fertility bulls. *J. Dairy Sci.* 100, 5824–5836. doi:10.3168/jds.2016-12484

Kumaresan, A., Johannisson, A., Humblot, P., and Bergqvist, A-S. (2012). Oviductal fluid modulates the dynamics of tyrosine phosphorylation in cryopreserved boar spermatozoa during capacitation. *Mol. Reprod. Dev.* 79, 525–540. doi:10.1002/mrd.22058

Kumaresan, A., Siqueira, A. P., Hossain, M. S., and Bergqvist, A. S. (2011). Cryopreservation-induced alterations in protein tyrosine phosphorylation of spermatozoa from different portions of the boar ejaculate. *Cryobiology* 63, 137–144. doi:10.1016/j.cryobiol.2011.08.002

Malik, A., Laily, M., and Zakir, M. I. (2015). Effects of long term storage of semen in liquid nitrogen on the viability, motility and abnormality of frozen thawed Frisian Holstein bull spermatozoa. *Asian Pac. J. Reproduction* 4, 22–25. doi:10.1016/S2305-0500(14)60052-X

Mańkowska, A., Brym, P., Paukszto, L., Jastrzębski, J. P., and Fraser, L. (2020). Gene polymorphisms in boar spermatozoa and their associations with post-thaw semen quality. *Int. J. Mol. Sci.* 21, 1902. doi:10.3390/ijms21051902

Marzano, G., Chiriacò, M. S., Primiceri, E., Dell'Aquila, M. E., Ramalho-Santos, J., Zara, V., et al. (2020). Sperm selection in assisted reproduction: A review of established methods and cutting-edge possibilities. *Biotechnol. Adv.* 40, 107498. doi:10.1016/j.biotechadv.2019.107498

Moye, A. R., Bedoni, N., Cunningham, J. G., Sanzhaeva, U., Tucker, E. S., Mathers, P., et al. (2019). Mutations in ARL2BP, a protein required for ciliary microtubule structure, cause syndromic male infertility in humans and mice. *PLoS Genet.* 15, e1008315. doi:10.1371/journal.pgen.1008315

Nag, P., Kumaresan, A., Akshaya, S., Manimaran, A., Rajendran, D., Paul, N., et al. (2021). Sperm phenotypic characteristics and oviduct binding ability are altered in breeding bulls with high sperm DNA fragmentation index. *Theriogenology* 172, 80–87. doi:10.1016/j.theriogenology.2021.06.006

Ni, Z., Olsen, J. B., Guo, X., Zhong, G., Ruan, E. D., Marcon, E., et al. (2011). Control of the RNA polymerase II phosphorylation state in promoter regions by CTD interaction domain-containing proteins RPRD1A and RPRD1B. *Transcription* 2, 237–242. doi:10.4161/trns.2.5.17803

Ostermeier, G. C., Miller, D., Huntriss, J. D., Diamond, M. P., and Krawetz, S. A. (2004). Reproductive biology: Delivering spermatozoan RNA to the oocyte. *Nature* 429, 154. doi:10.1038/429154a

Parthipan, S., Selvaraju, S., Somashekar, L., Kolte, A. P., Arangasamy, A., and Ravindra, J. P. (2015). Spermatozoa input concentrations and RNA isolation methods on RNA yield and quality in bull (*Bos taurus*). *Anal. Biochem.* 482, 32–39. doi:10.1016/j.ab.2015.03.022

Paul, N., Kumaresan, A., Das Gupta, M., Nag, P., Guvvala, P. R., Kuntareddi, C., et al. (2020). Transcriptomic profiling of buffalo spermatozoa reveals dysregulation of functionally relevant mRNAs in low-fertile bulls. *Front. Vet. Sci.* 7, 609518. doi:10.3389/fvets.2020.609518

Prakash, M. A., Kumaresan, A., Ebenezer Samuel King, J. P., Nag, P., Sharma, A., Sinha, M. K., et al. (2021). Comparative transcriptomic analysis of spermatozoa from high- and low-fertile crossbred bulls: Implications for fertility prediction. *Front. Cell Dev. Biol.* 9, 647717. doi:10.3389/fcell.2021.647717

Rather, H. A., Kumaresan, A., Nag, P., Kumar, V., Nayak, S., Batra, V., et al. (2020). Spermatozoa produced during winter are superior in terms of phenotypic characteristics and oviduct explants binding ability in the water buffalo (Bubalus bubalis). *Reprod. Domest. Anim.* 55, 1629–1637. doi:10.1111/rda.13824

Riesco, M. F., and Robles, V. (2013). Cryopreservation causes genetic and epigenetic changes in zebrafish genital ridges. *PLoS ONE* 8, e67614. doi:10.1371/journal.pone.0067614

Said, T. M., and Land, J. A. (2011). Effects of advanced selection methods on sperm quality and ART outcome: A systematic review. *Hum. Reprod. Update* 17, 719–733. doi:10.1093/humupd/dmr032

Salvolini, E., Buldreghini, E., Lucarini, G., Vignini, A., Lenzi, A., Di Primio, R., et al. (2013). Involvement of sperm plasma membrane and cytoskeletal proteins in human male infertility. *Fertil. Steril.* 99, 697–704. doi:10.1016/j.fertnstert.2012.10.042

Saraf, K. K., Kumaresan, A., Sinha, M. K., and Datta, T. K. (2021). Spermatozoal transcripts associated with oxidative stress and mitochondrial membrane potential differ between high- and low-fertile crossbred bulls. *Andrologia* 53, e14029. doi:10.1111/and.14029

Saraf, K. K., Singh, R. K., Kumaresan, A., Nayak, S., Chhillar, S., Lathika, S., et al. (2019). Sperm functional attributes and oviduct explant binding capacity differs between bulls with different fertility ratings in the water buffalo (Bubalus bubalis). *Reprod. Fertil. Dev.* 31, 395–403. doi:10.1071/RD17452

Singh, R. K., Kumaresan, A., Chhillar, S., Rajak, S. K., Tripathi, U. K., Nayak, S., et al. (2016). Identification of suitable combinations of *in vitro* sperm-function test for the prediction of fertility in buffalo bull. *Theriogenology* 86, 2263–2271. doi:10.1016/j.theriogenology.2016.07.022

Valcarce, D. G., Cartón-García, F., Herráez, M. P., and Robles, V. (2013). Effect of cryopreservation on human sperm messenger RNAs crucial for fertilization and early embryo development. *Cryobiology* 67, 84–90. doi:10.1016/j.cryobiol.2013.05.007

Vignesh, K., Murugavel, K., Antoine, D., Prakash, M. A., Saraf, K. K., Nag, P., et al. (2020). The proportion of tyrosine phosphorylated spermatozoa in cryopreserved semen is negatively related to crossbred bull fertility. *Theriogenology* 149, 46–54. doi:10.1016/j.theriogenology.2020.03.020

Vijayalakshmy, K., Kumar, D., Virmani, M., Jacob, N., and Kumar, P. (2018). Sperm transcriptomics: An emerging technique to assess male fertility. *Int. J. Curr. Microbiol. Appl. Sci.* 7, 1188–1200. doi:10.20546/ijcmas.2018.709.141

Watson, P. F. (1995). Recent developments and concepts in the cryopreservation of spermatozoa and the assessment of their post-thawing function. *Reprod. Fertil. Dev.* 7, 871–891. doi:10.1071/rd9950871

Watson, P. F. (2000). The causes of reduced fertility with cryopreserved semen. *Anim. Reprod. Sci.* 60–61, 481–492. doi:10.1016/s0378-4320(00)00099-3

Wishart, M. J., and Dixon, J. E. (2002). The archetype STYX/dead-phosphatase complexes with a spermatid mRNA-binding protein and is essential for normal sperm production. *Proc. Natl. Acad. Sci. U. S. A.* 99, 2112–2117. doi:10.1073/pnas.251686198

Wu, K., Mei, C., Chen, Y., Guo, L., Yu, Y., and Huang, D. (2019). C-type natriuretic peptide regulates sperm capacitation by the cGMP/PKG signalling pathway via Ca2+ influx and tyrosine phosphorylation. *Reprod. Biomed. Online* 38, 289–299. doi:10.1016/j.rbmo.2018.11.025

Xia, H., Chen, Y., Wu, K-J., Zhao, H., Xiong, C-L., and Huang, D-H. (2016). Role of C-type natriuretic peptide in the function of normal human sperm. *Asian J. Androl.* 18, 80–84. doi:10.4103/1008-682X.150254

Yeste, M. (2016). Sperm cryopreservation update: Cryodamage, markers, and factors affecting the sperm freezability in pigs. *Theriogenology* 85, 47–64. doi:10.1016/j.theriogenology.2015.09.047

Zhang, K., Wang, H., Rajput, S. K., Folger, J. K., and Smith, G. W. (2018). Characterization of H3.3 and HIRA expression and function in bovine early embryos. *Mol. Reprod. Dev.* 85, 106–116. doi:10.1002/mrd.22939

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 31 October 2022
DOI 10.3389/fgene.2022.988488

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Tsukasa Fukunaga,
Waseda University, Japan
Luis Javier Chueca,
University of the Basque Country, Spain

*CORRESPONDENCE
Vinod Kumar,
vinodk@kisr.edu.kw

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

RECEIVED 07 July 2022
ACCEPTED 06 October 2022
PUBLISHED 31 October 2022

CITATION
Karam Q, Kumar V, Shajan AB, Al-Nuaimi S, Sattari Z and El-Dakour S (2022), De-novo genome assembly and annotation of sobaity seabream Sparidentex hasta.
Front. Genet. 13:988488.
doi: 10.3389/fgene.2022.988488

De-novo genome assembly and annotation of sobaity seabream *Sparidentex hasta*

Qusaie Karam[1], Vinod Kumar[2]*, Anisha B. Shajan[2], Sabeeka Al-Nuaimi[1], Zainab Sattari[3] and Saleem El-Dakour[3]

[1]Crises Management and Decision Support Program, Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, Kuwait City, Kuwait, [2]Biotechnology Program, Environment and Life Sciences Research Center, Kuwait Institute ForScientific Research, Kuwait City, Kuwait, [3]Aquaculture Program, Environment and Life Sciences Research Center, Kuwait Institute ForScientific Research, Kuwait City, Kuwait

Sparidentex hasta (Valenciennes, 1830) of the Sparidae family, is an economically important fish species. However, the genomic studies on *S. hasta* are limited due to the absence of its complete genome. The goal of the current study was to sequence, assemble, and annotate the genome of *S. hasta* that will fuel further research related to this seabream. The assembled draft genome of *S. hasta* was 686 Mb with an N50 of 80 Kb. The draft genome contained approximately 22% repeats, and 41,201 genes coding for 44,555 transcripts. Furthermore, the assessment of the assembly completeness was estimated based on the detection of ~93% BUSCOs at the protein level and alignment of >99% of the filtered reads to the assembled genome. Around 68% of the predicted proteins (n = 30,545) had significant BLAST matches, and 30,473 and 13,244 sequences were mapped to Gene Ontology annotations and different enzyme classes, respectively. The comparative genomics analysis indicated *S. hasta* to be closely related to *Acanthopagrus latus*. The current assembly provides a solid foundation for future population and conservation studies of *S. hasta* as well as for investigations of environmental adaptation in Sparidae family of fishes. Value of the Data: This draft genome of *S. hasta* would be very applicable for molecular characterization, gene expression studies, and to address various problems associated with pathogen-associated immune response, climate adaptability, and comparative genomics. The accessibility of the draft genome sequence would be useful in understanding the pathways and functions at the molecular level, which may further help in improving the economic value and their conservation.

KEYWORDS

draft genome, fisheries and aquaculture, food security, Kuwait, assembly and annotation

Introduction

Sobaity seabream, *Sparidentex hasta* (Valenciennes, 1830) belongs to the Sparidae family, which comprises 35 genera, 132 species, and 10 subspecies (de la Herran et al., 2001). The species has a wide geographic distribution extending from the Arabian Gulf to the sea of Oman and the western Indian Ocean, and the Indian coasts (Carpenter et al., 1997). *S. hasta* is recognized as one of the most promising species for aquaculture, because of its good adaptation to captivity, rapid growth, and high market price. Further, it is of high economic significance in Kuwait and the Arabian Gulf regions.

The anthropogenic and fishing activities around the coastal regions are affecting marine fauna including the population of many commercially important fish species (Bukola et al., 2015). However, genomics and molecular biology research on Sobaity seabream is limited due to the absence of its complete genome sequence. The DNA barcoding of several commercial seabreams including *S. hasta* was reported (Al-Zaidan et al., 2021). Most of the other studies on *S. hasta* are focused on the dietary effects of different feed combinations on *S. hasta* (Hossain et al., 2017; Yaghoubi et al., 2018; Hekmatpour et al., 2019). Furthermore, a study on the response of *S. hasta* larvae to the toxicity of dispersed and undispersed crude oil was reported (Karam et al., 2021).

There is an increasing demand for fish in Kuwait as fisheries only fulfill about 30% of local fish demand, as the other 70% is met through imports. However, there is a global decline in fisheries (Hossain et al., 2017) and to compensate for this decline and to assure future food security in Kuwait, aquaculture technologies were developed for *S. hasta* to fulfill the demands of the local market (Teng et al., 1987; Abdullah et al., 1989). Sobaity was chosen to be the first candidate species for commercial production in Kuwait because of its survival capability and tolerance in captivity (Al-Abdul-Elah et al., 2010; Torfi Mozanzadeh et al., 2017) and it is the second most favorable commercial seabream species in Kuwait after the yellowfin seabream (*Acanthopagrus latus*) (Al-Zaidan et al., 2021). The selection of Sobaity for aquaculture in the early 80's was primarily attributed to its ability to spawn in captivity, its tolerance to different culture conditions, and its fast growth rate (Yousif et al., 2003). *S. hasta* can exercise a wide range of tolerance to changes in water quality parameters such as dissolved oxygen, temperature, pH, and salinity which are reflective of natural ambient conditions. However, extreme and abrupt changes in those environmental parameters can result in the deteriorating health of juvenile Sobaity in culture tanks (European Food Safety Authority, 2008; Zainal and Altuama, 2020).

Also, Sobaity is sought for its nutritional value as a healthy seafood commodity. Fishes containing a certain type of fatty acids are known to reduce the risk of coronary heart disease (Kris-Etherton et al., 2002). In particular, Sobaity is rich in n-3 polyunsaturated fatty acids (PUFA), docosahexaenoic acid (DHA), and eicosapentaenoic acid (EPA). Interestingly, the wild-caught Sobaity contains a higher n-3 PUFA than their cultured counterparts (Hossain et al., 2017; Hossain et al., 2019). Moreover, the highest muscle lipid content recorded for Sobaity was during the pre-spawning and spawning seasons.

An extinction risk assessment of marine fishes, mainly for seabreams, conducted recently based on the dataset of the International Union for Conservation of Nature's Red List indicated that around 25 species are in threatened/near-threatened condition as shown by their body weight (Comeros-Raynal et al., 2016). In this context, the availability of the complete genome sequence may help in understanding the detailed pathways and functions at the molecular level, which may further help in improving the economic value of the fish as well as pave better ways for their conservation.

Next-generation sequencing has propelled the construction of draft genome sequences of various important organisms (Goodwin et al., 2016) including many fish species from the Sparidae family (Shin et al., 2018; Zhu et al., 2021). The complete genome sequence is available for very few species of Sparidae family that include *Sparus aurata* (Pauletto et al., 2018), *Spondyliosoma cantharus* (GCA_900302685), *Pagrus major* (Shin et al., 2018), and the most recent one *Acanthopagrus latus* (Zhu et al., 2021). The genome of *S. aurata* is approximately 830 Mb and had a GC content of 42% (GCA_900880675.2). The genome of *P. major* is ~875 Mb with a GC content of 38%. The draft genome contained a total of 886,260 scaffolds with an N50 of 4.6 Mb (GCA_002897255.1). *S. cantharus* genome is approximately 680 Mb in length containing 47,064 scaffolds (GCA_900302685.1), whereas the size of *A. latus* genome (GCF_904848185.1) is ~685 Mb contained within 66 scaffolds. The study on *A. latus* presented a chromosome-level genome assembly and explored the molecular basis of sex reversal and the characteristics of the osmoregulation in this species (Zhu et al., 2021).

In the current study, our goal was to sequence, assemble, and annotate the draft genome of *S. hasta*. The draft genome assembly will facilitate future investigations of the biology of this species and provide a valuable resource for the conservation and breeding management of *S. hasta*.

Materials and methods

DNA isolation from fin tissues of sobaity fish

The genomic DNA was isolated from 4 months old female Sobaity fish collected from the Mariculture and Fisheries Department, Kuwait Institute for Scientific Research, Salmiya, Kuwait. DNA isolation was performed from the fin tissues (80 mg) using GenElute Plant Genomic DNA Miniprep Kit.

The quantity of the genomic DNA was estimated using a Nanodrop spectrophotometer and Qubit fluorometer 3.0, and quality was checked by the A260/280 value and 0.8% agarose gel electrophoresis.

RNA isolation

The sobaity seabream larvae were reared in aerated tanks with six air stones, illuminated by natural sunlight and fluorescent light (40 W) with 1,500 lux light intensity in the day and 1,000 lux at night time (Al-Abdul-Elah, 1984; Teng et al., 1999). The stock density in *S. hasta* rearing tanks was 40 larvae/L seawater. The 24 h post-hatch larvae were transferred from Mariculture and Fisheries Department and acclimated to laboratory conditions at the Ecotoxicology Laboratory, Kuwait Institute for Scientific Research, Kuwait. Total RNA from 100 mg of the larvae was extracted using the Invitrogen TRIzol reagent (Life Technologies Corporation, United States) following the instructions provided by the manufacturer. Genomic DNA contamination in the extracted RNA samples was removed using the On-Column DNase 1 Digestion Set (DNASE70, Sigma-Aldrich, United States).

Library preparation and sequencing

One microgram of genomic DNA was randomly fragmented by Covaris. The fragmented genomic DNA was selected by Agencourt AMPure XP-Medium kit to an average size of 200–400 bp. Fragments were end-repaired and then 3' adenylated. Adaptors were ligated to the ends of these 3' adenylated fragments and the fragments were amplified using PCR. The PCR products were purified by the Agencourt AMPure XP-Medium kit. The double-stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single-strand circle DNA (ssCir DNA) were considered as the final library. The library was validated on the Agilent Technologies 2100 bioanalyzer. The qualified libraries were sequenced by BGISEQ-500: ssCir DNA molecule formed a DNA nanoball (DNB) containing more than 300 copies through rolling-cycle replication. The DNBs were loaded into the patterned nanoarray by using a high-density DNA nanochip technology. Finally, 150 bp pair-end reads were obtained by combinatorial Probe-Anchor Synthesis (cPAS). The next-generation sequencing was performed at BGI, Hongkong.

De-novo genome assembly

The high-quality paired-end DNA sequencing data was used for *de novo* assembly of the *S. hasta* genome using MaSuRCA-4.0.3 (Zimin et al., 2013). The MaSuRCA assembler combines the benefits of *deBruijn* graph and Overlap-Layout-Consensus assembly approaches. It aids in different types of analysis by integrating various tools for genome size estimation, error correction, assembly scaffolding, polishing, and has been widely used by the scientific community. In addition, MaSuRCA has been suggested to be at least equal to or better than most of the genome assemblers in terms of assembly quality and completeness by the comparative studies performed on eukaryotic genomes (Mikheenko et al., 2018; Sohn and Nam, 2018). The paired-end reads were error corrected using QuorUM (Marçais et al., 2015), and then used for the construction of k-unitigs. Further, the paired-end reads were extended to form super reads with the help of unitigs. After creating the super-reads, contiging and scaffolding was performed using a modified version of CABOG assembler. Finally, gaps in the scaffold assembly were filled. All the steps were performed using the MaSuRCA assembler. The genome size was estimated using the jellyfish mer-counter, integrated within MaSuRCA. Additionally, we have used backmap tool (Schell et al., 2017; Pfenninger et al., 2022) which estimates the genome size based on the reads mapped to the assembly. The primary assembly was filtered to remove scaffolds shorter than 500 bp.

Repeat annotation and masking

A *de novo* repeat library for *S. hasta* filtered assembly was constructed using RepeatModeler (Flynn et al., 2020), which employs three repeat-finding methods; RECON (Bao and Eddy, 2002), RepeatScout (Price et al., 2005), and TRF (Benson, 1999). The repeat library was then subjected to RepeatMasker to find and mask the repeats in the assembled genome using RMBlast as the default search engine.

Gene prediction, annotation, and assembly completeness

The BRAKER2 pipeline (Brůna et al., 2021) was used to perform gene prediction by integrating *ab initio* gene prediction, RNA-seq based prediction, and predictions based on vertebral protein sequences, which combined the advantages of both GeneMark-ET and AUGUSTUS. The RNA-seq data was generated from the post-hatch fish larvae of Sobaity-seabream (BioProject Accession: PRJNA748027). The filtered RNA-seq reads were aligned to the repeat masked assembly using TopHat2 (Kim et al., 2013) with default parameters. The vertebral protein sequences from various species (n = 4,937,339) used for gene prediction were downloaded from the OrthoDB database (Kriventseva et al., 2019). The ProtHint (Brůna et al., 2020) protein mapping pipeline was used for generating required hints from the vertebral protein sequences for BRAKER. The assembled scaffolds along with the aligned

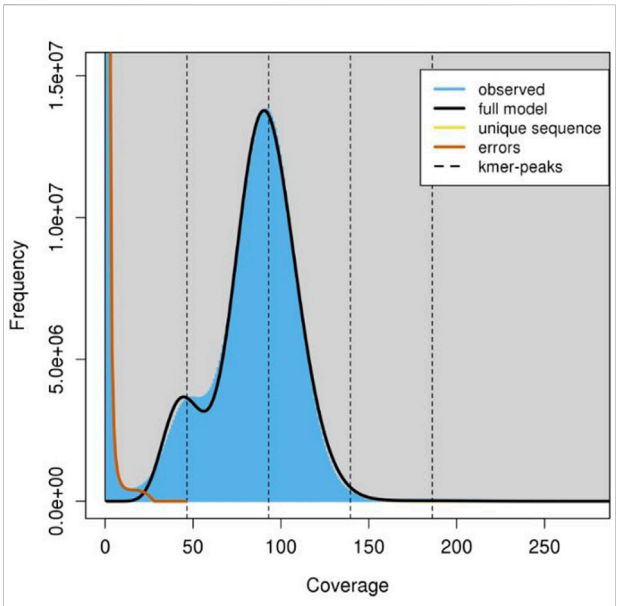


FIGURE 1
K-mer frequency coverage. The graph indicates the frequency of observed kmers using which the genome size is estimated. The jellyfish histo file was used as the input and the graph was constructed using the GenomeScope tool (http://qb.cshl.edu/genomescope/).

TABLE 1 Statistics of the final filtered assembly.

No. of scaffolds	20,442
GC-content	42.1%
L50	2,427
L90	9,117
N50 (bp)	80,670
N90 (bp)	18,310
Min. length	500
Max. length	770,404
Mean length	33,588
Median length	14,545
No. of bases	686609404
No. of 'As'	198736919
No. of 'Cs'	144604718
No. of 'Gs'	144492944
No. of 'Ts'	198522703
No. of 'Ns'	252,120

The L50, L90, N50 and N90 statistics indicate the assembly quality. L50 and L90: Count of smallest number of contigs whose length sum makes up 50% and 90% of the genome size, respectively. N50 and N90: 50% and 90% of the entire assembly is contained in scaffolds that are equal to or larger than these values.

reads (BAM files) and generated hints from the protein sequences were used for generating initial gene structures using the GeneMark-ET tool (Lomsadze et al., 2014). The initial gene structures were then used for training by AUGUSTUS to produce the final gene predictions (Stanke and Waack, 2003). The predicted genes were submitted to Blast2GO tool (Conesa et al., 2005) for annotation.

The raw reads were aligned back to the filtered scaffolds to assess the quality of the genome assembly using Bowtie 2 (Langmead and Salzberg, 2012). Furthermore, the predicted genes from the BRAKER2 pipeline were subjected to BUSCO version 5.2.2 (Manni et al., 2021) to evaluate the completeness of the assembled genome, based on the vertebrata_odb10 database.

Comparative genomics and phylogenetic analysis

We used OrthoMCL v2.0.9 (Li et al., 2003) for ortholog analysis based on protein datasets from the BRAKER2 pipeline and four other fish species: *Diplodus sargus* (txid: 38,941), *Spondyliosoma cantharus* (taxid: 50,595), *Sparus aurata* (txid: 8175), *Acanthopagrus latus* (txid: 8177). For *S. aurata* (GCA_900880675.1) and *A. latus* (GCA_904848185.1), the protein sequences were downloaded from the RefSeq database and used for the phylogenetic analysis. However, for *D. sargus* (GCA_903131615.1) and *S. cantharus* (GCA_900302685.1), the genome sequences were downloaded from the NCBI and proteins were predicted using BRAKER2 pipeline. These protein sequences were then used for the phylogenetic analysis. CD-HIT (Li and GodzikCd-hit, 2006) was used to remove redundant sequences (≥90% identity) in each organism. The protein sequences were further filtered to remove poor quality sequences using 'orthomclFilterFasta' command using default parameters. Then, the non-redundant filtered protein sequences were subjected to all-against-all BLASTp (Altschul et al., 1990) with an E-value of 1e-5. The blast results were used to identify single-copy orthologs using OrthoMCL across the species. The single copy ortholog sequences were then used for multiple sequence alignment by MAFFT, the result of which was used for the construction of phylogenetic tree using FastTree (Price et al., 2009).

Results

Draft genome assembly of *S. hasta*

A total of approximately 550 million paired-end reads were used for constructing the genome assembly of *S. hasta*. The genome size of *S. hasta* was estimated to be around 703 Mb based on *k*-mer statistics using jellyfish *k*-mer counter (Figure 1) integrated within MaSuRCA and 688.8 Mb based on backmap tool. The slight difference in the estimated genome size by both the tools could be attributed to the different approaches used by these tools. The size of the assembled genome was ~687 Mb

TABLE 2 Repeat annotation of the assembly.

Type of repeats		Number of elements	Length (bp)	% of sequence
Retroelements		59,487	13902715	2.02
	SINEs	6,590	859,207	0.13
	Penelope	3,800	518,117	0.08
	LINEs	42,621	10345032	1.51
	L2/CR1/Rex	29,189	6459372	0.94
	R1/LOA/Jockey	2,317	327,028	0.05
	R2/R4/NeSL	152	87,166	0.01
	RTE/Bov-B	5,405	1978982	0.29
	L1/CIN4	1,310	681,130	0.1
	LTR elements	10,276	2698476	0.39
	BEL/Pao	612	282,964	0.04
	Gypsy/DIRS1	2,343	1199000	0.17
	Retroviral	3,156	479,547	0.07
DNA transposons		213,033	34752653	5.06
	hobo-Activator	97,375	15466145	2.25
	Tc1-IS630-Pogo	29,790	5793811	0.84
	PiggyBac	3,453	383,971	0.06
	Tourist/Harbinger	9,907	2052324	0.3
	Other (Mirage, P-element, Transib)	2,563	498,544	0.07
Unclassified		547,666	82948695	12.08
Total interspersed repeats			131604063	19.17
Rolling-circles		12,167	2543620	0.37
Small RNA		4,738	772,793	0.11
Satellites		3,281	493,762	0.07
Simple repeats		383,453	16415048	2.39
Low complexity		41,575	2219709	0.32

SINE: Short interspersed nuclear element; LINE: Long interspersed nuclear element; LTR: Long terminal repeat.

contained within 22,741 scaffolds. The assembly was filtered to remove the scaffolds shorter than 500 Mb, and the final filtered assembly contained 20,442 scaffolds. The size of the filtered assembly was ~686 Mb. The final assembly contained very low N content (~0.04%). Furthermore, the alignment of the cleaned reads indicated successful matching of 99% of the raw reads back to the filtered assembly, suggesting the completeness of the assembly. The filtered assembly was used for further analysis. The complete statistics of the filtered assembly is provided in Table 1.

Repeat identification, annotation, and masking

The repeat sequences in the filtered assembly were predicted by RepeatModeller and masked using RepeatMasker. The total length of repetitive sequences was ~153.4 Mb, accounting for ~22.35% of the draft genome size. Among these, ~12% of the repeats were unclassified. DNA transposons corresponded to ~5%, whereas retroelements corresponded to ~2% of the genome. A complete list of different repeats along with their content in the draft genome has been shown in Table 2.

Gene prediction and annotation of the draft assembly

Gene prediction using BRAKER2 pipeline based on *ab-initio* method, and RNA-seq and ortholog protein sequence-based prediction resulted in a total of 41,201 genes coding for 44,555 transcripts. The mean length of the coding sequence was 1,249 bp, whereas that of protein sequence was 416 aa. Among the protein sequences, a total of 30,545 sequences (68.5% of all the protein sequences) had significant BLAST matches. Many sequences had a mean similarity score of more than 50 (Figure 2A).

Furthermore, 30,473 and 13,244 sequences were mapped to Gene Ontology annotations and different enzyme classes, respectively (Figure 2B). The assessment of the assembly

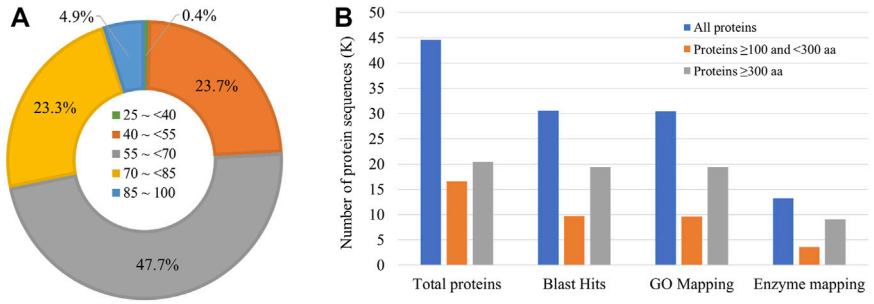


FIGURE 2
Annotation of *Sparidentex* hasta protein sequences. **(A)** Doughnut chart indicating the percentage of sequences with different blast similarity means. **(B)** Annotation of the protein sequences. Different coloured bars indicate the annotation of proteins of varying lengths.

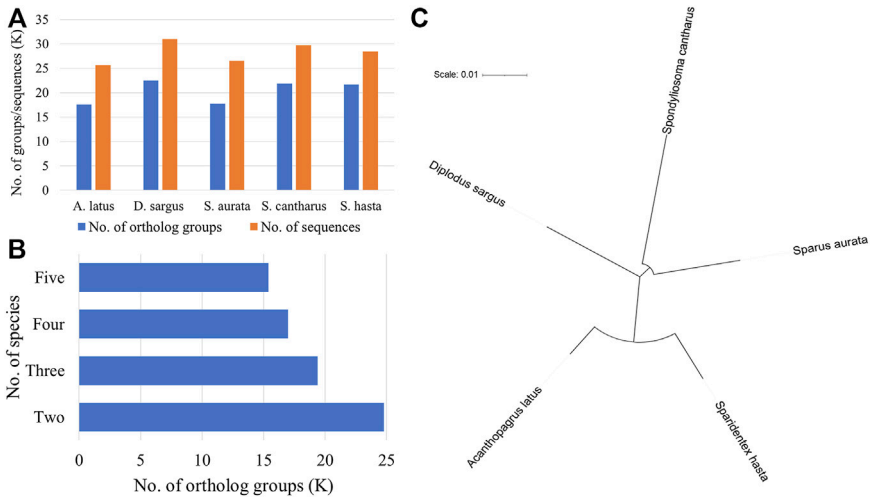


FIGURE 3
Phylogenetic analysis using protein sequences across five fish species. **(A)** No. of ortholog groups and corresponding sequences in each species. The x-axis represents the names of fish species, whereas y-axis represents the number of groups or sequences in thousand. **(B)** No. of ortholog groups shared across the five species. **(C)** Phylogenetic tree of five fish species based on single-copy orthologs.

completeness indicated the detection of 93.4% complete BUSCOs (Benchmarking Universal Single-Copy Orthologs) at the protein level, with the single-copy, duplicated, fragmented, and missing accounting for 82.8, 10.6, 5.1, and 1.5%, respectively.

Comparative genomics

The phylogenetic analysis was performed to understand the relationship among five fish species (*D. sargus*, *S. cantharus*, *S. aurata*, *A. latus*, and *S. hasta*) at the sequence level. The ortholog analysis revealed a total of 24,784 ortholog groups across the five species. *D. sargus* sequences were clustered into most number of ortholog groups (Figure 3A).

Further, there were 15,389 groups that were shared across all five species (Figure 3B). There were a total of 10,785 single-copy orthologs across the five species. The sequences corresponding to these ortholog groups were considered for phylogenetic tree construction. The phylogenetic tree indicated the relationship among the five seabreams and showed that *S. hasta* is closer to *A. latus*, the yellowfin seabream (Figure 3C).

Discussion

In the current study, we assembled the draft genome sequence of Sobaity seabream, *S. hasta*, belonging to Sparidae family. The Sparidae family of fishes are economically important due to their good meat quality and good adaptability to captivity. Currently, very

few species of this family have been completely sequenced at the genome level. The assembled genome size of *S. hasta* was ~680 Mb, closely comparable to the genome of other sequenced seabreams. For instance, the genome size of *P. major* and *A. latus* was estimated to be ~800 Mb. Our assembled genome was shown to be of high quality in terms of completeness, which was indicated based on the overall assembly statistics, such as number of Ns, read alignment and assembly completeness. The N50 statistics of our assembly was comparatively lower than that of the recently published genomes of other closely related species. For instance, the N50 of our assembly was 80 Kb, while this value for *P. major* and *A. latus* contig assembly was 2.8 and 2.6 Mb, respectively (Shin et al., 2018; Zhu et al., 2021).

The lower N50 statistics of *S. hasta* could be attributed to the unavailability of long-read/mate-pair sequences and is a limitation of the current study. Long-read sequences produced from technologies, such as PacBio and Oxford Nanopore can readily traverse the most repetitive regions and help in filling the gaps between contigs, thus increasing the length of assembled sequences, in turn improving N50 statistics (Logsdon et al., 2020). The draft genome of *S. hasta* contained a moderate number of repeats (~22%). This was in agreement with the results from other species of the Sparidae family. The draft genome sequence of *A. latus* contained approximately 19% repeats, among which 14% were unclassified (Zhu et al., 2021). Similarly, the *S. aurata* genome contained around 20% repeats (Pauletto et al., 2018). The genome of *P. major*, however, contained a comparatively higher number of repeat sequences, which corresponded to 31% of its genome (Shin et al., 2018). Furthermore, the genome of *A. latus* contains ~19,600 genes, whereas, the genomes of *S. aurata* and *P. major* have approximately 30,500 and 28,300 protein-coding genes, respectively. In the current study, we estimated *S. hasta* genome to contain a total of 41,201 genes (with 44,555 transcripts), slightly higher than that reported in other seabreams, and approximately, 70% of the protein sequences were significantly aligned to other protein sequences using BLAST. Further, we showed that our assembly quality was good based on the single copy orthologs and alignment of the reads to the assembled genome. The annotation results were in agreement with that of the other published seabream studies. The *S. aurata* genome contained 90% single copy genes and 91% complete BUSCO groups. The BUSCO score for *P. major* was ~98%, whereas, in *A. latus* genome, more than 92% of BUSCO genes were identified. We detected approximately 93% of complete BUSCOs in *S. hasta* genome. The phylogenetic analysis of *S. hasta* and four other seabreams revealed a close relationship between *S. hasta* and *A. latus*.

In summary, we report the first draft assembly of *S. hasta* genome. The size of the filtered assembly was ~686 Mb with 20,442 scaffolds. The repeat sequences were accounted for ~22% of the genome sequence. The assembly contained a total of 44,555 transcript sequences with a mean length of 1,249 bp. Approximately 68% of the protein sequences (n = 30,545) had orthologs based on significant BLAST matches, and 30,473 sequences mapped to Gene Ontology annotations. Furthermore, the comparative genome analysis indicated that *S. hasta* is closer to *A. latus*, a yellowfin seabream. The current assembly provides a solid foundation for future population and conservation studies of *S. hasta* as well as for investigations of environmental adaptation in Sparidae family of fishes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article.

Ethics statement

Necessary permits for sampling and handling fish were obtained from Public Authority for Agriculture and Fish Resources (PAAFR) Kuwait.

Author contributions

QK and VK: Conceptualization, sampling, project administration, manuscript preparation. VK and ABS: DNA and RNA isolation, sample processing, data analysis. QK: funding acquisition. SA-N, ZS, and SE-D: fish culture, rearing, and supply.

Funding

The authors gratefully acknowledge Kuwait Foundation for the Advancement of Sciences (KFAS) and Kuwait Institute for Scientific Research (KISR) for funding the project (Grant No. PR18-12SL-01).

Conflict of interest

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

Publisher's note

References

Abdullah, M., Onn, W., and Lennox, A. (1989). *Culture of marketable sobaity*. Kuwait: Kuwait Institute for Scientific Research. Report No KISR3253.

Al-Abdul-Elah, K., Al-Albani, S., Abu-Rezq, T., El-Dakour, S., Al-Marzouk, A., and James, C. (2010). *Effects of changing photoperiods and water temperature on spawning season of sobaity, Sparidentex hasta*. Kuwait: Kuwait Institute for Scientific Research. Report No KISR10029.

Al-Abdul-Elah, K. (1984). *Procedures and problems of marine fish hatcheries with special reference to Kuwait*. Master of Science Dissertation. University of Stirling.

Al-Zaidan, A. S. Y., Akbar, A., Bahbahani, H., Al-Mohanna, S. Y., Kolattukudy, B., and Balakrishna, V. (2021). Landing, consumption, and DNA barcoding of commercial seabream (Perciformes: Sparidae) in Kuwait. *Aquat. Conserv.* 31 (4), 802–817. doi:10.1002/aqc.3476

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Bao, Z., and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12 (8), 1269–1276. doi:10.1101/gr.88502

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

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-ep+ and AUGUSTUS supported by a protein database. *Nar. Genom. Bioinform.* 3 (1), lqaa108. doi:10.1093/nargab/lqaa108

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *Nar. Genom. Bioinform.* 2 (2), lqaa026. doi:10.1093/nargab/lqaa026

Bukola, D., Zaid, A., Olalekan, E. I., and Falilu, A. (2015). *Consequences of anthropogenic activities on fish and the aquatic environment*. Poultry, Fisheries & Wildlife Sciences.

Carpenter, K., Krupp, F., Jones, D., and Zajonz, U. (1997). "FAO species identification field guide for fishery purposes," in *The living marine resources of Kuwait, Eastern Saudi Arabia, Bahrain, Qatar, and the United Arab Emirates. FAO species identification field guide for fishery purposes the living marine resources of Kuwait, Eastern Saudi Arabia, Bahrain, Qatar, and the United Arab Emirates*.

Comeros-Raynal, M. T., Polidoro, B. A., Broatch, J., Mann, B. Q., Gorman, C., Buxton, C. D., et al. (2016). Key predictors of extinction risk in sea breams and porgies (Family: Sparidae). *Biol. Conserv.* 202, 88–98. doi:10.1016/j.biocon.2016.08.027

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18), 3674–3676. doi:10.1093/bioinformatics/bti610

de la Herran, R., Rejon, C. R., Rejon, M. R., and Garrido-Ramos, M. A. (2001). The molecular phylogeny of the Sparidae (Pisces, Perciformes) based on two satellite DNA families. *Hered. (Edinb)* 87 (6), 691–697. doi:10.1046/j.1365-2540.2001.00967.x

European Food Safety Authority (2008). Animal welfare aspects of husbandry systems for farmed European seabass and gilthead seabream-Scientific Opinion of the Panel. *EFSA J.* 6 (11), 844. doi:10.2903/j.efsa.2008.844

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* 117 (17), 9451–9457. doi:10.1073/pnas.1921046117

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17 (6), 333–351. doi:10.1038/nrg.2016.49

Hekmatpour, F., Kochanian, P., Marammazi, J. G., Zakeri, M., and Mousavi, S. M. (2019). Changes in serum biochemical parameters and digestive enzyme activity of juvenile sobaity sea bream (Sparidentex hasta) in response to partial replacement of dietary fish meal with poultry by-product meal. *Fish. Physiol. Biochem.* 45 (2), 599–611. doi:10.1007/s10695-019-00619-4

Hossain, M., Al-Abdul-Elah, K., and Yaseen, S. (2019). Seasonal variations in proximate and fatty acid composition of sobaity sea bream (Sparidentex hasta) in Kuwait waters. *J. Mar. Biol. Assoc. U. K.* 99 (4), 991–998. doi:10.1017/s0025315418000991

Hossain, M. A., Al-Abdul-Elah, K. M., and El-Dakour, S. (2017). Evaluation of different commercial feeds on grow-out silver black porgy, Sparidentex hasta (Valenciennes), for optimum growth performance, fillet quality, and cost of production. *Saudi J. Biol. Sci.* 24 (1), 71–79. doi:10.1016/j.sjbs.2015.09.018

Karam, Q., Annabi-Trabelsi, N., Al-Nuaimi, S., Ali, M., Al-Abdul-Elah, K., Beg, M. U., et al. (2021). The response of sobaity sea bream *Sparidentex hasta* larvae to the toxicity of dispersed and undispersed oil. *Pol. J. Environ. Stud.* 30 (6), 5065–5077. doi:10.15244/pjoes/133231

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14 (4), R36. doi:10.1186/gb-2013-14-4-r36

Kris-Etherton, P. M., Harris, W. S., and Appel, L. J. (2002). Fish consumption, fish oil, omega-3 fatty acids, and cardiovascular disease. *circulation* 106 (21), 2747–2757. doi:10.1161/01.cir.0000038493.65177.94

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., et al. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47 (D1), D807-D811–d11. doi:10.1093/nar/gky1053

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

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

Li, W., and GodzikCd-hit, A. (2006). Cd-Hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi:10.1093/bioinformatics/btl158

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21 (10), 597–614. doi:10.1038/s41576-020-0236-x

Lomsadze, A., Burns, P. D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 42 (15), e119. doi:10.1093/nar/gku557

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38 (10), 4647–4654. doi:10.1093/molbev/msab199

Marçais, G., Yorke, J. A., and Zimin, A. (2015). QuorUM: An error corrector for illumina reads. *PLoS One* 10 (6), e0130821. doi:10.1371/journal.pone.0130821

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34 (13), i142–i150. doi:10.1093/bioinformatics/bty266

Pauletto, M., Manousaki, T., Ferraresso, S., Babbucci, M., Tsakogiannis, A., Louro, B., et al. (2018). Genomic analysis of *Sparus aurata* reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. *Commun. Biol.* 1, 119. doi:10.1038/s42003-018-0122-7

Pfenninger, M., Schönnenbeck, P., and Schell, T. (2022). ModEst: Accurate estimation of genome size from next generation sequencing data. *Mol. Ecol. Resour.* 22 (4), 1454–1464. doi:10.1111/1755-0998.13570

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

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26 (7), 1641–1650. doi:10.1093/molbev/msp077

Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M., et al. (2017). An annotated draft genome for Radix auricularia (Gastropoda, Mollusca). *Genome Biol. Evol.* 9 (3), 585–592. doi:10.1093/gbe/evx032

Shin, G. H., Shin, Y., Jung, M., Hong, J. M., Lee, S., Subramaniyam, S., et al. (2018). First draft genome for red sea bream of family Sparidae. *Front. Genet.* 9, 643. doi:10.3389/fgene.2018.00643

Sohn, J. I., and Nam, J. W. (2018). The present and future of de novo whole-genome assembly. *Brief. Bioinform.* 19 (1), 23–40. doi:10.1093/bib/bbw096

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 (2), ii215–25. doi:10.1093/bioinformatics/btg1080

Teng, S-K., El-Zahr, C., Al-Abdul-Elah, K., and Almatar, S. (1999). Pilot-scale spawning and fry production of blue-fin porgy, Sparidentex hasta (Valenciennes), in Kuwait. *Aquaculture* 178 (1-2), 27–41. doi:10.1016/s0044-8486(99)00039-3

Teng, S. K., James, C. M., Al-Ahmad, T., Rasheed, V., and Shehadeh, Z. (1987). *Development of technology for commercial culture of Sobaity fish in*

Kuwait. Vol. III. Recommended technology for commercial application. Kuwait Institute for Scientific Research. Report No KISR2269, Kuwait.

Torfi Mozanzadeh, M., Marammazi, J. G., Yaghoubi, M., Agh, N., Pagheh, E., and Gisbert, E. (2017). Macronutrient requirements of silvery-black porgy (Sparidentex hasta): A comparison with other farmed sparid species. *Fishes* 2 (2), 5. doi:10.3390/fishes2020005

Yaghoubi, M., Mozanzadeh, M. T., Safari, O., and Marammazi, J. G. (2018). Gastrointestinal and hepatic enzyme activities in juvenile silvery-black porgy (Sparidentex hasta) fed essential amino acid-deficient diets. *Fish. Physiol. Biochem.* 44 (3), 853–868. doi:10.1007/s10695-018-0475-3

Yousif, O. M., Ali, A., and Kumar, K. (2003). *Spawning and hatching performance of the silvery black porgy Sparidentex hasta under hypersaline conditions*.

Zainal, K., and Altuama, R. (2020). The instantaneous growth rate of maricultured Sparedentix hasta (Valenciennes, 1830) and *Sparus aurata* (Linnaeus, 1758). *Arab Gulf J. Sci. Res.* 38 (3), 208–221. doi:10.51758/agjsr-03-2020-0012

Zhu, K. C., Zhang, N., Liu, B. S., Guo, L., Guo, H. Y., Jiang, S. G., et al. (2021). A chromosome-level genome assembly of the yellowfin seabream (Acanthopagrus latus; Hottuyn, 1782) provides insights into its osmoregulation and sex reversal. *Genomics* 113 (4), 1617–1627. doi:10.1016/j.ygeno.2021.04.017

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29 (21), 2669–2677. doi:10.1093/bioinformatics/btt476

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 November 2022
DOI 10.3389/fgene.2022.1039838

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Il-Youp Kwak,
Chung-Ang University, South Korea
Satish Kumar Illa,
Sri Venkateswara Veterinary University, India

*CORRESPONDENCE
Ruimin Qiao,
qrm480@163.com
Panyang Hu,
hpy9809.@163.com

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

RECEIVED 08 September 2022
ACCEPTED 26 October 2022
PUBLISHED 09 November 2022

CITATION
Qiao R, Zhang M, Zhang B, Li X, Han X, Wang K, Li X, Yang F and Hu P (2022), Population genetic structure analysis and identification of backfat thickness loci of Chinese synthetic Yunan pigs. *Front. Genet.* 13:1039838. doi: 10.3389/fgene.2022.1039838

Population genetic structure analysis and identification of backfat thickness loci of Chinese synthetic Yunan pigs

Ruimin Qiao*, Menghao Zhang, Ben Zhang, Xinjian Li, Xuelei Han, Kejun Wang, Xiuling Li, Feng Yang and Panyang Hu*

College of Animal Science and Technology, Henan Agricultural University, Zhengzhou, China

Yunan is a crossed lean meat pig breed in China. Backfat thickness is the gold standard for carcass quality grading. However, over 14 years after breed registration, the backfat of Yunan thickened and the consistency of backfat thickness decreased. Meanwhile, no genetic study has been ever performed on Yunan population. So, in this study we collected all the 120 nucleus individuals of Yunan and recorded six backfat traits of them, carried out population genetic structure analysis, selection signals analysis and genome-wide association study of Yunan pigs with the help of their founder population Duroc and Chinese native Huainan pigs, to determine the genomic loci on backfat of Yunan. Genetic diversity indexes suggested Yunan pigs had no inbreeding risk while population genetic structure showed they had few molecular pedigrees and were stratified. A total of 71 common selection signals affecting growth and fat deposition were detected by F_{ST} and XP-CLR methods. 34 significant loci associated with six backfat traits were detected, among which a 1.40 Mb region on SSC4 (20.03–21.43 Mb) were outstanding as the strong region underlying backfat. This region was common with the results of selection signature analysis, former reported QTLs for backfat and was common for different kinds of backfat traits at different development stage. *ENPP2*, *EXT1* and *SLC30A8* genes around were fat deposition related genes and were of Huainan pig's origin, among which Type 2 diabetes related gene *SLC30A8* was the most reasonable for being in a 193.21 Kb haplotype block of the 1.40 Mb region. Our results had application value for conservation, mating and breeding improvement of backfat thickness of Yunan pigs and provided evidence for a human function gene might be reproduced in pigs.

KEYWORDS

synthetic pig breed, genetic diversity, population structure, selection signature, GWAS, backfat thickness

Introduction

Huainan is one of the oldest northern China native pig breeds. It is native to south of the upper reaches of Huai River and North of Dabie Mountains in Henan Province. Henan is located in the middle and lower reaches of the Yellow River and is one of the earliest pig domestication areas in China (Zhang et al., 2007). Huainan is an all-black pig breed with large body size, large ears, short mouth, strong and robust limbs, high fertility, strong adaptability and delicious meat quality (Wang et al., 2005). People in and around Huainan pig producing areas have the habit of eating black pigs. Black pig is an essential ingredient of the famous local cuisine there. However, the poor growth rate of Huainan pigs limited its development.

Since 1980s, a large number of foreign commercial pigs with fast growth rate and high lean meat rate have been introduced to China to improve growth performance of Chinese native pig breeds (Xu, 2004). Among them, the American Duroc with golden coat color was widely used because of its high growth rate and high feed conversion rate (Figure 1A). So, in 1996, we started to use Huainan and American Duroc pigs as the base herd to intercross to develop a new high-performance black pig, Yunan pig (Figures 1B,C) (Zhu et al., 2009).

Yunan pig is a black crossbred pig breed obtained by nine generations crossbreeding between Huainan and American Duroc (Figure 2). In 2008, Yunan was registered as a new pig breed in China by National Livestock and Poultry Genetic Resources Committee. Theoretically, it contains 37.50% of Huainan's lineage and 62.50% of Duroc lineage. Yunan exhibits good growth performances from Duroc with an average daily gain of 648 g during the stage of body weight at 30 kg–90 kg and lean meat rate of 56%, and has good meat quality from Huainan with an excellent meat quality (intramuscular fat content as 4.11%). So, Yunan becomes a popular black pig breed in and around Henan.

Since the year of 2008, the systematic breeding of Yunan pigs have been no longer carried out which resulted in a decline of backfat consistency. Backfat thickness now is still the only gold standard for black pig carcass grading as same as the commercial pigs in China. Therefore, the decline of backfat performance of Yunan brings great economic loss to farmers. Meanwhile, with the emergence of the African swine fever in China in 2018, the population of Yunan pigs reduced seriously.

However, there has never been a genomic study on Yunan pigs. Only a few correlation analyses of several genes and comparative analysis of production traits of Yunan were reported (Li et al., 2016; Wang et al., 2021). Herein, in this study, we recorded six backfat thickness traits of the core group of Yunan population, and used SNP array genotyping data of Yunan, Huainan and Duroc with the following objectives: 1) determine the genetic diversity of Yunan by calculating the observed heterozygosity and the expected heterozygosity; 2) access the inbreeding state by detecting runs of homozygosity (ROH); 3) detect selection signals of Yunan by pairwise F_{ST} and XP-CLR methods; 4) investigate genomic evidence on the population structure of Yunan by phylogenetic tree, principal component and admixture analysis; 5) identify the genomic region that controlled backfat thickness of Yunan.

Material and methods

Animals sampling, genotyping and phenotyping

Ear tissues were collected from Yunan (n = 120) and Huainan (n = 33) pigs at Sungo Agricultural and Livestock Co. Henan, China. Genomic DNA was extracted using the phenol-chloroform method and genotyped by 50 K SNP (Compass) and 80 K SNP (NEOGEN). Five backfat thickness traits including shoulder backfat (SBF), sixth and seventh ribs backfat (SSRBF), last rib backfat (LRBF), lumbar joint backfat (LBF) and P2 backfat (P2BF) were measured and recorded for Yunan pigs using ultrasonic instrument (EXAGO, France). Average backfat (ABF) was calculated as the average of SBF, SSRBF, LBF and LRBF. Genotype data of American Duroc pigs (n = 40) (Illumina PorcineSNP60 Genotyping Bead Chip) was downloaded from the public Dryad database (http://dx.doi.org/10.5061/dryad.30tk6).

We used PLINK v.1.9 (Purcell et al., 2007) to perform the quality control (QC) of the total 193 individuals genotype data. SNPs without positions on pig reference genome (*Sus scrofa* 11.1), or with genotyping rate less than 90%, or with minor allele

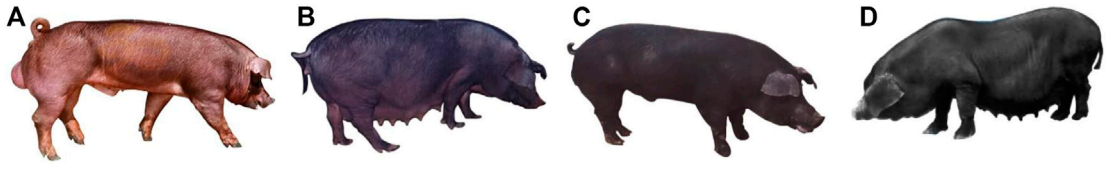


FIGURE 1
Yunan, Huainan, and Duroc used in this study. **(A)** Duroc. **(B,C)** Yunan **(D)** Huainan.

frequency (MAF) lower than 1% or on Y chromosome were excluded. Individuals with genotyping rate less than 90% were excluded. Genotypic data of the same SNP in Yunan, Huainan and Duroc populations were extracted for genetic diversity, genetic differentiation and population genetic structure analyses.

Genetic diversity and selection signal analysis

Genetic indicators including the observed heterozygosity (Ho), the expected heterozygosity (He) and MAF were calculated for Yunan, Huainan and Duroc by PLINK v.1.9 to compare the genetic diversity of these populations, command set is "-hardy". In addition, SNePv1.1 (Barbato et al., 2015) software was used to calculate the effective population size (Ne).

The length and frequency of ROH can reflect the group history. A long ROH indicates recent inbreeding, while a short ROH indicates ancient inbreeding. Genomic homozygous fragments of each individual were detected using runs of homozygosity of PLINK v.1.9. The following parameters were used of define ROHs: 1) the minimum length of ROHs was 500 Kb; 2) a sliding window of 50 SNPs across the genome; 3) each window allowed one heterozygous genotype and five missing SNPs (Xu et al., 2021) to avoid false negatives caused by occasional genotyping errors and missing genotypes. Then the ratio of the total length of the ROH fragment to the total length of autosomal genome was calculated to get the coefficient of inbreeding (Silió et al., 2013) using the following formula:

$$F_{ROH} = \frac{\sum L_{ROH}}{L_{auto}}, \quad (1)$$

where L_{ROH} is the sum of the lengths of the autosomal ROH fragments and L_{auto} is the total size of 18 autosomes of pigs covered by SNPs, which is 2.265 Gb (https://www.ncbi.nlm.nih.gov).

We used SNP data from three breeds, using genetic differentiation coefficient (F_{ST}) and cross-population compound likelihood ratio test (XP-CLR) for genome-wide selective detection. The genetic differentiation coefficient (F_{ST}) between populations was calculated using Vcftools v.0.1.13 (Sun et al., 2020), The command set was as follows: 1) window size was set to 500 Kb; 2) step length was set to 40 Kb. The F_{ST} values ranging from 0 to 0.05, 0.05 to 0.15, and 0.15 to 0.25 indicate that there are no genetic, moderate, and large differentiations among populations, respectively. XP-CLR was calculated using XP-CLR (Chen et al., 2010) software. XP-CLR models the difference in frequency of multiple alleles between the two populations. Further, we used the parameters ("-w1 0.005 200 2000 -p0 0.95") to calculate the XP-CLR score for each chromosome. The empirical cutoffs for the genomic windows with top 1% F_{ST} and XP-CLR values across the whole genome were considered as selective sweeps.

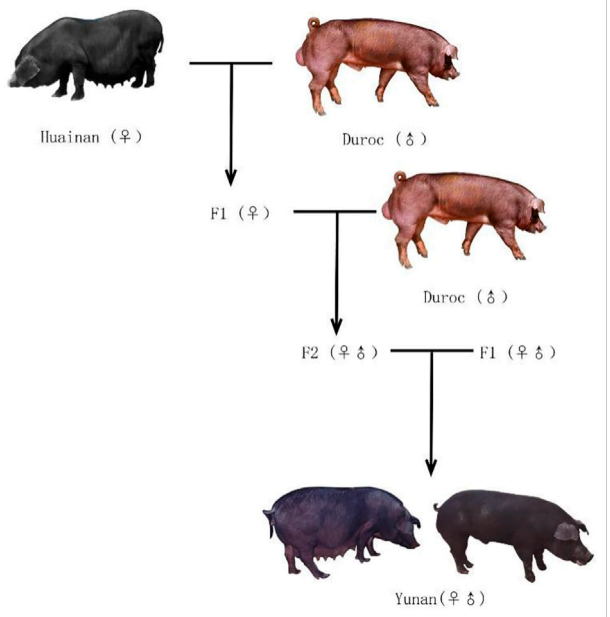


FIGURE 2
Schematic representation of the setup to generate Yunan using Huainan and Duroc.

Population genetic structure analysis

To assess the individual genetic distances between populations to illustrate the relationship between populations, we carried out principal components analysis (PCA) of the SNP dataset using PLINK v.1.9 and selected the first two principal components for visualization by ggplot package in R v.4.1.3 (Team, 2014). To gain more insight into the Yunan pig population, we performed t-distributed stochastic neighbour embedding (T-SNE) analysis on the dataset using the R package "Rtsne" (Krijthe, 2015). The phylogenetic tree was built by neighbor-joining (NJ) method and visualized using R v.4.1.3, the genetic distance matrix was constructed by PLINK v.1.9. To explore the population structure of Yunan, the rapid model of ADMIXTURE v.1.3 (Alexander et al., 2015) software was used to cluster all Yunan samples in a context of some related populations and the results were visialized using the "barplot" package of R v.4.1.3 software.

Genome-wide association analysis of backfat thickness

Genome-wide association analysis for backfat traits of Yunan was performed using a univariate mixed linear model of GEMMA v.0.98.5 (Zhou et al., 2012) as follows.

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \quad (2)$$

TABLE 1 Summary of population genetic diversity indexes.

Breed	No	Ne	MAF	H_O	He
Yunan	120	67	0.2813	0.3756	0.3685
Huainan	33	43	0.2337	0.3422	0.3193
Duroc	40	77	0.2462	0.3019	0.3277

where y is a vector of phenotypic values for the trait; W is a matrix of fixed effects including the first three PCs; α is a vector of corresponding coefficients including the intercept; x is a vector of genotypes for the marker; β is an effect size for the marker; u is an n-vector of random effects; and ε is an n-vector of errors.

The results of genome-wide association analysis were visualized using the rMVP (Yin et al., 2021) package in R software. The suggestive significance threshold was 1/N, where N is the number of SNPs used for the analysis. We used 1 Mb upstream and downstream significant SNPs region as traits related candidate region.

Results

Genotyping

There are 57,466 SNPs sites in the original data. After quality control, SNPs without positions on pig reference genome (Sus scrofa 11.1), or with genotyping rate less than 90%, or with minor allele frequency lower than 1% or on Y chromosome were excluded. Individuals with genotyping rate less than 90% were excluded. Finally, 50,717 SNPs were used for the following analysis.

Genetic diversity analysis

To assess genetic diversity of Yunan, we analyzed MAF, Ho, He and Ne of Yunan, Huainan and Duroc. The results were shown in Table 1. By comparison, the Ne of Huainan was lower than that of Yunan and Duroc. Indexes MAF, Ho and He of Yunan were higher than those of Huainan and Duroc indicating that Yunan was rich in genetic diversity.

ROH analysis results of Yunan, Huainan and Duroc were shown in Table 2. A total of 3,317 homozygous fragments were detected in the three populations. Duroc had the highest average number of ROH per animal (48.8 ± 6.55 with a range of 2–80 Mb) while Huainan had the lowest (7.21 ± 8.14 with a range of 1–105 Mb). Mean length of ROH (MGL_{ROH}) was maximum in Yunan (10.01 ± 4.28 Mb) and minimum in Duroc (8.91 ± 1.14 Mb). Duroc revealed the highest ROH based inbreeding (F_{ROH} = 0.1925 ± 0.037). F_{ROH} of Huainan (F_{ROH} = 0.0324 ± 0.043) and Yunan (F_{ROH} = 0.043 ± 0.033) were lower than Duroc. Yunan did not have severe inbreeding.

ROH fragment size of Yunan mainly concentrated in 5–15 Mb, accounting for 88.98% while that of Duroc and Huainan mainly concentrated in 1–10 Mb, accounting for 75.32% and 71.01%, respectively (Figure 3A). Chromosomes with the most ROH fragments in Yuanan, Duroc and Huainan were SSC1 (n = 170), SSC1 (n = 255) and SSC7 (n = 22), and chromosomes with the least ROH fragments were SSC17 (n = 15), SSC17 (n = 53) and SSC18 (n = 4) (Figure 3B). The total average length of ROH of Yunan, Duroc and Huainan were 98.37 Mb, 436.28 Mb and 73.32 Mb (Figure 3C).

Population genetic structure analysis of yunan population

The first and second principal components of Yunan, Huainan and Duroc explained 29.68% and 9.99% of the total variance. Yunan, Huainan, and Duroc were clearly separated (Figure 4A). The first PC clearly separated Yunan, Huainan and duroc. Yunan was located between Huainan and Duroc. We used t-SNE to best classify the populations to perform dimensionality reduction clustering analysis on all the breeds. From Figure 4B, these results indicate that different subpopulations exist in Yunan pig population. From the result of admixture analysis in Figure 4C, When K = 2, Yunan, Huainan and Duroc could be clearly distinguished. Yunan contained 33.78% of Huainan and 66.22% of Duroc blood. When K = 3, a new bloodline in red was appeared in Yunan. When K = 4-5, different degrees of differentiation appeared in Yunan population. To investigate genetic structure of Yunan pigs, we constructed the NJ tree of the 193 individuals, which indicated the existing Yunan pigs could be divided into four main branches, each of which was divided into two or more subbranches (Figure 4D).

Selection signal screening among yunan, Huainan and Duroc

To screen the selection signals among Yunan, Huainan and Duroc, and analyse the possibly origin of the signals, we divided Yunan, Huainan and Duroc into three groups (Yunan-Duroc Vs. Huainan, Yunan-Huainan Vs. Duroc and Yunan Vs. Huainan-Duroc) and used two methods of F_{ST} and XP-CLR. Top 1% regions of F_{ST} and XP-CLR were considered as salient loci under selected. F_{ST} screening results were shown in Figure 5A (Yunan-Duroc Vs. Huainan), **5C** (Yunan-Huainan Vs. Duroc) and **5E** (Yunan Vs. Huainan-Duroc). We detected several significant loci on SSC1-12, SSC14-17 and SSCX in Yunan-Duroc Vs. Yunan group, SSC1-16 and SSCX in Yunan-Huainan Vs. Duroc group, and SSC1-10, SSC12-18 and SSCX in Yunan-Duroc Vs. Yunan group. XP-CLR analysis results of the above three groups were shown in Figure 5B (Yunan-Duroc Vs. Huainan), **5D** (Yunan-Huainan Vs. Duroc) and **5F** (Yunan Vs. Huainan-Duroc). We

TABLE 2 Genomic distributions and descriptive statistics of ROH in Yunan, Huainan and Duroc.

Breeds	N_{ROH}	Range ROH (Mb)	NM_{ROH}	MGL_{ROH} (Mb)	F_{ROH}
Yunan	1,126	3–124	9.38 ± 4.55	10.01 ± 4.28	0.043 ± 0.033
Huainan	238	1–105	7.21 ± 8.14	9.49 ± 6.21	0.034 ± 0.043
Duroc	1953	2–80	48.82 ± 6.55	8.91 ± 1.14	0.1925 ± 0.037

N_{ROH}: Total number of ROH per breed; Range ROH: length range of ROH; NM_{ROH}: Mean number of ROH in a breed; MGL_{ROH}: Breed wise mean genome length covered by ROH in Mb; F_{ROH}: Inbreeding coefficient based on ROH

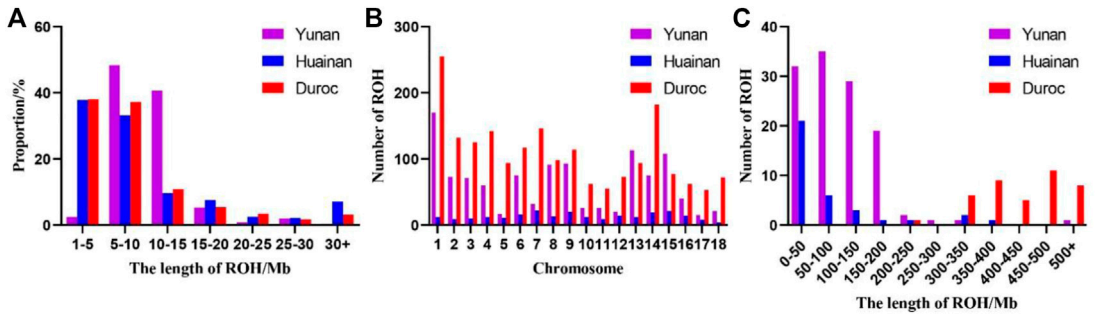


FIGURE 3
(A) Distribution of ROH length in Yunan, Huainan and Duroc. **(B)** Distribution of ROH quantity on Yunan, Huainan and Duroc chromosome. **(C)** Distribution of ROH samples number in Yunan, Huainan and Duroc.

obtained 71 significant loci on each of the 18 autosomes in each of the three groups.

Then we combined the outputs of F_{ST} and XP-CLR analysis. In Yunan-Duroc Vs. Huainan group, we found 37 significant regions that over threshold under selected were overlapped in F_{ST} and XP-CLR results. A total of 104 genes were annotated in these regions in the Ensembl (http://ensemble.org) database, 17 out of which were involved in fat production and metabolism, nine out of which were related to growth and development. These 26 genes under selected were thought to be of Duroc origin.

Similarly, in Yunan-Huainan Vs. Duroc group, we got 21 overlapping regions containing 103 annotated genes, 14 genes out of which were related to fat regulation while eight genes were related to growth development. These 22 genes were thought to be of Huainan origin. Also, 13 overlapping regions covering 116 genes were found in Yunan Vs. Huainan-Duroc. Only eight genes were involved in fat metabolism and adipose production, three genes were involved in skeletal muscle development. These outstanding 11 genes may have been selected by artificial selection in Yunan population.

To further look into former reported economic characters underlying these genomic regions, we extracted the common regions between these results and PigQTLdb (www.animalgenome.org/cgi-bin/QTLdb/index). QTLs that overlapped with potentially selected regions were QTLs of average daily gain (Fontanesi et al., 2014; Guo et al., 2017), backfat last rib (Gilbert et al., 2007) and days 110 kg (Wang et al., 2015), etc. as shown in Table 3.

Genome-wide association study on backfat traits in yunan pigs

To understand genetic background of backfat thickness variation in Yunan, Yunan pigs were divided into gilt (n = 48), 1st parity (n = 48) and 2nd parity (n = 24) according to the age. Descriptive statistics of the six backfat traits (P2BF, SBF, SSRBF, LRBF, LBF and ABF) of this three groups were shown in Table 4. From Table 4, Yunan was much fatter from the ideal body shape. For example, the average P2BF thickness of gilt, 1st parity and 2nd parity in Yunan was 36.90, 35.28 and 35.93 mm. This was much thicker than the commercial sows (usually 16–22 mm). Moreover, the coefficient variation (CV) of the backfat traits was larger, ranged from 12.80% to 34.96%.

With the goal of pinpointing genomic region associated with backfat phenotypes of Yunan, we performed genome-wide association studies of six backfat traits using a linear mixed model. We identified 34 suggestive significant SNPs associated with six backfat traits in total. Thereinto, nine significant SNPs on SSC1-2, SSC4, SSC11-12, SSC14 and SSC16 were detected in 1st parity pigs. 14 significant SNPs on a 1.4 Mb segment of SSC4

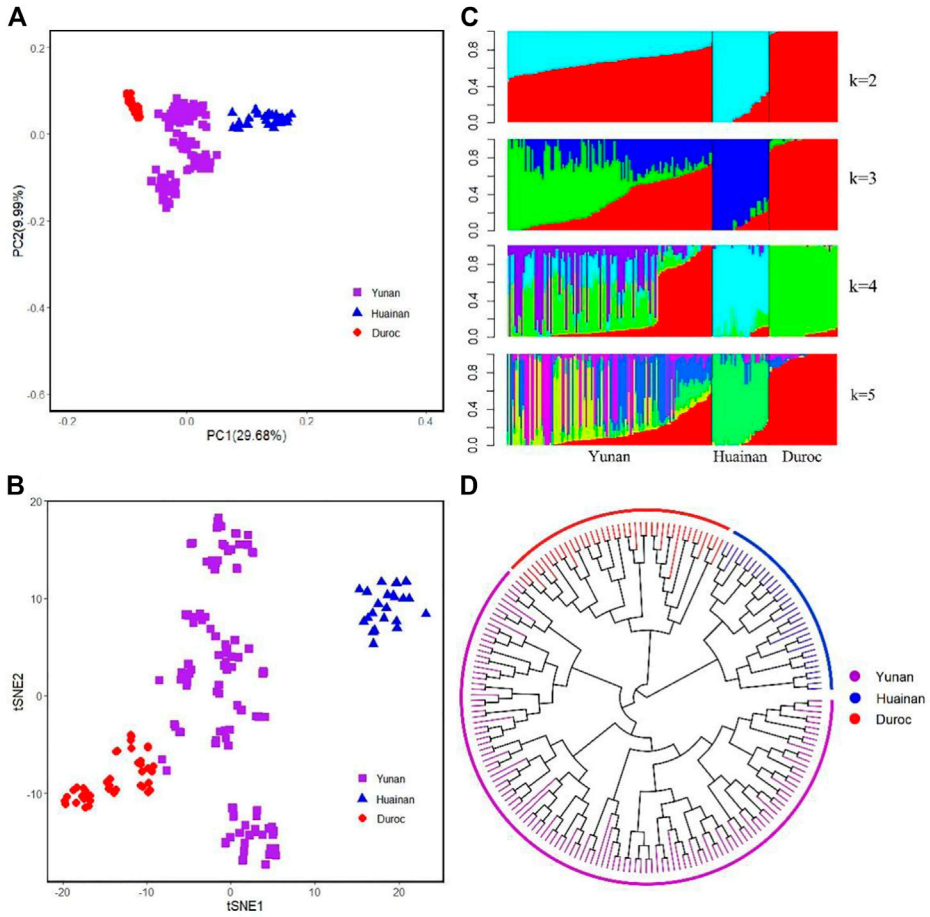


FIGURE 4
Population structure analyses for all pig individuals. **(A)** First and second principal components from a principal component analysis of all populations. **(B)** t-SNE plot for all breeds. **(C)** Population structure plots for all pig populations at K = 2, 3, 4, and 5. **(D)** Neighbor-joining tree for all individual pigs.

(20.03–21.43 Mb) were detected in 2nd parity pigs while 11 significant SNPs on SSC1-2, SSC8, SSC13 and SSCX were obtained in gilts. 202 genes were annotated in the 34 genomic regions in the Ensembl database, five, six and three genes of which were involved in fat metabolism or growth in gilt, 1st and 2nd parity pigs (Figure 6).

The genomic regions where these significant loci were located partially overlapped. The most significant associated SNP was the same for P2BF and LRBF on SSC8 (10.43 Mb) in gilts, for P2BF and ABF on SSC16 (21.97 Mb) in 1st parity, for P2BF, LRBF and ABF on SSC4 (20.03–21.43 Mb) and for SSRBF and ABF on SSC4 (106.29 Mb) in 2nd parity pigs. Then we combined these 34 regions with PigQTLdb database (www.animalgenome.org/cgi-bin/QTLdb/index) and found that significant loci on SSC4 (20.03–21.43 Mb) were overlapped with a reported average backfat thickness QTL (20,366,121–21,945,045 bp) in large white pigs (Bink et al., 2000) (Table 5).

In view of the existence of selective signals related to fat deposition and growth, and to investigate whether the genomic regions associated with backfat traits mapped by GWAS might have been selected in the population, we analyzed the regions common to the results of F_{ST}, XP-CLR and GWAS. The selected region on SSC4 (21.24–21.83 Mb) in Yunan-Huainan Vs. Duroc group which was thought to be of Huainan origin and the P2BF, LRBF and ABF associated region on SSC4 (20.03–21.43 Mb) in 2nd parity stood out. These two regions were overlapped, and *ENPP2* and *EXT1* genes (SSC4:19.02–21.06 Mb) in this region were functionally related to fatty acid production (Crespo-Piazuelo et al., 2020). Therefore, the genome region of 20.03–21.83 Mb on SSC4 was considered as an important candidate region for backfat thickness in Yunan pig population.

Finally, we used Haploview v4.2 (Barrett et al., 2005) software for linkage disequilibrium analysis and constructed haplotype modules for the SSC4 (20.03–21.83 Mb) under the default

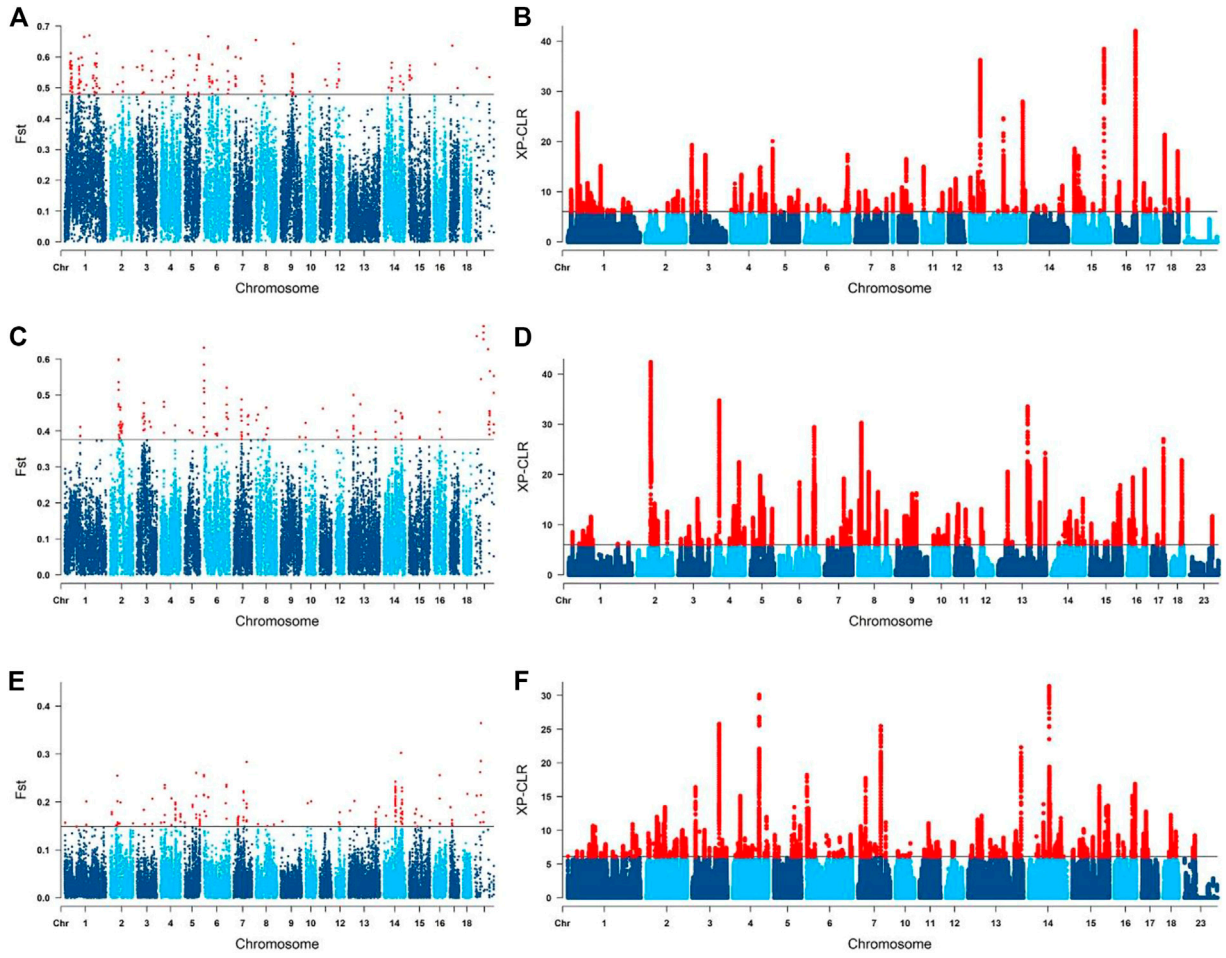


FIGURE 5
Selected signature detection for different groups on manhattan graphs. **(A,B)** yunan-duroc vs. huainan. **(C,D)** yunan-huainan vs. duroc. **(E,F)** yunan vs. huainan-duroc.

parameters. We found four linkage blocks (r^2≥0.8), block1 (21, 111, 825–21,196,785 bp), block2 (21, 208, 018–21,218,173 bp), block3 (21, 402, 451–21,421,498 bp) and block4 (21, 516, 432–21,709,640 bp) within the target area (Figure 7). Only one gene *SLC30A8* (21, 517, 486–21,555,757 bp) was found in block4. *SLC30A8* gene encodes zinc transporter, which transfers zinc from the cytoplasm of pancreatic β-cells to intracellular vesicles and functions in insulin secretion. Impaired insulin secretion and insulin resistance are the pathogenesis of type 2 diabetes mellitus (Witka et al., 2019).

Discussion

The genetic diversity of a population was a key factor in ensuring the survival and evolution of a species. Commercial pigs have been subjected to high-intensity artificial selection for a time, resulting in a decrease in genetic diversity, while local varieties were mainly selected by natural conditions such as environmental factors and geographical factors, thus maintaining a high level of genetic diversity. China's native pig breeds has a better genetically diverse (SanCristobal et al., 2006; Ai et al., 2013). The genetic diversity of Chinese hybrid pigs was higher than that of Chinese local pig breeds (Yang et al., 2003; Huang et al., 2020) and commercial pig breeds (Wang et al., 2021). There were many statistical methods to evaluate population genetic diversity, such as allele frequency and population heterozygosity. The higher the heterozygosity, the higher the genetic diversity. In this study, the MAF, Ho and He of the hybrid pig breed Yunan were higher than those of its founders Duroc and Huainan, indicating Yunan had a richer genetic diversity. Effective population size (Ne) is also an important indicator of diversity and species conservation. If the Ne was below 65, the breed might have a population crisis

TABLE 3 Candidate genes and previous reported QTLs of common selected regions from F_{ST} and XP-CLR analysis.

Group	SSC	Position (bp)	Gene		QTL
			Backfat	Growth	
Yunan-Duroc Vs. Huainan	1	44,480,000–4,4,931,794	GPRC6A		
	1	90,160,000–90,495,794		SENP6	
	1	97,080,000–97,995,794	SMAD2, ZBTBZ7C		Body weight (birth)
	3	2,770,524–3,132,524	SDK1	SLC26A2	Body weight (end of test)
	3	36,784,268–37,498,268		C16orf89	
	3	71,302,268–71,340,268	ZNF638		
	6	54,733,496–55,767,496	CPT1	CLEC11A	Average daily gain
	7	14,960,000–15,610,808	ID4		Backfat at last rib
	8	978,400–1036400	NSD2	NELFA	Leaf fat weight
	9	76,400,000–7734,000	TAC1	DLX5	
	9	109,103,272–109,460,000	SFRP5	CRTAC1	
	15	3,800,000–4,580,000		MBD5, ACVR2A	Average backfat thickness
	17	17,009,180–17,211,180	PLCB1		
Yunan-Huainan Vs. Duroc	2	56,423,768–56,919,768	NLRP3		
	2	5,8,071,768–60,123,768	LPAR2, GATAD2A, SUGP1, NCAN, INSL3, JAK3, CRTC1	CLIP2, MEF2B, SLC25A42, CRLF1, GDF15	Body depth
					Body width
	2	68,667,768–68,780,000	PIN1, OLFM2	UBL5	
	2	69,440,001–69,625,768	DNM2, CARM1	TGFBR3L	Backfat at first rib
	4	21,240,001–21,830,268	SLC30A8		Average backfat thickness
	6	149,794,156–150,244,156	ANGPTL3		Body weight
					Days to 100 kg
	13	33,334,864–33,480,864		DOCK3	
Yunan Vs. Huainan-Duroc	2	44,416,175–44,614,028	PDE3B		
	6	4,808,255–5,832,188	CDH13		Body length
	6	29,366,624–29,413,703	AMFR		
	7	28,300,808–29,064,808	BEND6	RAB23	Average daily gain
	12	22,829,286–23,387,286	MED1, PIP4K2B		
	14	75,501,272–77,161,272	MCU, PLA2G12B, ANXA7	MSS51, MYOZ1	
	14	112,560,001–113,175,272	LDB1		

(Simon, 1999). The Ne of Yunan was 67. This suggested that although having experienced the threat of African swine fever, Yunan did not have a group crisis now.

ROH segments contains information about population inbreeding, and its length and frequency can reflect the population history. Compared with pedigree inbreeding number, the calculation of genomic inbreeding number based on ROH was more accurate and can better reflect the real inbreeding number of an individual (Purfield et al., 2012; Zanella et al., 2016; Deniskova et al., 2019; Bhati et al., 2020). A longer ROH usually indicates a closer genetic relationship while a shorter ROH indicates an ancient inbreeding. A larger number of ROHs and fragments means a higher probability of inbreeding (Kirin et al., 2010). Here, the average total length of ROH of Yunan was between Huainan and Duroc. Duroc had the highest average number of ROH per animal while Huainan had the lowest. Mean length of ROH was maximum in Yunan and minimum in Duroc. Duroc had the highest ROH based on inbreeding. F_{ROH} of Huainan and Yunan were lower than Duroc. Compared with some other local breed pigs (Mashen and Chun'an) in China (Cai et al., 2021; Dai et al., 2021) and European commercial breeds (Zhan et al., 2020), the average inbreeding coefficient of Yunan was relatively lower, and Yunan black pigs did not show obvious inbreeding.

From few large branches in NJ trees, population stratification in PCA, and having no one main lineage when K = 3 to 5 of

TABLE 4 Descriptive statistical of backfat traits in Yunan pigs.

	Number	Trait	Mean (mm)	SD	Min	Max	CV %
Gilt	48	P2BF	36.90	9.21	20.70	53.90	34.96
		SBF	15.70	3.24	9.40	26.90	20.64
		SSRBF	40.00	9.29	23.20	58.90	23.23
		LRBF	34.64	8.37	20.10	53.90	24.16
		LBF	20.58	3.24	15.00	27.60	15.74
		ABF	27.70	5.04	19.58	37.93	18.19
Paity 1	48	P2BF	35.28	7.37	20.70	50.80	20.89
		SBF	18.40	2.73	13.80	26.30	14.84
		SSRBF	35.77	8.74	20.70	54.50	24.43
		LRBF	31.96	7.50	18.20	52.60	23.47
		LBF	21.78	3.99	13.80	36.40	18.32
		ABF	26.98	4.79	18.00	39.03	17.75
Paity 2	24	P2BF	35.93	6.74	22.60	52.00	18.76
		SBF	18.36	2.35	13.80	24.40	12.80
		SSRBF	39.67	9.94	21.30	60.30	25.06
		LRBF	32.87	6.54	21.30	50.10	19.90
		LBF	21.76	2.88	17.50	30.10	13.24
		ABF	28.17	5.03	19.43	40.43	17.86

P2BF, P2 point backfat; SBF, Shoulder backfat; SSRBF, Six and seven rib backfat; LRBF, Last rib backfat; LBF, Lumbosacral backfat; ABF, Average backfat; SD, Standard Deviation; CV, Coefficient of variance.

ADMIXTURE implied much more breeding work should be done in Yunan. In addition, different subsets of Yunan may exist according to PCA, which was also indicated by t-SNE analysis. So, even if Yunan pigs had no risk of inbreeding and recession according to the population diversity and inbreeding analysis, but this diversity may partly be due to stratification within groups. Therefore, in the future breeding work of Yunan pigs, we should carry out mating among lineages according to the NJ molecular pedigree to avoid multiple invalid matings, otherwise the consistency of the population would not be improved.

Selection signals can reflect loci and genes that have been strongly selected in a population during long-term domestication. The selective signals we detected in Yunan black pigs were not only related to fat deposition, but also to growth. Both directions had both Duroc origin (Yunan-Duroc VS. Huainan) and Huainan origin (Yunan-Huainan VS. Duroc). For example, *GPRC6A* (Mukai et al., 2021) and *MBD5* (Du et al., 2012) for fatness and growth of Duroc while *ANGPTL3* (Jiang et al., 2018) and *DOCK3* (Reid et al., 2020) of Huainan. These results indicated that both Duroc and Huainan pigs had genetic variants affecting growth and fat deposition. In addition, some genes, such as *CDH13* (Göddeke et al., 2018) and *RAB23* (Hasan et al., 2020), were found to be involved in fat deposition and growth in the Yunan VS. Huainan-Duroc group. These loci might be derived from the founder effect of the initial breeding group of Yunan or might be the result of the 14 years of breeding process of Yunan.

Genomic loci associated with backfat thickness partially overlapped between different parities and different traits. This suggested the influence of some genes on backfat was a long-term process, there's no time and space specificity. When it came to the nonoverlapped genes or regions, we could not come to a conclusion of time or space specificity. Because although the population used in this study covers the core population of Yunan black pigs, the number of populations was too small. This reminded us that under the epidemic environment, it was necessary to adopt multi-site conservation for species, especially for local genetic resources.

Three genes related to fat deposition, *ENPP2*, *EXT1* and *SLC30A8*, were found in the upper and lower 1 Mb of SSC4 (20.03–21.43 Mb) (Du et al., 2009; Hutley et al., 2011; Nishimura et al., 2014). Among the four linkage disequilibrium blocks in the 1.40 Mb interval of SSC4, only one gene, *SLC30A8*, was present in the fourth block (193.21 Kb). This gene is the star gene of type 2 diabetes. In 2009, there was a study that attempted to find the causal mutation for human type 2 diabetes by analyzing the correlation between the mutation of this gene with fat deposition in pigs (Du et al., 2009) but failed. Further analysis of this gene in Yunan might be needed.

In addition, QTL has been widely used in the study of important economic traits in pigs. The hunt for QTL in pigs has been ongoing for nearly 2 decades, beginning with the first publication of a QTL for fatness on pig chromosome 4 in 1994 (Andersson et al., 1994). In crossbreeding analysis, SNP markers

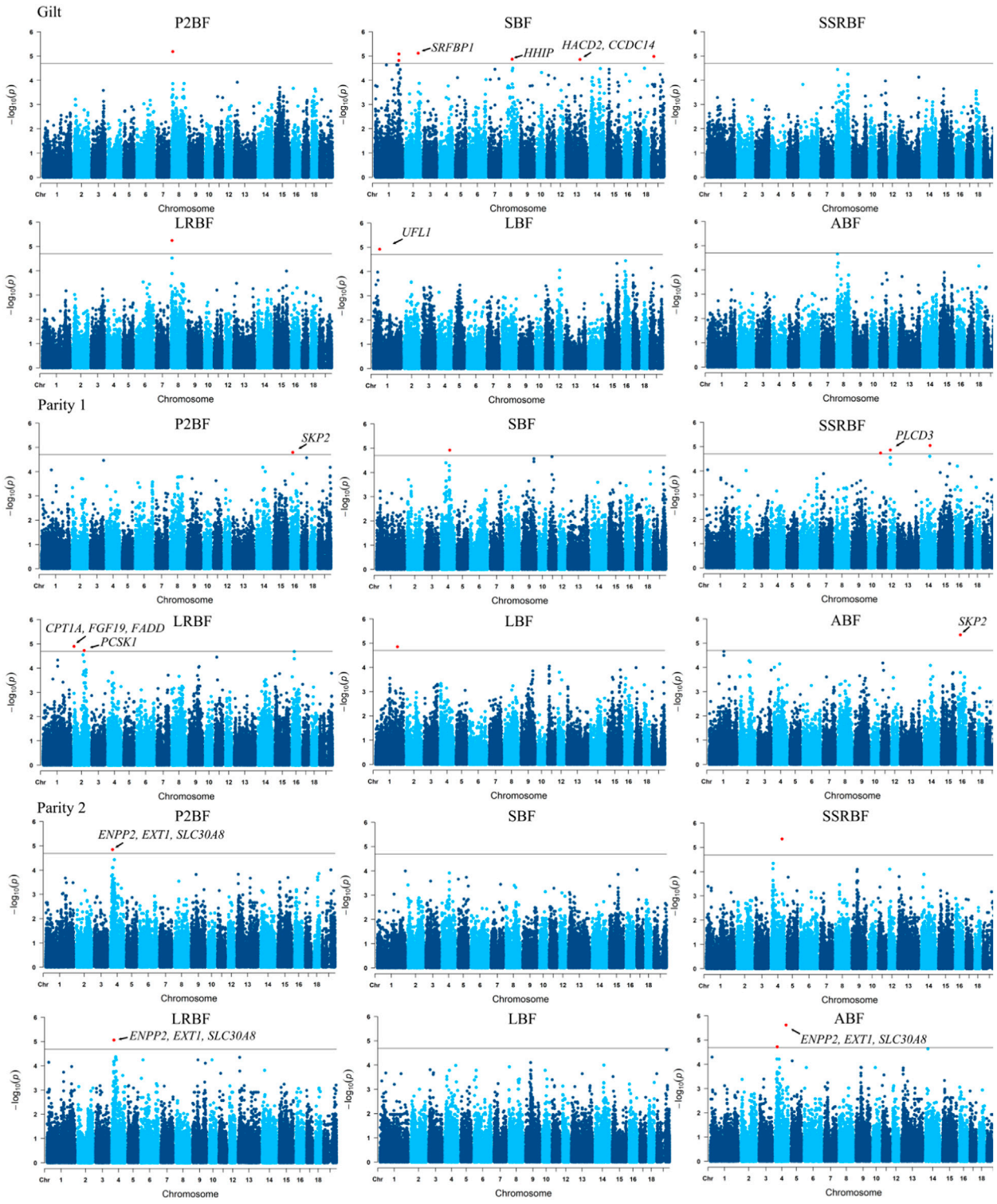


FIGURE 6
Manhattan plot of genome-wide association analysis of P2-backfat, shoulder backfat, sixth and seventh ribs backfat, last rib backfat, lumbar joint backfat and average backfat at different parity. The annotated genes represent the relevant candidate genes.

TABLE 5 Summary of genomic regions significantly associated with six backfat traits in Yunan pigs.

Parity	Phenotype	SSC	Position (Mb)	Number of significant SNPs	Most significant SNP		Minor allele frequency	Candidate gene name
					Position (bp)	*p*-value		
Gilt	P2BF	8	10.43	1	10,430,709	6.58E-06	0.2396	
	SBF	1	233.79–235.18	4	234,643,280	8.30E-06	0.125	
		2	125.35	1	125,353,968	7.77E-06	0.375	*SRFBP1*
		8	83.12	1	83,125,662	1.37E-05	0.08333	*HHIP*
		13	136.08	1	136,088,769	1.40E-05	0.2812	*HACD2, CCDC14*
		23	23.89	1	23,893,572	1.05E-05	0.05208	
	LRBF	8	10.43	1	10,430,709	5.69E-06	0.2396	
	LBF	1	63.41	1	63,418,505	1.21E-05	0.09375	*UFL1*
Parity 1	P2BF	16	21.97	1	21,976,238	1.61E-05	0.4271	*SKP2*
	SBF	4	83.26	1	83,269,208	1.21E-05	0.4375	
	SSRBF	11	26.09	1	26,093,623	1.84E-05	0.1354	
		12	17.86	1	17,867,041	1.38E-05	0.1562	*PLCD3*
		14	85.98	1	85,986,081	9.06E-06	0.0625	
	LRBF	2	3.52	1	3,528,858	1.27E-05	0.1458	*CPT1A, FGF19, FADD*
		2	102.23	1	102,231,770	1.89E-05	0.1562	*PCSK1*
	LBF	1	213.80	1	213,808,326	1.42E-05	0.01042	
	ABF	16	21.97	1	21,976,238	4.61E-06	0.4271	*SKP2*
Parity 2	P2BF	4	20.03–21.43	4	20,038,718	1.44E-05	0.4792	*ENPP2, EXT1, SLC30A8*
	SSRBF	4	106.29	1	106,293,182	4.47E-06	0.1250	
	LRBF	4	20.03–21.43	4	20,038,718	8.73E-06	0.4792	*ENPP2, EXT1, SLC30A8*
	ABF	4	106.29	1	106,293,182	2.44E-06	0.1250	
		4	20.03–21.43	4	20,038,718	1.91E-05	0.4792	*ENPP2, EXT1, SLC30A8*

Significance of SNP.

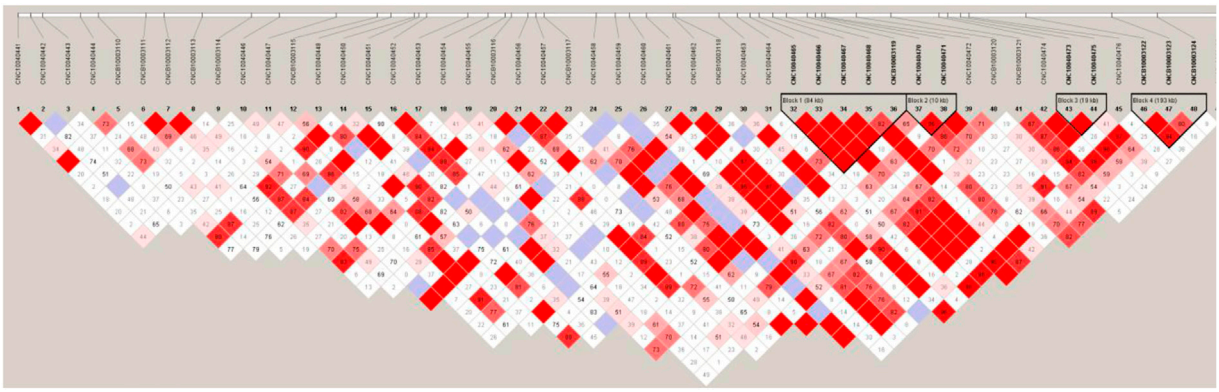


FIGURE 7
Haploview of linkage disequilibrium of SNPs on chromosome four (20.03–21.83 Mb) in Yunan pigs.

and QTL linkage analysis can be used to identify genomic regions with complementary effects from potential resource groups, so as to increase the degree of genetic complementarity between varieties or lines in a planned way, and thus improve the economic traits of hybrids and the genetic diversity of varieties.

In conclusion, we analyzed genetic structure and mapped genomic regions affecting backfat thickness of Yunan black pigs.

Although there was no risk of inbreeding depression, the stratification phenomenon existed in Yunan population. The increasing backfat thickness and decreasing consistency had a particular genetic basis. The 1.40 Mb interval of SSC4 (20.03–21.43 Mb) was a strong candidate region associated with backfat thickness. *ENPP2*, *EXT1,* and *SLC30A8*, particular *SLC30A8*, were strong candidate genes. Our findings are helpful for the subsequent breeding and conservation, as well as genomic improvement of backfat thickness of Yunan black pigs and suggested the importance of multi-site breeding conservation method.

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://bigd.big.ac.cn, GVM000386.

Author contributions

RQ contributed to the conception of the study and directed the wrote of the manuscript. PH performed date analysis and wrote the manuscript. MZ and BZ contributed to sample collection and data collection. XnL, XH, KW, XuL and FY contributed to revise the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (U1904115).

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

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PLoS One* 8 (2), e56001. doi:10.1371/journal.pone.0056001

Alexander, D. H. Shringarpure, S. S. Novembre, J., and Lange, K. L. (2015). *Admixture 1.3 software manual.*

Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., JohanssonM., Andersson, K., et al. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Sci. (New York, N.Y.)* 263 (5154), 1771–1774. doi:10.1126/science.8134840

Barbato, M., Orozco-terWengel, P., Tapio, M., and Bruford, M. W. (2015). SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* 6, 109. doi:10.3389/fgene.2015.00109

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263–265. doi:10.1093/bioinformatics/bth457

Bhati, M., Kadri, N. K., Crysnanto, D., and Pausch, H. (2020). Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics* 21 (1), 27. doi:10.1186/s12864-020-6446-y

Bink, M. C., Te Pas, M. F., Harders, F. L., and Janss, L. L. (2000). A transmission/disequilibrium test approach to screen for quantitative trait loci in two selected lines of large white pigs. *Genet. Res.* 75 (1), 115–121. doi:10.1017/s0016672399004061

Cai, C., Zhang, X., Zhang, W., Yang, Y., Gao, P., Guo, X., et al. (2021). Evaluation of genetic structure in mashen pigs conserved population based on SNP chip. *Chin. J. Animal Veterinary Sci.* 52 (04), 920–931.

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20 (3), 393–402. doi:10.1101/gr.100545.109

Crespo-Piazuelo, D., Criado-Mesas, L., Revilla, M., Castelló, A., Noguera, J. L., Fernández, A. I., et al. (2020). Identification of strong candidate genes for backfat and intramuscular fatty acid composition in three crosses based on the Iberian pig. *Sci. Rep.* 10 (1), 13962. doi:10.1038/s41598-020-70894-2

Dai, L., Zhu, X., Chen, X., Yang, N., Huang, S., and Xu, R. (2021). Analysis of genetic diversity and genetic structure in chun'an spotted pigs conserved population based on SNP chip. *Swine Prod.* (06), 59–64. doi:10.13257/j.cnki.21-1104/s.2021.06.018

Deniskova, T., Dotsev, A., Lushihina, E., Shakhin, A., Kunz, E., Medugorac, I., et al. (2019). Population structure and genetic diversity of sheep breeds in the Kyrgyzstan. *Front. Genet.* 10, 1311. doi:10.3389/fgene.2019.01311

Du, Y., Liu, B., Guo, F., Xu, G., Ding, Y., Liu, Y., et al. (2012). The essential role of Mbd5 in the regulation of somatic growth and glucose homeostasis in mice. *PLoS One* 7 (10), e47358. doi:10.1371/journal.pone.0047358

Du, Z. Q., Fan, B., Zhao, X., Amoako, R., and Rothschild, M. F. (2009). Association analyses between type 2 diabetes genes and obesity traits in pigs. *Obes. (Silver Spring, Md.)* 17 (2), 323–329. doi:10.1038/oby.2008.557

Fontanesi, L., Schiavo, G., Galimberti, G., Calò, D. G., and Russo, V. (2014). A genomewide association study for average daily gain in Italian Large White pigs. *J. Anim. Sci.* 92 (4), 1385–1394. doi:10.2527/jas.2013-7059

Gilbert, H., Le Roy, P., Milan, D., and Bidanel, J. P. (2007). Linked and pleiotropic QTLs influencing carcass composition traits detected on porcine chromosome 7. *Genet. Res.* 89 (2), 65–72. doi:10.1017/S0016672307008701

Göddeke, S., Knebel, B., Fahlbusch, P., Hörbelt, T., Poschmann, G., van de Velde, F., et al. (2018). CDH13 abundance interferes with adipocyte differentiation and is a novel biomarker for adipose tissue health. *Int. J. Obes.* 42 (5), 1039–1050. doi:10.1038/s41366-018-0022-4

Guo, Y., Huang, Y., Hou, L., Ma, J., Chen, C., Ai, H., et al. (2017). Genome-wide detection of genetic markers associated with growth and fatness in four pig populations using four approaches. *Genet. Sel. Evol.* 49 (1), 21. doi:10.1186/s12711-017-0295-4

Hasan, M. R., Takatalo, M., Ma, H., Rice, R., Mustonen, T., and Rice, D. P. (2020). RAB23 coordinates early osteogenesis by repressing FGF10-pERK1/2 and GLI1. *eLife* 9, e55829. doi:10.7554/eLife.55829

Huang, M., Yang, B., Chen, H., Zhang, H., Wu, Z., Ai, H., et al. (2020). The fine-scale genetic structure and selection signals of Chinese indigenous pigs. *Evol. Appl.* 13 (2), 458–475. doi:10.1111/eva.12887

Hutley, L. J., Newell, F. S., Kim, Y. H., Luo, X., Widberg, C. H., Shurety, W., et al. (2011). A putative role for endogenous FGF-2 in FGF-1 mediated differentiation of human preadipocytes. *Mol. Cell. Endocrinol.* 339 (1-2), 165–171. doi:10.1016/j.mce.2011.04.012

Jiang, Y., Tang, S., Wang, C., Wang, Y., Qin, Y., Wang, Y., et al. (2018). A genome-wide association study of growth and fatness traits in two pig populations with different genetic backgrounds. *J. Anim. Sci.* 96 (3), 806–816. doi:10.1093/jas/skx038

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., and Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5 (11), e13996. doi:10.1371/journal.pone.0013996

Krijthe, J. H. (2015). *Rtsne:. T-distributed Stoch. neighbor Embed. using a Barnes-hut Implement*.

Li, X., Qiao, R., Li, X., Guo, J., Han, X., Zhang, H., et al. (2016). The genetic characteristics of meat quality and nutritional components of yunan black pig and its hybrid pigs. *J. Domest. Animal Ecol.* 37 (03), 20–26.

Mukai, S., Mizokami, A., Otani, T., Sano, T., Matsuda, M., Chishaki, S., et al. (2021). Adipocyte-specific GPRC6A ablation promotes diet-induced obesity by inhibiting lipolysis. *J. Biol. Chem.* 296, 100274. doi:10.1016/j.jbc.2021.100274

Nishimura, S., Nagasaki, M., Okudaira, S., Aoki, J., Ohmori, T., Ohkawa, R., et al. (2014). ENPP2 contributes to adipose tissue expansion and insulin resistance in diet-induced obesity. *Diabetes* 63 (12), 4154–4164. doi:10.2337/db13-1694

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Purfield, D. C., Berry, D. P., McParland, S., and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genet.* 13, 70. doi:10.1186/1471-2156-13-70

Reid, A. L., Wang, Y., Samani, A., Hightower, R. M., Lopez, M. A., Gilbert, S. R., et al. (2020). DOCK3 is a dosage-sensitive regulator of skeletal muscle and Duchenne muscular dystrophy-associated pathologies. *Hum. Mol. Genet.* 29 (17), 2855–2871. doi:10.1093/hmg/ddaa173

SanCristobal, M., Chevalet, C., Haley, C. S., Joosten, R., Rattink, A. P., Harlizius, B., et al. (2006). Genetic diversity within and between European pig breeds using microsatellite markers. *Anim. Genet.* 37 (3), 189–198. doi:10.1111/j.1365-2052.2005.01385.x

Silió, L., Rodríguez, M. C., Fernández, A., Barragán, C., Benítez, R., Óvilo, C., et al. (2013). Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. *J. Anim. Breed. Genet.* 130 (5), 349–360. doi:10.1111/jbg.12031

Simon, D. L. (1999). European approaches to conservation of farm animal genetic resources. *Anim. Genet. Resour. Inf.* 25, 77–97. doi:10.1017/S1014233900005794

Sun, T., Shen, J., Achilli, A., Chen, N., Chen, Q., Dang, R., et al. (2020). Genomic analyses reveal distinct genetic architectures and selective pressures in buffaloes. *Gigascience* 9 (2), giz166. doi:10.1093/gigascience/giz166

Team, R. C. (2014). R: A language and environment for statistical computing. *MSOR Connect.* 1.

Wang, K., Liu, D., Hernandez-Sanchez, J., Chen, J., Liu, C., Wu, Z., et al. (2015). Genome wide association analysis reveals new production trait genes in a male Duroc population. *PloS one* 10 (9), e0139207. doi:10.1371/journal.pone.0139207

Wang, Q., Wang, M., Pang, Y., Wang, Z., and Liu, H. (2005). Studies on the growth and development and fattening performance in huainan pig. *J. Henan Agric. Sci.* (05), 10–74.

Wang, X., Zhang, C., Yue, L., Zhang, H., Cao, T., Hu, Y., et al. (2021a). Sequence analysis of MC1R gene in Yunan black pig. *Heilongjiang Animal Sci. Veterinary Med.* (02), 43–48 159160. doi:10.13881/j.cnki.hljxmsy.2019.09.0228

Wang, Y., Zhao, X., Wang, C., Wang, W., Zhang, Q., Wu, Y., et al. (2021). High-density single nucleotide polymorphism chip-based conservation genetic analysis of indigenous pig breeds from Shandong Province, China. *Anim. Biosci.* 34 (7), 1123–1133. doi:10.5713/ajas.20.0339

Witka, B. Z., Oktaviani, D. J., Marcellino, M., Barliana, M. I., and Abdulah, R. (2019). Type 2 diabetes-associated genetic polymorphisms as potential disease predictors. *Diabetes Metab. Syndr. Obes.* 12, 2689–2706.

Xu, W. (2004). *Ancient and modern agriculture*Introduction and domestication of European breeds of pig in modern China.

Xu, Z., Mei, S., Zhou, J., Zhang, Y., Qiao, M., Sun, H., et al. (2021). Genome-wide assessment of runs of homozygosity and estimates of genomic inbreeding in a Chinese composite pig breed. *Front. Genet.* 12, 720081. doi:10.3389/fgene.2021.720081

Yang, S. L., Wang, Z. G., Liu, B., Zhang, G. X., Zhao, S. H., Yu, M., et al. (2003). Genetic variation and relationships of eighteen Chinese indigenous pig breeds. *Genet. Sel. Evol.* 35 (6), 657–671. doi:10.1186/1297-9686-35-7-657

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 (4), 619–628. doi:10.1016/j.gpb.2020.10.007

Zanella, R., Peixoto, J. O., Cardoso, F. F., Cardoso, L. L., Biegelmeyer, P., Cantão, M. E., et al. (2016). Genetic diversity analysis of two commercial breeds of pigs using genomic and pedigree data. *Genet. Sel. Evol.* 48, 24. doi:10.1186/s12711-016-0203-3

Zhan, H., Zhang, S., Zhang, K., Peng, X., Xie, S., Li, X., et al. (2020). Genome-wide patterns of homozygosity and relevant characterizations on the population structure in piétrain pigs. *Genes (Basel)* 11 (5), E577. doi:10.3390/genes11050577

Zhang, B., Hu, J., Zhang, L., and Winters, M. (2007). Genetics: Breeding to improve dairy cow fertility. *Livestock* 12, 50–52. doi:10.1111/j.2044-3870.2007.tb00109.x

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–824. doi:10.1038/ng.2310

Zhu, X., Li, X., and Zhao, C. (2009). Breeding history and breed superiority of yunan black pig. *China Anim. Ind.* (04), 36–38.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 November 2022
DOI 10.3389/fgene.2022.969959

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Andras Gaspardy,
University of Veterinary Medicine Budapest, Hungary
杰 于,
Northwest A&F University, China

*CORRESPONDENCE
Changfa Wang,
wangcf1967@163.com

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

RECEIVED 15 June 2022
ACCEPTED 31 October 2022
PUBLISHED 22 November 2022

CITATION
Liu Z, Wang T, Shi X, Wang X, Ren W, Huang B and Wang C (2022), Identification of *LTBP2* gene polymorphisms and their association with thoracolumbar vertebrae number, body size, and carcass traits in Dezhou donkeys.
Front. Genet. 13:969959.
doi: 10.3389/fgene.2022.969959

Identification of *LTBP2* gene polymorphisms and their association with thoracolumbar vertebrae number, body size, and carcass traits in Dezhou donkeys

Ziwen Liu, Tianqi Wang, Xiaoyuan Shi, Xinrui Wang, Wei Ren, Bingjian Huang and Changfa Wang*

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

The number of thoracolumbar vertebrae in Dezhou donkeys varies from 22 to 24 and is associated with body size and carcass traits. In mammals, the latent transforming growth factor beta binding protein 2 (*LTBP2*) has been found to have some functions in the development of thoracolumbar vertebrae. The relationship between *LTBP2* and TLN (the number of thoracolumbar vertebrae) of Dezhou donkeys is yet to be reported. The purposes of this study are as follows: 1) to quantify the effect of thoracolumbar vertebrae number variation of Dezhou donkeys on body size and carcass trait; 2) to study the distribution of single nucleotide variants (SNVs) in the *LTBP2* gene of Dezhou donkeys; and 3) to explore whether these SNVs can be used as candidate sites to study the mechanism of Dezhou donkey muti-thoracolumbar vertebrae development. The TLN, body size, and carcass traits of 392 individuals from a Dezhou donkey breed were recorded. All animals were sequenced for *LTBP2* using GBTS liquid chip and 16 SNVs were used for further analysis. We then analyzed the relationship between these SNVs with TLN, body size, and carcass traits. The results showed that: 1) c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, and c.2782 + 3975 A > G were complete genetic linkages and significantly associated with thoracic vertebrae number (TN) ($p < 0.05$) (wild-type homozygotes had more TN than heterozygotes); 2) c.1381 + 768 T > G and c.1381 + 763 G > T were significantly associated with lumber vertebrae number (LN) ($p < 0.05$); 3) c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, and c.812 + 22526 T > G were significantly associated with chest circumference (CHC), front carcass weight (CWF), after carcass weight (CWA), and carcass weight (CW) ($p < 0.05$) (wild-type homozygotes were larger than other genotypes in CHC, CWF, CWA, and CW); and 4) the effect of variation is not consistent in c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T. These results provide useful information that the polymorphism of *LTBP2* is significantly associated with TLN, body size, and carcass traits in Dezhou donkeys, which can serve as a molecule marker to improve donkey production performance.

KEYWORDS

thoracolumbar vertebrae, Dezhou donkey, *LTBP2*, single nucleotide variants (SNVs), association analysis

1 Introduction

In mammals, the vertebral column consists of cervical, thoracic, lumbar, sacral, and caudal vertebrae. The number of vertebrae, especially the thoracolumbar vertebrae number, is an important economic trait in domestic animals because it is related to the length of the carcass and the size of the body (Borchers et al., 2004). The number of thoracic vertebrae (TN) is equal to the number of ribs, which are among the most valuable parts of the meat in the China market, and the number of lumbar vertebrae (LN) is bound up with abdominal muscle. The variation in the number of thoracolumbar vertebrae (TLN) has been observed in many mammalian species, such as in sheep; there are 13 thoracic vertebrae and 6 lumbar vertebrae (T13L6) for the majority of sheep, while the carcass length of T13L7 and T14L6 increases by 2.22 cm and 2.93 cm, respectively, compared with normal T13L6, and carcass weight (CW) increases about 1.68 kg and 1.90 kg (Donaldson et al., 2013; Li et al., 2017). In pig breeds, modern Western pig breeds, such as Duroc, Large White, and Landrace, have more (n = 21–23) TLN than Chinese indigenous breeds (n = 19), generally (King and Roberts, 2010; Mikawa et al., 2011; Liu et al., 2020), and one extra vertebrae expands the carcass length by 80 mm in bacon pigs (King and Roberts, 2010).

The vertebral column originates from the pre-somatic mesoderm under regulation of the notochord, which means the number of vertebrae in a region may vary connaturally, and this has been accurately determined in the embryonic stage (Greene and Copp, 2009). The variability of the vertebral column may arise from cranial-caudal border shifts, which take place when there is a somatic shift from the typical distribution of vertebral segments in a region; this may be one of the reasons for the inconsistent TLN (Thawait et al., 2012). The multi-vertebrae number trait is complex and regulated by genetics; up to now, several genes have been proved to have the function of regulating the development of vertebrae. Two quantitative trait loci (QTL) have been identified to regulate the vertebrate development in pigs: one on chromosome 1 and another one on chromosome 7; through more in-depth research, the nuclear receptor subfamily 6, group A, member 1 (*NR6A1*) gene (c.748 Pro 192 Leu) and vertebrae development-associated (*VRTN*) gene (g.20311-20312 ins291) have been proved to be candidate genes affecting spinal development (Mikawa et al., 2005; Mikawa et al., 2007; Mikawa et al., 2011; Ren et al., 2012; Fan et al., 2013; Burgos et al., 2015; Zhang et al., 2016). Association analysis has revealed significant associations between the single nucleotide variants (SNVs) (rs414302710: A > G) in the exon-8 of the *NR6A1* gene with the number of lumbar vertebrae (Zhang et al., 2019). In addition to this, many genes have been found to be associated with the development of the vertebrae, such as the homeobox gene family (Rijli et al., 1995; Wellik, 2007; Gomez et al., 2008), *GDF11* (Li et al., 2010), and *Btg2* (Park et al., 2004), but most of these studies have focused on model animals, such as mice and fish.

The multi-vertebrae trait has also attracted donkey breeding researchers since it was discovered several decades ago owing to of multi-vertebrae advantages. Furthermore, the multi-vertebrae trait is highly heritable [heritability is 0.62 in pigs (Borchers et al., 2004)]. For the majority of Dezhou donkeys, there are 18 thoracic vertebrae and 5 lumbar vertebrae, usually labeled as T18L5. To date, we have observed five thoracic-lumbar vertebrae combination types in Dezhou donkeys (T17L5, T17L6, T18L5, T18L6, and T19L5); we have not observed the T19L6 type, nor have we seen any article reporting this. We collected skeletal specimens of three thoracic vertebrae types and two lumbar vertebrae types from the Dezhou donkey slaughterhouse (Figure 1). Many articles have reported that the *LTBP2* gene may have the function of regulating the vertebrae development, but the mechanism is not clear. Zhang et al. (2016) found that growth differentiation factor 11 (*Gdf11*) can increase the number of ribs in mice, while some scholars have found that *LTBP2* can inhibit the extracellular processing of *Gdf11* through basic amino acid specific preprotein convertase 5/6 (*PC5/6*). In our previous studies, through the genome-wide association study (GWAS), we predicted that *LTBP2* may be associated with the vertebrae number in Dezhou donkeys (Wang et al., 2020). Therefore, the current study aims to detect mutations in the *LTBP2* gene and analyze their association with the TLN, body size, and carcass traits of Dezhou donkeys. This study contributes to the understanding of the mechanism based of TLN differences and provides theoretical support for multi-vertebrae molecular marker-assisted selection in donkey breeding.

2 Materials and methods

2.1 Experimental animals and data acquisition

The study collected 392 Dezhou donkeys in Dezhou Town, Shandong Province, China, from 2019 to 2021 (the average body size data are shown in Supplementary Table S1). All donkeys were 2-year-old jackasses, fattened for meat production, with unknown pedigree information. During this study, the animal feed consisted of grass and hay *ad libitum* and water. The was no food intake within 12 h before slaughter, but water was unlimited. A 10 ml blood sample was collected from each Dezhou donkey using an EDTA anticoagulated blood

10.3389/fgene.2022.969959

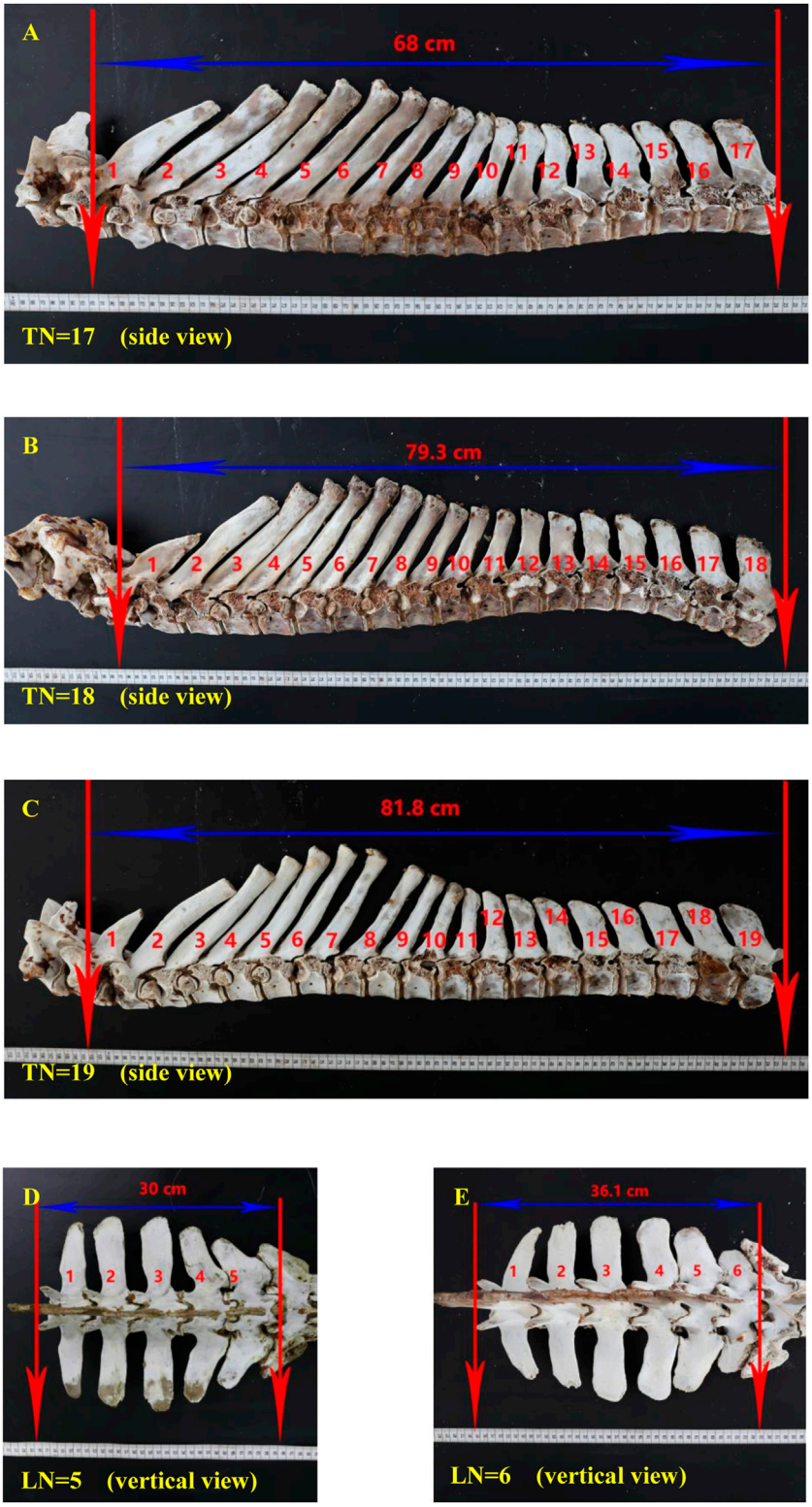


FIGURE 1
Thoracic and lumbar vertebrae specimens of Dezhou donkeys. Notes: **(A)** the side view of a 17 thoracic vertebrae specimen; **(B)** the side view of an 18 thoracic vertebrae specimen; **(C)** the side view of a 19 thoracic vertebrae specimen; **(D)** the vertical view of 5 lumbar vertebrae specimen; **(E)** the vertical view of 6 lumbar vertebrae specimen. In each figure, the red arrows mark the starting and ending points of the thoracic or lumbar vertebrae, respectively. The blue arrows indicate the straight-line length of the thoracic or lumbar vertebrae. Red numbers with units indicate the straight-line length value of the thoracic or lumbar vertebrae. Red numbers without units indicate the order of the thoracic or lumbar vertebrae. "LN" means the number of lumbar vertebrae; "TN" means the number of thoracic vertebrae.

collection tube and stored at −80°C. The body size [body height (BH), the vertical distance from the highest point of bun nail to the ground; body length (BL), the linear distance from the fore-end of humeral carina to the last internal carina of ischial tubercle; and chest circumference (CHC), make a vertical line from the back of the scapula and measure the circumference of its chest] of these 392 Dezhou donkeys was measured using Zhang et al.'s (2021) method. The measurement for collecting the thoracolumbar vertebrae number and carcass traits were as follows. The warm carcass weight [removing the head, viscus, skin, hooves, penis, testicles, and tail (CW)], the front part of the carcass weight [cut off from the boundary between the last thoracic vertebra and the first lumbar vertebra, and the carcass where the thoracic vertebra are located (CWF)], and the hind part of the carcass weight [cut off from the boundary between the thoracic vertebra and the lumbar vertebra, and the carcass where the lumbar vertebra are located (CWA)] were measured immediately after slaughter. The information regarding thoracolumbar vertebrae was measured [thoracic vertebrae number (TN), the total length of thoracic vertebrae (TL), lumbar vertebrae number (LN), and the total length of lumbar vertebrae (LL)] at the abattoir in the cold-storage room after slaughter on the left half of the carcass (Liu et al., 2022). The TIANAMP Genomic DNA Extraction Kit (DP304, TIANGEN, Beijing, China) was utilized to extract the genomic DNA from blood samples. The A260/280 ratios of all DNA samples were determined with a NanoDrop (ND 2000, NanoDrop, United States). After that, genomic DNA were diluted to a common centration 50 ng/μL and stored at −20°C.

2.2 Genotyping

All animals were sequenced for the whole *LTBP2* gene with GBTS (genotyping by targeted sequencing) liquid chip (using GenoBaits and GenoPlexs technology), which was developed by our laboratory and the Shijiazhuang Breeding Biotechnology Co., Ltd. In order to confirm the accuracy of the chip sequencing results, eight pairs of specific primer sequences were designed using Primer Premier 5.0 software to amplify some polymorphic loci of the *LTBP2* gene (50%). The primers are listed in Supplementary Table S2. The fragments contained 16 SNVs (c.5547 + 860 C > T, c.5251 + 281 A > C, c.1381 + 768 T > G, c.1003 + 704 C > T, c.812 + 22526 T > G, c.812 + 6591 T > C, c.565 + 3453 C > T, and c.494 + 5808 C > T) of *LTBP2*. The PCR system (25 μL) included: 2 μL genomic DNA (25 ng/μL); 1 μL of each primer (10 μM); 12.5 μL 2 × MasterMix (TIANGEN, Beijing, China); and 8.5 μL double-distilled H_2O. The PCR protocol was as follows: 95°C for 10 min; 35 cycles of denaturing at 95°C for 30 s; annealing at Tm°C (Supplementary Table S2) for 30 s; and extension at 72°C for 1 min, with a final extension at 72°C for 10 min. After amplification, PCR products were first electrophoresed on 2% agarose gels with MF079-M5 Hipure Gelred nucleic acid strain (Mei5 Biotechnology, Co., Ltd., Beijing, China) and then sent to Sangon Biotech Biotechnology Co., Ltd. for sequencing in both directions and the sequencing results were compared using DNAMAN software (Version 5.2, Lynnon Biosoft, Vaudreuil, Canada).

2.3 Statistical analysis

The genotype frequencies of all target loci (the frequency of each genotype is greater than 5%) were calculated. Using these SNVs, haplotypes were constructed, and haplotype frequency was calculated; haplotypes with frequencies greater than 5% were used for subsequent analysis. The genotype frequencies, allelic frequencies, and Hardy–Weinberg equilibrium partial values were determined by direct counting. At the same time, the population genetic parameters were estimated using online software (http://analysis.bio-x.cn/myAnalysis.php) and (http://www.msrcall.com/Gdicall.aspx), including observed heterozygosity (Obs-Het), predicted heterozygosity (Pred-Het), homozygosity (Ho), effective number of alleles (Ne) (Nei and Roychoudhury, 1974), and polymorphism information content (PIC) (Botstein et al., 1980). The linkage relationship between D' (LD' coefficient) and r^2 (correlation coefficient) based on the alleles of each site was calculated using Haploview software (Version 4.2, Daly Lab at the Broad Institute Cambridge, United States) and haplotype frequency was estimated between the loci (Barrett et al., 2005). In order to explore the differences in SNVs' genotype and haplotypes in *LTBP2*, which are important genetic factors for several traits, the linear model in SPSS 22.0 software (Statistical Product and Service Solutions, Version 22.0 Edition, IBM, Armonk, NY, United States) was used for correlation analysis:

$$y_{ij} = \mu + a_i + e_{ij},$$

where y_{ij} is the measured value of the related trait, μ is the population mean, a_i is the genotype effect, and e_{ij} is the random residual.

3 Results

3.1 Variation in vertebral number in Dezhou donkeys

The variation in vertebral number among 392 Dezhou donkeys is shown in Figure 2. The TN, LN, and TLN ranged from 17 to 19, 5 to 6, and 22 to 24, respectively; the dominated type was 18 (83.2%), 5 (78.3%), and 23 (87.5%), respectively. The most common proportion among the experimental

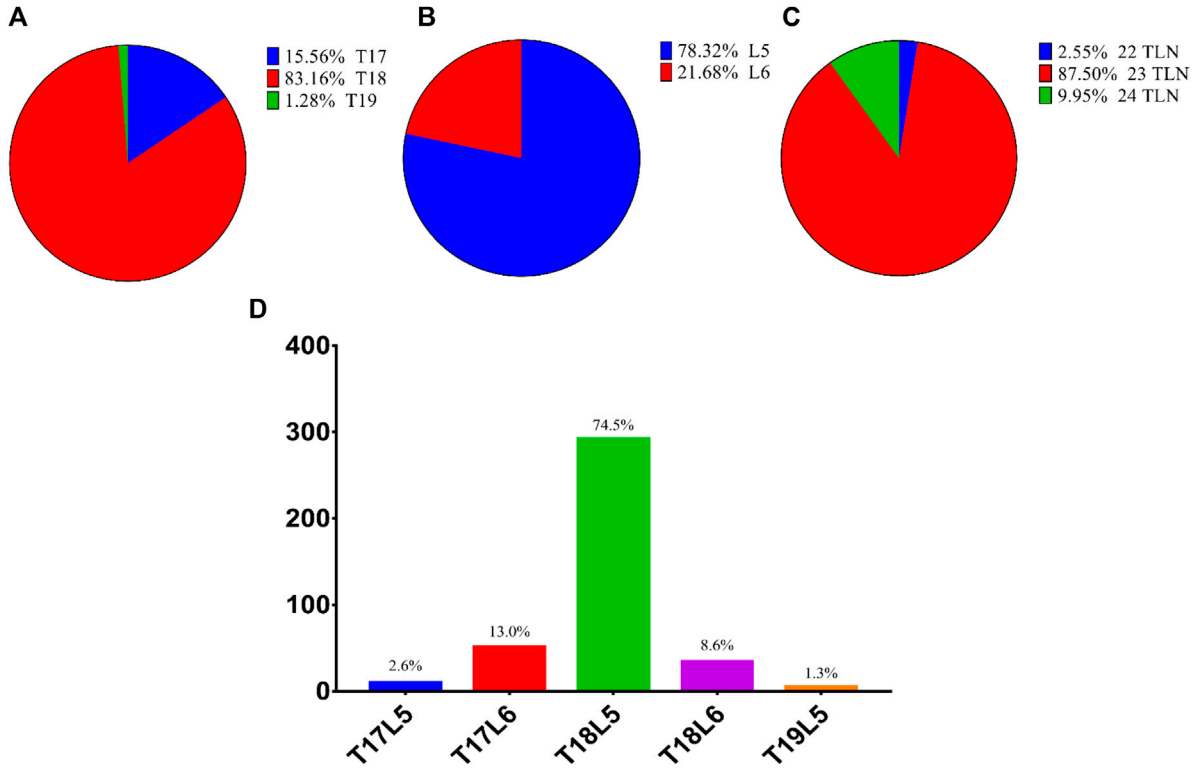


FIGURE 2
Variation in vertebral number of Dezhou donkeys. Notes: **(A)** variation in TN; **(B)** variation in LN; **(C)** variation in TLN; **(D)** variation in TL-type.

samples was the T18L5 thoracic-lumbar vertebrae combinational type (74.5%), and the rarest was T19L5 (1.3%).

3.2 Effect of TLN variation on body size and carcass traits

The number of vertebrae is associated with body length, carcass weight, and other economical traits in livestock. However, for Dezhou donkeys, the effect of the vertebrae number increase on economic traits is still unclear. To evaluate their relationship, the effects of single TL, LN, TLN, and thoracic-lumbar vertebrae combinational types on carcass weight were analyzed successfully. In order to accurately calculate the effect of adding one type of vertebrae on body length and carcass weight, we studied the effect of variations in the number of thoracic vertebrae in the same population of lumbar vertebrae and *vice versa* (Supplementary Table S3). When LN remains unchanged, adding a thoracic vertebra means a significant increase in TL by 5 cm according to the weighted average method (TN = 17, TL = 68.73 ± 0.35; TN = 18, TL = 73.61 ± 0.18; TN = 19, TL = 75.2 ± 1.59), significant at $p < 0.05$. When TN remains unchanged, adding a lumbar vertebra means a significant increase in LL by 4 cm according to the weighted average method (LN = 5, LL = 23.27 ± 0.07; LN = 6, LL = 27.19 ± 0.15, $p < 0.05$). From the perspective of TLN, adding a thoracolumbar vertebra means the total length of thoracolumbar vertebrae (TLL) will increase significantly (TLN = 22, TLL = 91.6 ± 1.07; TLN = 23, TLL = 96.75 ± 0.22; TLN = 24, TLL = 100.59 ± 0.71; $p < 0.05$). Individuals with 23 or 24 TLN were about 10 kg heavier than 22 TLN individuals in carcass weight, but no significant difference was detected. Theoretically, there should be a T19L6 combinational type, but this has not been observed in our long-term observations; maybe we need a larger sample size.

3.3 Polymorphisms of 16 multi-vertebrae trait causal loci

A total of 16 mutations (c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, c.2782 + 3975 A > G, c.1381 + 768 T > G, c.1381 + 763 G > T, c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, c.812 + 22526 T > G, c.812 + 10937 C > T, c.812 + 6591 T > C, c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T) were identified as candidate sites for multi-vertebrae and carcass traits (Figure 3). In order to further confirm their association with vertebrae

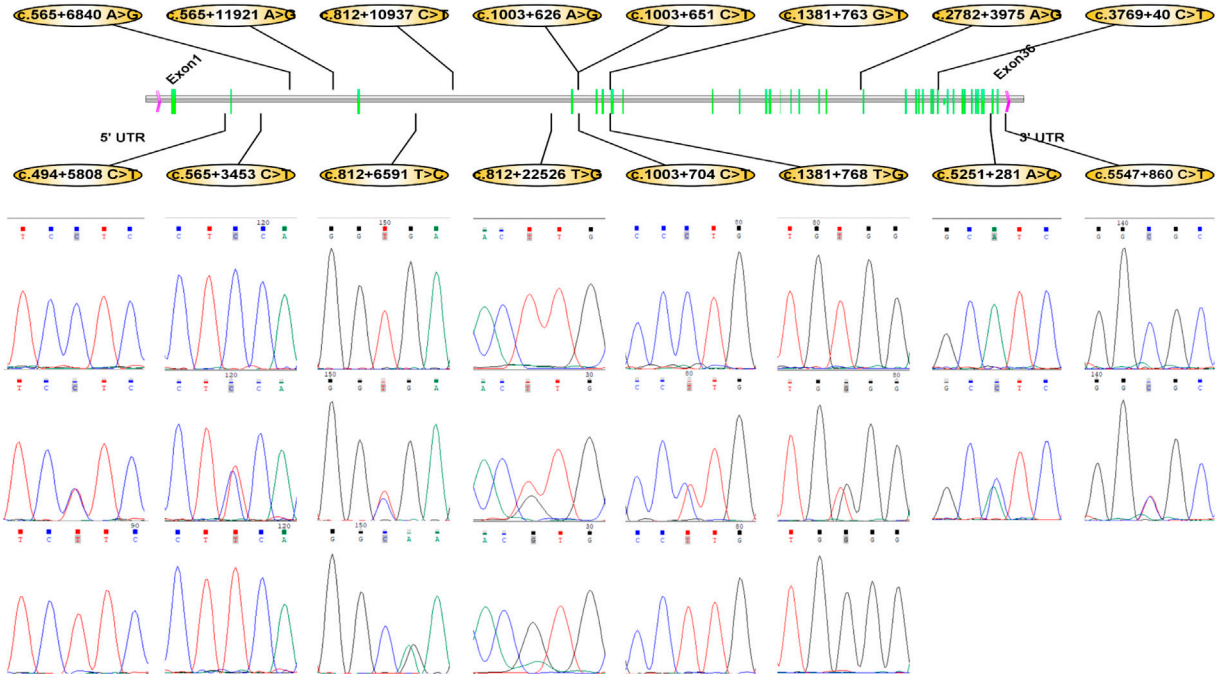


FIGURE 3
Sixteen SNVs of the *LTBP2* gene in Dezhou donkeys.

number, body size, and carcass traits, all samples were genotyped for these 16 alleles using GBTS liquid chip. The sequencing results of the PCR products amplified by specific primers were consistent with those of the GBTS liquid chip.

In Supplementary Table S4, three genotype frequencies (wild-type homozygote, heterozygotes, and mutant-type homozygote) and genetic indices (Obs-Het, Pred-Het, Ho, and PIC) in Dezhou donkeys' *LTBP2* gene are shown. As shown in Supplementary Table S4, we found that c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, and c.2782 + 3975 A > G did not have a mutant-type homozygote. The other 12 loci had three genotypes in total. The Hardy–Weinberg test results showed that the population was genetically balanced and belonged to the Mendelian population. The genetic diversity index showed that the PIC of Dezhou donkeys was 0.123–0.555, and the polymorphism was 1 < PIC < 0.25, which is in the moderate-high polymorphism state.

3.4 Association analysis of donkey vertebrae number, body size, and carcass traits with 16 SNV polymorphisms

The effects of the *LTBP2* SNVs on the number of vertebrae, body size, and carcass traits were evaluated and are presented in Supplementary Table S5. As shown in Supplementary Table S5, individuals with the wild-type homozygote had significantly more TN than heterozygotes ($p < 0.05$) in c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, and c.2782 + 3975 A > G loci, which means the main effect of these four loci are significant in TN. The lumbar vertebrae number in Dezhou donkeys with wild-type homozygous and mutant-type homozygous genotypes of c.1381 + 768 T > G and c.1381 + 763 G > T was 5.05 ± 0.05 and 5.25 ± 0.03, respectively, and the difference was significant ($p < 0.05$). The other two mutant loci significantly related to the number of lumbar vertebrae are c.812 + 10937 C > T and c.812 + 6591 T > C. In c.812 + 10937 C > T, individuals with mutant-type homozygous T/T and heterozygous C/T genotypes had more LN than wild-type homozygotes, while in c.812 + 6591 T > C, individuals with mutant-type homozygous C/C had significantly fewer LN than wild-type homozygotes and heterozygotes. As described in Supplementary Table S5, in the c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, and c.812 + 22526 T > G mutational sites, the effect of variation on CHC, CWF, CWA, and CW was consistent. Individuals with wild-type homozygotes had significantly higher data than other genotypes among the aforementioned four traits ($p < 0.05$), while others (BH, BL, SW, TN, TL, STL, LN, LL, SLL, TLN, TLL, and STLL) were not significant. There are five mutational sites significantly associated with TLN (c.812 + 10937 C > T, c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T). The effect of c.812 + 10937 C > T in TLN was

consistent with that in LN. In c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T mutational sites, individuals with the heterozygous genotype had the most TLN (23.12 ± 0.03) compared to other genotypes (significantly).

3.5 Linkage disequilibrium and haplotype analysis of *LTBP2* gene mutational sites

In order to reveal the linkage relationship between the 16 mutational sites of the *LTBP2* gene, the linkage disequilibrium between these loci was estimated. The analysis results showed that three blocks exist between these 16 SNVs ($r^2 > 0.33$) (Ardlie et al., 2002). Block1 is composed of c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, and c.2782 + 3975 A > G, block2 is composed of c.1381 + 768 T > G, c.1381 + 763 G > T, c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, and c.812 + 22526 T > G, and block3 is composed by c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T. The linkage coefficients for other pairs of SNVs were low ($r^2 < 0.33$).

Haplotype analysis was performed for the three blocks, and three haplotypes with a frequency greater than 0.05 were identified and labeled as Hap1 (-CACAGTTTGGAATT-), Hap2 (-CACAGT TTGGGGCC-), and Hap3 (-CACATGCCATGGCC-). The haplotype frequencies of Hap1, Hap2, and Hap3 in Dezhou donkeys account for 37.6%, 28.3%, and 10.2%, respectively.

3.6 Association analysis between the haplotype combinations and TLN, body size, and carcass traits

In order to explore the association between SNV polymorphism and the traits of Dezhou donkeys, haplotype combination was carried out on the basis of constructed haplotypes. Six haplotype combinations (Hap1Hap1, Hap1Hap2, Hap1Hap3, Hap2Hap2, Hap2Hap3, and Hap3Hap3) were constructed, respectively. The results are shown in Supplementary Table S6; individuals with Hap2Hap3 and Hap3Hap3 made up less than 5% of the total sample size, so no follow-up analysis was conducted for these two haplotype combinations. The correlation analysis between four haplotypes and TLN, body size, and carcass traits was calculated, but no significant difference was observed.

4 Discussion

With the development of agricultural mechanization, the causative function of donkeys is gradually weakening, which is the reason why the stock of donkeys in China decreased rapidly from 1990 to 2016 (Statistics, 2021). In recent years, luckily, with the further study of the special functions of donkey meat, milk, and skin, the situation has gradually improved (D'Auria et al., 2011; Li et al., 2006; Polidori et al., 2015; Polidori et al., 2008; Yvon et al., 2018). Donkeys are raised as a small special livestock in China today; the government has enacted many policies to support the development of the donkey industry, especially in breed preservation and improvement. Vertebrae number is an important characteristic in livestock, such as pigs, cattle, sheep, and donkeys, because a higher vertebrae number means a longer body length and a heavier carcass weight (Yang et al., 2016; Zhang et al., 2017). In this study, we observed that the TN, LN, and TLN in Dezhou donkeys is 17–19, 5–6, and 22–24, respectively; no T19L6 thoracolumbar vertebrae combination type individuals were observed, which is similar to what has been reported in previous studies (Jamdar and Ema, 1982; Gao, 2020; Gao et al., 2021). The effect of increasing TN or LN is not consistent; one more TN increases TLL by about 5.14 cm, while one more LN increases TLL by about 4.05 cm (significant at $p < 0.05$). Interestingly, the effect of one more TN in different LN populations is unequal, such as in the 5 lumber vertebrae number population, the CW of individuals with 18 TN is 11.1 kg more than individuals with 17 TN, while in 6 lumbar vertebrae number population, the CW of individuals with 18 TN is 3 kg more than individuals with 17 TN; the disparity between the increase between the two is 8.1 kg. Similarly, the effect of one more LN in different TN populations is also unequal.

The heritability of the vertebrae is high; for example, the heritability of vertebrae in pigs is 0.62 (Borchers et al., 2004). Therefore, we believe that directional breeding of Dezhou donkeys with multiple thoracolumbar vertebrae numbers and studying the molecular mechanism of multiple thoracolumbar vertebrae numbers is worthy of attention. The development of vertebrae is a complex process, which requires the regulation of multiple signal molecules and specific components. To date, there are several regions in the genome that have been identified as having the function of regulating the development of vertebrae. Mikawa et al. (2005), Mikawa et al. (2007), and Mikawa et al. (2011) identified two QTLs' influence on the number of vertebrae, mapped to SSC1 and SSC7. There is a 479 kb region on SSC7 including nine annotated genes, identified as a critical fragment influencing the TLN (Liu et al., 2020). The Notch receptor 1 (Notch-1) gene is a critical gene for somite development, while the activation of FOS could inhibit the expression of Notch-1 (Portanova et al., 2013; Liao and Oates, 2017).

In previous studies, many mutation sites in *LTBP2* have been proved to have some functions in the development of primary congenital glaucoma (c.3028G > A, p.Asp1010Asn; c.3427delC, p.Gln1143Argfs*35) (Rauf et al., 2020), cardiomyocytes (c.2206G > A, p.Asp736Asn) (Chen et al., 2020), and lung fibroblast-to-myofibroblast differentiation (Zou et al., 2021). *LTBP2* has also been found to be related to the number of teats in pigs (Martins et al., 2022). In this study, *LTBP2* was used as a candidate gene associated with TLN in Dezhou donkeys

based on related studies (Zhao et al., 2020; Wang et al., 2022). The liquid chip was designed based on published data from several databases and was used for genotyping all mutational loci of all samples. PCR primers were designed, based on a published donkey *LTBP2* sequence (GenBank accession NC-052183.1), to amplify the eight mutational loci (Wang et al., 2020). The alignments of the genotyping result between GBTS liquid chip and PCR products with specific primers is 100%. From this, we can see that the GBTS liquid chip sequencing results are reliable.

We actually detected a total of 544 mutational sites in the whole *LTBP2* gene of Dezhou donkeys; however, most of the loci cannot meet the requirements of correlation analysis, so we finally screened out 16 SNVs for subsequent analysis. Among these loci, the adjacent loci are significantly related to a certain trait ($p < 0.05$), centrally. For example, for c.5547 + 860 C > T, c.5251 + 281 A > C, c.3769 + 40 C > T, and c.2782 + 3975 A > G, the distance between them was 8,247 bp, 6,018 bp, and 9,253 bp, respectively, genome wide; they were all significantly associated with TN ($p < 0.05$). The analysis results from the Haploview software showed a strong linkage disequilibrium between them, which means when the C allele is mutated to the T allele at c.5547 + 860 C > T, the A allele at c.5251 + 281 A > C will mutate to C allele, the C allele at c.3769 + 40 C > T will mutate to T allele, and the A allele at c.2782 + 3975 A > G will mutate to G allele. In this block, individuals with wild-type homozygous had more 0.16 pieces of TN than heterozygotes, so we speculate that the mutation here has a negative effect on TN. The same situation also occurs in block2 (c.1381 + 768 T > G, c.1381 + 763 G > T, c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, and c.812 + 22526 T > G), but the function of block2 is aimed at LN, CHC, CWF, CWA, and CW. For the development of TN, mutations at c.1381 + 768 T > G and c.1381 + 763 G > T sites are beneficial, while for CHC, CWF, CWA, and CW, mutations at c.1003 + 704 C > T, c.1003 + 651 C > T, c.1003 + 626 A > G, and c.812 + 22526 T > G are negative. In block3, four SNVs (c.565 + 11921 A > G, c.565 + 6840 A > G, c.565 + 3453 C > T, and c.494 + 5808 C > T) were all significantly associated with TLN; individuals with the heterozygous genotype had higher TLN than others ($p < 0.05$), but the difference between wild-type homozygotes and mutant-type homozygotes was different.

In order to determine the linkage disequilibrium of 16 SNV markers located on the Dezhou donkey *LTBP2* gene, a total of three haplotype blocks among 16 SNVs were identified, examining haplotypes above 5%, spanning approximately 30 kb. The most frequent haplotypes within the blocks comprised the linkage of nucleotides CACA (96.4%) for block1, GTTTGG (72.8%) for block2, and AATT (53.6%) for block3. To demonstrate the total effect of a single locus, haplotype combinations were combined for further study. The haplotype combination types with a sample size less than 5% of the total sample were removed. Our results showed no dominant type in the four haplotype combinations; the reason may be the effect of these three blocks counteracting each other. These results indicate that the *LTBP2* gene is a potential functional gene for regulating the development of vertebrae in Dezhou donkeys, and this finding may provide important implications for further studies on the regulation of the TLN in Dezhou donkeys. It must be noted that our research has not explored the molecular mechanism of SNPs in *LTBP2* regarding how they effect TLN development.

5 Conclusion

In summary, the effects of TLN variation on body size and carcass traits of Dezhou donkeys were quantified, and 16 trait-related SNVs have been identified in the *LTBP2* gene of Dezhou donkeys. These loci have low-to-high polymorphism and there are significant differences between different genotypes among different traits. The results of this study will play an important role in the further study of the molecular mechanism of Dezhou donkey molecular breeding for muti-vertebrae breeds.

Data availability statement

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found at: NCBI; PRJNA878942.

Ethics statement

The animal study was reviewed and approved by the Animal Welfare and Ethics Committee of Institute of Animal Sciences, Liaocheng University (No. LC2019-1). Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

CW and ZL provided the study concept and design, ZL wrote and revised the manuscript, and TW, XS, XW, WR, and BH collected and analyzed the data. All authors have read and approved the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China (grant no. 31671287), the Well-bred Program of Shandong Province (grant no. 2017LZGC020), Taishan Leading Industry Talents, Agricultural Science of Shandong Province (grant no. LJNY201713), and Shandong Province Modern Agricultural Technology System Donkey Industrial Innovation Team (grant no. SDAIT-27).

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.2022.969959/full#supplementary-material

References

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3 (4), 299–309. doi:10.1038/nrg777

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263–265. doi:10.1093/bioinformatics/bth457

Borchers, N., Reinsch, N., and Kalm, E. (2004). The number of ribs and vertebrae in a Piétrain cross: Variation, heritability and effects on performance traits. *J. Anim. Breed. Genet.* 121, 392–403. doi:10.1111/j.1439-0388.2004.00482.x

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32 (3), 314–331.

Burgos, C., Latorre, P., Altarriba, J., Carrodeguas, J. A., Varona, L., and Lopez-Buesa, P. (2015). Allelic frequencies of NR6A1 and VRTN, two genes that affect vertebrae number in diverse pig breeds: A study of the effects of the VRTN insertion on phenotypic traits of a Duroc×Landrace–large white cross. *Meat Sci.* 100, 150–155. doi:10.1016/j.meatsci.2014.09.143

Chen, H. X., Yang, Z. Y., Hou, H. T., Wang, J., Wang, X. L., Yang, Q., et al. (2020). Novel mutations of TCTN3/LTBP2 with cellular function changes in congenital heart disease associated with polydactyly. *J. Cell. Mol. Med.* 24 (23), 13751–13762. doi:10.1111/jcmm.15950

D'Auria, E., Mandelli, M., Ballista, P., Di Dio, F., and Giovannini, M. (2011). Growth impairment and nutritional deficiencies in a cow's milk-allergic infant fed by unmodified donkey's milk. *Case Rep. Pediatr.* 2011, 103825. doi:10.1155/2011/103825

Donaldson, C. L., Lambe, N. R., Maltin, C. A., Knott, S., and Bunger, L. (2013). Between- and within-breed variations of spine characteristics in sheep. *J. Anim. Sci.* 91 (2), 995–1004. doi:10.2527/jas.2012-5456

Fan, Y., Xing, Y., Zhang, Z., Ai, H., Ouyang, Z., Ouyang, J., et al. (2013). A further look at porcine chromosome 7 reveals VRTN variants associated with vertebral number in Chinese and Western pigs. *Plos One* 8 (4), e62534. doi:10.1371/journal.pone.0062534

Gao, Q. (2020). *Study on Dezhou donkey germplasm*. Shandong: Shandong Teachers' University.

Gao, Q., Wang, J., Li, Y., Sun, Y., Yang, C., Li, H., et al. (2021). Preliminary study on the production performance of Dezhou donkey. *J. Livest. Ecol.* 42 (02), 56–61.

Gomez, C., Ozbudak, E. M., Wunderlich, J., Baumann, D., Lewis, J., and Pourquie, O. (2008). Control of segment number in vertebrate embryos. *Nature* 454 (7202), 335–339. doi:10.1038/nature07020

Greene, N. D., and Copp, A. J. (2009). Development of the vertebrate central nervous system: formation of the neural tube. *Prenat. Diagn.* 29 (4), 303–311. doi:10.1002/pd.2206

Jamdar, M. N., and Ema, A. N. (1982). A note on the vertebral formula of the donkey. *Br. Vet. J.* 138 (3), 209–211. doi:10.1016/s0007-1935(17)31084-9

King, J. W. B., and Roberts, R. C. (2010). Carcass length in the bacon pig; its association with vertebrae numbers and prediction from radiographs of the young pig. *Anim. Prod.* 2 (01), 59–65. doi:10.1017/s0003356100033493

Li, C., Zhang, X., Cao, Y., Wei, J., You, S., Jiang, Y., et al. (2017). Multi-vertebrae variation potentially contribute to carcass length and weight of Kazakh sheep. *Small Ruminant Res.* 150, 8–10. doi:10.1016/j.smallrumres.2017.02.021

Li, H., Huang, M. J., Zhang, S. Q., Ye, M. Y., and Rao, P. F. (2006). Major constituent proteins in donkey hide and their interaction. *Zhongguo Zhong Yao Za Zhi* 31 (8), 659–663.

Li, Z., Kawasumi, M., Zhao, B., Moisyadi, S., and Yang, J. (2010). Transgenic over-expression of growth differentiation factor 11 propeptide in skeleton results in transformation of the seventh cervical vertebra into a thoracic vertebra. *Mol. Reprod. Dev.* 77 (11), 990–997. doi:10.1002/mrd.21252

Liao, B. K., and Oates, A. C. (2017). Delta-Notch signalling in segmentation. *Arthropod Struct. Dev.* 46 (3), 429–447. doi:10.1016/j.asd.2016.11.007

Liu, Q., Yue, J., Niu, N., Liu, X., Yan, H., Zhao, F., et al. (2020). Genome-wide association analysis identified BMPR1A as a novel candidate gene affecting the number of thoracic vertebrae in a Large White × Minzhu intercross pig population. *Animals.* 10 (11), E2186. doi:10.3390/ani10112186

Liu, Z., Gao, Q., Wang, T., Chai, W., Zhan, Y., Akhtar, F., et al. (2022). Multi-thoracolumbar variations and NR6A1 gene polymorphisms potentially associated with body size and carcass traits of Dezhou donkey. *Animals.* 12 (11), 1349. doi:10.3390/ani12111349

Martins, T. F., Braga Magalhães, A. F., Verardo, L. L., Santos, G. C., Silva Fernandes, A. A., Gomes Vieira, J. I., et al. (2022). Functional analysis of litter size and number of teats in pigs: From GWAS to post-GWAS. *Theriogenology* 193, 157–166. doi:10.1016/j.theriogenology.2022.09.005

Mikawa, S., Hayashi, T., Nii, M., Shimanuki, S., Morozumi, T., and Awata, T. (2005). Two quantitative trait loci on *Sus scrofa* chromosomes 1 and 7 affecting the number of vertebrae. *J. Anim. Sci.* 83 (10), 2247–2254. doi:10.2527/2005.83102247x

Mikawa, S., Morozumi, T., Shimanuki, S., Hayashi, T., Uenishi, H., Domukai, M., et al. (2007). Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (NR6A1). *Genome Res.* 17 (5), 586–593. doi:10.1101/gr.6085507

Mikawa, S., Sato, S., Nii, M., Morozumi, T., Yoshioka, G., Imaeda, N., et al. (2011). Identification of a second gene associated with variation in vertebral number in domestic pigs. *BMC Genet.* 12, 5. doi:10.1186/1471-2156-12-5

Nei, M., and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* 76 (2), 379–390. doi:10.1093/genetics/76.2.379

Park, S., Lee, Y. J., Lee, H. J., Seki, T., Hong, K. H., Park, J., et al. (2004). B-cell translocation gene 2 (Btg2) regulates vertebral patterning by modulating bone morphogenetic protein/Smad signaling. *Mol. Cell. Biol.* 24 (23), 10256–10262. doi:10.1128/MCB.24.23.10256-10262.2004

Polidori, P., Pucciarelli, S., Ariani, A., Polzonetti, V., and Vincenzetti, S. (2015). A comparison of the carcass and meat quality of Martina Franca donkey foals aged 8 or 12 months. *Meat Sci.* 106, 6–10. doi:10.1016/j.meatsci.2015.03.018

Polidori, P., Vincenzetti, S., Cavallucci, C., and Beghelli, D. (2008). Quality of donkey meat and carcass characteristics. *Meat Sci.* 80 (4), 1222–1224. doi:10.1016/j.meatsci.2008.05.027

Portanova, P., Notaro, A., Pellerito, O., Sabella, S., Giuliano, M., and Calvaruso, G. (2013). Notch inhibition restores TRAIL-mediated apoptosis via AP1-dependent upregulation of DR4 and DR5 TRAIL receptors in MDA-MB-231 breast cancer cells. *Int. J. Oncol.* 43 (1), 121–130. doi:10.3892/ijo.2013.1945

Rauf, B., Irum, B., Khan, S. Y., Kabir, F., Naeem, M. A., Riazuddin, S., et al. (2020). Novel mutations in LTBP2 identified in familial cases of primary congenital glaucoma. *Mol. Vis.* 26, 14–25.

Ren, D. R., Ren, J., Ruan, G. F., Guo, Y. M., Wu, L. H., Yang, G. C., et al. (2012). Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc × Chinese Erhualian intercross resource population. *Anim. Genet.* 43 (5), 545–551. doi:10.1111/j.1365-2052.2011.02313.x

Rijli, F. M., Matyas, R., Pellegrini, M., Dierich, A., Gruss, P., Dollé, P., et al. (1995). Cryptorchidism and homeotic transformations of spinal nerves and vertebrae in Hoxa-10 mutant mice. *Proc. Natl. Acad. Sci. U. S. A.* 92 (18), 8185–8189. doi:10.1073/pnas.92.18.8185

Statistics, C. B. o. (2021). *China statistical yearbook*. Beijing, China: China Statistics Press.

Thawait, G. K., Chhabra, A., and Carrino, J. A. (2012). Spine segmentation and enumeration and normal variants. *Radiol. Clin. North Am.* 50 (4), 587–598. doi:10.1016/j.rcl.2012.04.003

Wang, C., Li, H., Guo, Y., Huang, J., Sun, Y., Min, J., et al. (2020). Donkey genomes provide new insights into domestication and selection for coat color. *Nat. Commun.* 11 (1), 6014. doi:10.1038/s41467-020-19813-7

Wang, X., Ran, X., Niu, X., Huang, S., Li, S., and Wang, J. (2022). Whole-genome sequence analysis reveals selection signatures for important economic traits in Xiang pigs. *Sci. Rep.* 12 (1), 11823. doi:10.1038/s41598-022-14686-w

Wellik, D. M. (2007). Hox patterning of the vertebrate axial skeleton. *Dev. Dyn.* 236 (9), 2454–2463. doi:10.1002/dvdy.21286

Yang, J., Huang, L., Yang, M., Fan, Y., Li, L., Fang, S., et al. (2016). Possible introgression of the VRTN mutation increasing vertebral number, carcass length and teat number from Chinese pigs into European pigs. *Sci. Rep.* 6, 19240. doi:10.1038/srep19240

Yvon, S., Olier, M., Leveque, M., Jard, G., Tormo, H., Haimoud-Lekhal, D. A., et al. (2018). Donkey milk consumption exerts anti-inflammatory properties by normalizing antimicrobial peptides levels in Paneth's cells in a model of ileitis in mice. *Eur. J. Nutr.* 57 (1), 155–166. doi:10.1007/s00394-016-1304-z

Zhang, L. C., Yue, J. W., Pu, L., Wang, L. G., Liu, X., Liang, J., et al. (2016). Genome-wide study refines the quantitative trait locus for number of ribs in a Large White × Minzhu intercross pig population and reveals a new candidate gene. *Mol. Genet. Genomics* 291 (5), 1885–1890. doi:10.1007/s00438-016-1220-1

Zhang, X., Li, C., Li, X., Liu, Z., Hu, S., Cao, Y., et al. (2019). Association analysis of polymorphism in the NR6A1 gene with the lumbar vertebrae number traits in sheep. *Genes Genomics* 41 (10), 1165–1171. doi:10.1007/s13258-019-00843-5

Zhang, Z., Sun, Y., Du, W., He, S., Liu, M., and Tian, C. (2017). Effects of vertebral number variations on carcass traits and genotyping of Vertnin candidate gene in Kazakh sheep. *Asian-Australas. J. Anim. Sci.* 30 (9), 1234–1238. doi:10.5713/ajas.16.0959

Zhang, Z., Zhan, Y., Han, Y., Liu, Z., Wang, Y., and Wang, C. (2021). Estimation of liveweight from body measurements through best fitted regression model in Dezhou donkey breed. *J. Equine Vet. Sci.* 101, 103457. doi:10.1016/j.jevs.2021.103457

Zhao, F., Deng, T., Shi, L., Wang, W., Zhang, Q., Du, L., et al. (2020). Genomic scan for selection signature reveals fat deposition in Chinese indigenous sheep with extreme tail types. *Animals.* 10 (5), E773. doi:10.3390/ani10050773

Zou, M., Zou, J., Hu, X., Zheng, W., Zhang, M., and Cheng, Z. (2021). Latent transforming growth factor-β binding protein-2 regulates lung fibroblast-to-myofibroblast differentiation in pulmonary fibrosis via NF-κB signaling. *Front. Pharmacol.* 12, 788714. doi:10.3389/fphar.2021.788714

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fvets.2022.1045589

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
George R. Wiggans,
Council on Dairy Cattle Breeding, United States
Juan José Arranz,
Universidad de León, Spain

*CORRESPONDENCE
Amir Rashidi
arashidi@uok.ac.ir
Jalal Rostamzadeh
j.rostamzadeh@uok.ac.ir
Mohammad Razmkabir
m.razmkabir@uok.ac.ir

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

RECEIVED 15 September 2022
ACCEPTED 02 November 2022
PUBLISHED 23 November 2022

CITATION
Mahmoudi P, Rashidi A, Nazari-Ghadikolaei A, Rostamzadeh J, Razmkabir M and Huson HJ (2022) Genome-wide association study reveals novel candidate genes for litter size in Markhoz goats.
Front. Vet. Sci. 9:1045589.
doi: 10.3389/fvets.2022.1045589

Genome-wide association study reveals novel candidate genes for litter size in Markhoz goats

Peyman Mahmoudi[1], Amir Rashidi[1]*, Anahit Nazari-Ghadikolaei[2], Jalal Rostamzadeh[1]*, Mohammad Razmkabir[1]* and Heather Jay Huson[3]

[1]Department of Animal Science, Faculty of Agriculture, University of Kurdistan, Sanandaj, Iran, [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3]Department of Animal Science, Cornell University, Ithaca, NY, United States

Introduction: The Markhoz goat is the only breed that can produce high-quality fiber called mohair in Iran; however, the size of its population has faced a dramatic decline during the last decades, mainly due to the reluctance of farmers to rear Markhoz goats caused by a reduction in goat production income. Litter size at birth (LSB) and weaning (LSW) are two economically important reproductive traits for local goat breeders and have the potential of increasing the population growth rate. The present study was aimed to identify possible genomic regions that are associated with LSB and LSW in Markhoz goats using a genome-wide association study (GWAS).

Methods: To this end, 136 Markhoz goats with record(s) of kidding were selected for GWAS using the Illumina Caprine 50K bead chip. The individual breeding values (BV) of available LSB and LSW records estimated under an animal mixed model were used as the dependent variable in the GWAS, thereby incorporating repeated categorical variables of litter size.

Results: Four SNPs on chromosomes 2, 20 and 21 were identified to be significantly associated (FDR $p < 0.05$) with LSB after multiple testing correction under a Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) model. Least-square analysis was performed to investigate the effects of detected genotypes on LSB. Ultimately, the GWAS results introduced six candidate genes, including *GABRA5, AKAP13, SV2B, PPP1R1C, SSFA2* and *TRNAS-GCU* in a 100 kb adjacent region of the identified SNPs. Previous studies proposed functional roles of *GABRA5* and *AKAP13* genes in reproductive processes; however, the role of other candidate genes in reproduction is not clear.

Conclusion: These findings warrant further investigation for use in marker-assisted selection programs in Markhoz goats.

KEYWORDS

genome-wide association study, litter size, Markhoz goat, reproduction, prolificacy

Introduction

Throughout history, goats have been a primary production species for mankind due to their ability to withstand variable and harsh environmental conditions and their desirable production of meat, milk, fiber, and skin. Improving economic traits in goats, such as prolificacy and viability, can be very profitable for rural people in countries with low-quality grazing lands, where goat farming is one of the primary sources of income.

Markhoz goats are the only mohair producing breed in Iran with different coat colors such as white, black, and various shades of brown (1). The population size of the Markhoz goat underwent a considerable decrease during the last two decades (2), such that only about 1,000 head remain in their main native habitat in western and northwestern Iran. The main reasons for the decreased population of the indigenous Markhoz goat include changes in the management system for how animals are reared and the generally low income gained by goat farming in local regions. Hence, increasing the number of kids born per kidding and subsequently increasing the total income through the selling of kids, fiber, and meat may encourage ranchers toward goat production.

Litter size at birth (LSB) and litter size at weaning (LSW) are two reproductive traits that are known to be controlled by several underlying genes in goats (3). Specifically, genetic variants with significant effects on LSB have been identified in genes including, *GDF9, BMP15, GnRH1, KISS1, KITLG, NGF, POU1F1, PRLR* and promoter of *miR-9* gene in various breeds of goats (4–11). However, there may be other genes that affect LSB and LSW that have remained unknown. Nowadays, with the development and availability of SNP genotyping technologies, conducting genome-wide association studies (GWASs) and detecting candidate genes and genetic variants that may have a significant association with economic traits, have become much easier and faster. In this regard, many GWASs have been conducted using the Illumina Caprine 50K beadchip [Illumina Inc., San Diego, CA (12)] for various economically important traits in different goat breeds including coat color and mohair traits (13), body morphological traits (14), conformation and milk yield (15), and resistance to nematodes (16). Similarly, multiple researchers have conducted GWAS on liter size at birth in sheep (17–19). However, there is only one GWAS for the number of kids alive per kidding in goat (20) and no GWAS for the number of kids alive till weaning in sheep or goat.

Detection of significantly associated SNPs with LSB and LSW in goat would lead to enhanced efficacy of animal selection in breeding strategies by reducing the cost and time required to raise and phenotypically characterize animals as they mature. As complex traits, LSB and LSW are known to be controlled by many genes with variants having a small effect. Hence, the possibility of incorporating genetic variants into selection strategies will likely accelerate the rate of improvement of such reproductive traits as compared to traditional phenotypic selection. The purpose of this study was to conduct GWAS to detect possible genomic regions and variants associated with LSB and LSW in Markhoz goats, with the potential of applying results in genomic selection.

Materials and methods

Animals and phenotypes

All female goats ($n = 184$) existing at the Markhoz Goat Performance Testing Station in Sanandaj, Kurdistan, Iran, were selected for inclusion in the study. The herd is reared under a semi-intensive management system, in which animals graze on natural pastures from spring to early autumn and fed a diet consisting of alfalfa and wheat straw for the rest of the year. At the age of 16–18 months, does are mated for the first time. The kidding season starts in late winter and ends in early spring. Litter size at birth (LSB) and litter size at weaning (LSW) were the two reproductive traits evaluated. LSB described the number of live kids born to the doe. LSW described the number of live kids at weaning for each doe, typically evaluating kids at 22–27 weeks old. LSB and LSW were categorical variables potentially repeated per doe as they aged and had subsequent litters.

Statistical analyses

Prediction of breeding values

Predicted breeding values (PBV) for LSB and LSW were generated for use in the GWAS to capture the repeated categorical values of LSB and LSW. A total of 3,410 litter size records for the Markhoz goats were collected from 1994 to 2019 for use in predicting breeding values. Accuracy of EBVs is based on the amount of performance information available on the animal and its close relatives. Selection using EBVs is more accurate, especially for low heritable traits like litter size, which allows for faster genetic gain compared to mass selection using phenotypes. Breeding values have the advantage that they are free of systematic environmental effects on measured phenotypes, as these effects are considered in the statistical model used for the estimation of EBVs. Additionally, they reflect the genetic makeup more accurately because they do not solely rely on own records but include information from all measured relatives. The pedigree file included 5,396 animals with 1,533 dams and 252 sires. The number of founders, individuals with progeny, and individuals without progeny were 343, 1,785 and 3,611, respectively. Breeding values for each individual was estimated, applying a repeatability threshold animal model using ASReml 2.0 (21) as follows:

$$y = Xb + Za + Wpe + e$$

where y is a vector of phenotypic value for LSB/LSW, b, a, pe and e are vectors of fixed effects including year of kidding (2010–2019), age of dam (2–9 years) and parity (1–7), random animal effects, random permanent environmental effects and random residual effects, respectively. X, Z, and W are design matrices that relate records to fixed, animal and permanent environmental effects, respectively.

Genotyping and quality control

All animal procedures were approved by the Cornell University Institutional Animal Care and Use Committee prior to sampling (protocol #2014-0121). Vacutainer tubes containing K_2EDTA as an anticoagulant were used to collect whole blood (5 ml) samples from the jugular vein of goats. Samples were immediately stored at −20°C until DNA extraction. A standard Phenol-Chloroform DNA extraction method was utilized for extracting genomic DNA (22). The Illumina Caprine 50K beadchip (Illumina, Inc., San Diego, CA, United States), including 53,353 SNPs, was used for genotyping samples (VHL Genetics, Wageningen, Netherlands). Golden Helix SVS v8.3.4 (Golden Helix, Bozeman, MT, United States) software was used for quality control process as follows: (1) 624 SNPs were removed for a call rate less than 0.9; (2) 2,540 SNPs with a minor allele frequency less than 0.03 were excluded; (3) 810 SNPs were not assigned to a genomic location; thus they were removed; (4) five samples were removed for a genotyping call rate less than 0.9; and (5) three samples with an estimated identity-by-state (IBS) score greater than 0.9 were removed to eliminate the possible effects of substantial relatedness between individuals on the overall results. Furthermore, 39 samples were excluded because they had no history of kidding in the data set. After the quality control process, 49,764 SNPs and 136 animals remained for GWAS.

Genome-wide association studies

The GAPIT v3.0 R package was used for investigating the association between genomic regions and PBVs for LSB and LSW as phenotypes. Several models including General Linear Model (GLM), Mixed Linear Model (MLM), Multi Locus Mixed Model (MLMM), Compressed Mixed Linear Model (CMLM), Enriched Compressed Mixed Linear Model (ECMLM), Factored spectrally transformed Linear Mixed Model (Fast-LMM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and random model Circulating Probability Unification (FarmCPU), Efficient Mixed-Model Association (EMMA), Efficient Mixed Model Association eXpedited (EMMAX) and BLINK were performed to find the best model for fitting PBV LSB and LSW data (Figure 2). The BLINK (23) and FarmCPU (24) models showed less deviation from expectation in the Q-Q plots than other models. Despite FarmCPU identifying more significantly associated SNPs in the model comparison, BLINK was selected for the final GWAS because it controlled both false positives and false negatives effectively, showing a sharp upward deviated tail and a straight line close to the 1:1 line. In the BLINK method, markers in LD ($r^2 > 0.7$) with the most significant marker are excluded from the analysis and the maximum likelihood of a random effect model is approximated by using the Bayesian Information Content (BIC) of a fixed-effects model. By applying that, the most significant markers will be selected among all markers that remained after LD exclusion and then used as cofactors in the model to test all markers across the genome. Because of the abovementioned reasons, the BLINK method had improved statistical power and better control on false positives in comparison with kinship-based methods (23). In the present study, the first 10 PCs explained about 23% of stratification, of which 12% was explained by the first three PCs. For considering population structure and avoiding biases due to population stratification in the present GWAS, only the first three PCs were included as covariates in the model, because no difference observed in the results when 4–10 top PCs (explained about 11% of stratification) were included in the analyses.

Significance of marker association was determined using a false discovery rate (FDR) adjusted p-value of less than 0.05. The ggplot R package was used to generate Q–Q and Manhattan plots (25). The genomic heritability was estimated using LDAK software v5.1 (26).

Least-square analysis and correlations between studied traits

Least-square analyses were conducted to investigate the negative/positive effects of detected genotypes on studied traits using PROC MIXED of SAS v8.2 software (27). Phenotypic and genotypic correlations between LSB and LSW were estimated using CORR procedure of SAS v8.2 and GCTA software tool (28), respectively.

Identifying candidate genes

A region with a distance of 100 kbp up-stream and 100 kbp down-stream of significant SNPs was explored to detect the nearest gene using *Capra hircus* ARS1 assembly (29) in the NCBI database. The LD decay pattern of the Markhoz goat population was estimated and the explored distance was selected based on the intersection point between the LD line and the r^2 threshold determined the LD decay value ($r^2 < 0.1$). Finally, gene databases, such as Genecards, National Center for Biotechnology Information (NCBI), Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were scrutinized to find the functions and pathway of identified genes.

TABLE 1 Descriptive statistics and estimated breeding values for litter size at birth and litter size at weaning in the Markhoz goat.

Statistic	LSB	LSB PBV	LSW	LSW PBV
n	137	5,396	137	5,396
Mean (± SD)	1.16 (± 0.37)	0.0031 (± 0.0146)	0.99 (± 0.50)	0.0073 (± 0.0381)
Min	1	−0.029	0	−0.124
Max	3	0.068	2	0.142

LSB, litter size at birth; PBV, predicted breeding value; LSW, litter size at weaning.

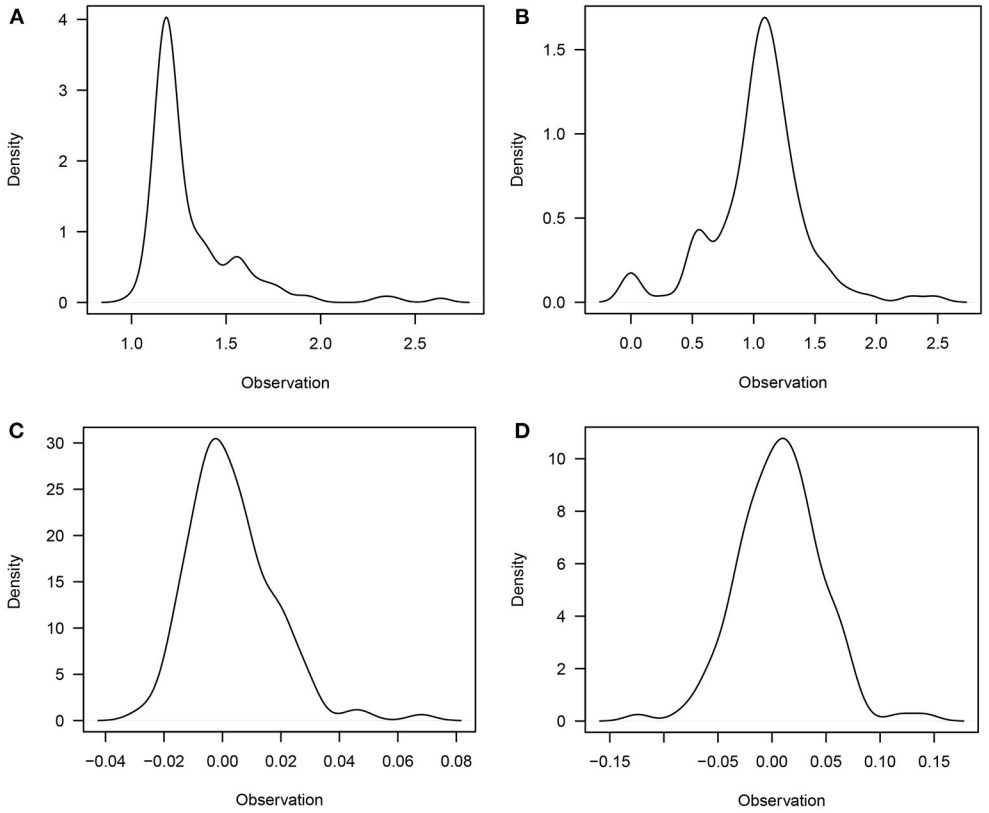


FIGURE 1
The distribution of litter size at birth **(A)**; litter size at weaning **(B)**; PBVs for litter size at birth **(C)**; and PBVs for litter size at weaning **(D)**.

Results

Descriptive statistics and predicted breeding values

The descriptive statistics and predicted breeding values for LSB and LSW traits are presented in Table 1. The studied population had an average LSB of 1.16, while only 0.99 of them remained alive at the age of weaning (LSW). The mean PBV for LSB and LSW was 0.0031 and 0.0073, respectively. Furthermore, the distribution of studied traits is depicted in Figure 1.

Genome-wide association studies

The Manhattan and Q-Q plots for PBV of LSB and LSW are depicted in Figure 2. Furthermore, the Q-Q plot for litter size at birth fitting other models is indicated in Figure 3. Four significantly associated SNPs on chromosome 2 (rs268267345, unadjusted p-value = 1.35e−07), chromosome 20 (rs268258357, unadjusted p-value = 1.33 e−08) and chromosome 21 (rs268288690, unadjusted p-value = 5.41e−13; rs268256209, unadjusted p-value = 2.56e−07) were identified for PBV LSB (Figure 2A). The FDR adjusted p-values for the identified SNPs were 0.002, 0.0003, 2.69e−08 and 0.003,

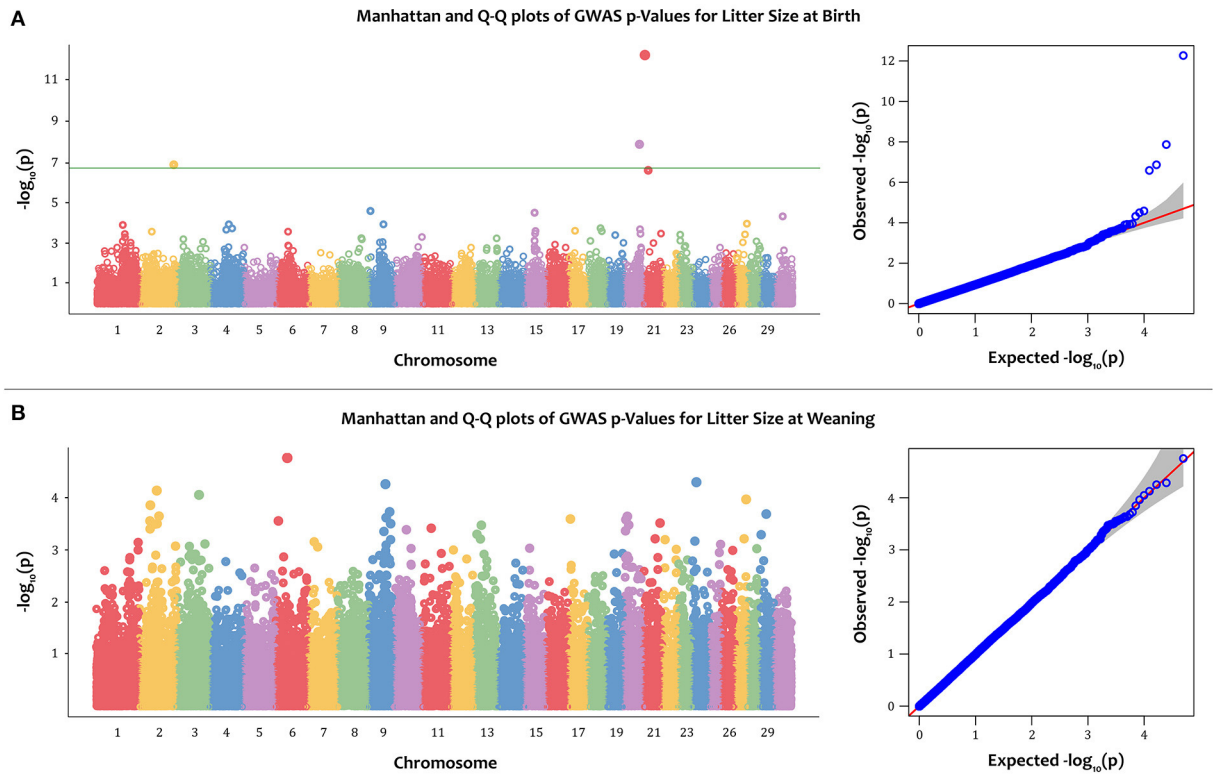


FIGURE 2
The Manhattan and quantile–quantile (Q–Q) plots of -log 10 (*p*-value) for litter size at birth (LSB) **(A)**; and litter size at weaning (LSW) **(B)** in the Markhoz goats using the BLINK GWAS model. The green horizontal line indicates the FDR adjusted *p*-value threshold of 0.05.

respectively. The Q–Q plot for LSB showed a very slight genomic inflation factor with $\lambda_{GC} = 1.01$. In contrast, no significantly associated SNPs were found for PBV LSW after multiple testing correction under the various models tested for this trait. The detailed information, including location, alleles, and p-values, for significantly associated SNPs on the PBV LSB trait is provided in Table 2. Considering that the distance between the significantly associated SNPs from their neighboring SNPs was farther than 500 kb and the BLINK model uses a minimum distance of 300 kb for exclusion of markers, linkage disequilibrium analysis was not performed.

Least-square analysis of identified genotypes for LSB and phenotypic and genotypic correlations

The results of the least-square analyses for LSB are provided in Table 3. Results show that markers rs268267345, rs268258357, and rs268288690 in the genome of the Markhoz goats significantly resulted in increased litter size in goats having one or two mutated alleles. In contrast, rs268256209 SNP has a significant negative effect on LSB in goats that carry two G alleles.

The phenotypic correlation between LSB and LSW was 0.697 ($P < 0.01$). Furthermore, analysis of the genotypic correlation between these traits showed a strong genetic relationship of 0.725 ($P < 0.01$).

Estimated genomic heritability

The estimated heritability using recorded data and pedigree for LSB and LSW of the entire population was 0.018 and 0.32e−6, respectively. However, the genomic heritability for studied traits was 0.011 and 0.21e−7, respectively. The detailed information for variance components of LSB and LSW traits is presented in Table 4.

Detected candidate genes

Six candidate genes were found within the 100 kb windows up- and down-stream of the identified SNPs located on the genome of *Capra hircus* in the NCBI database (ARS1 assembly,

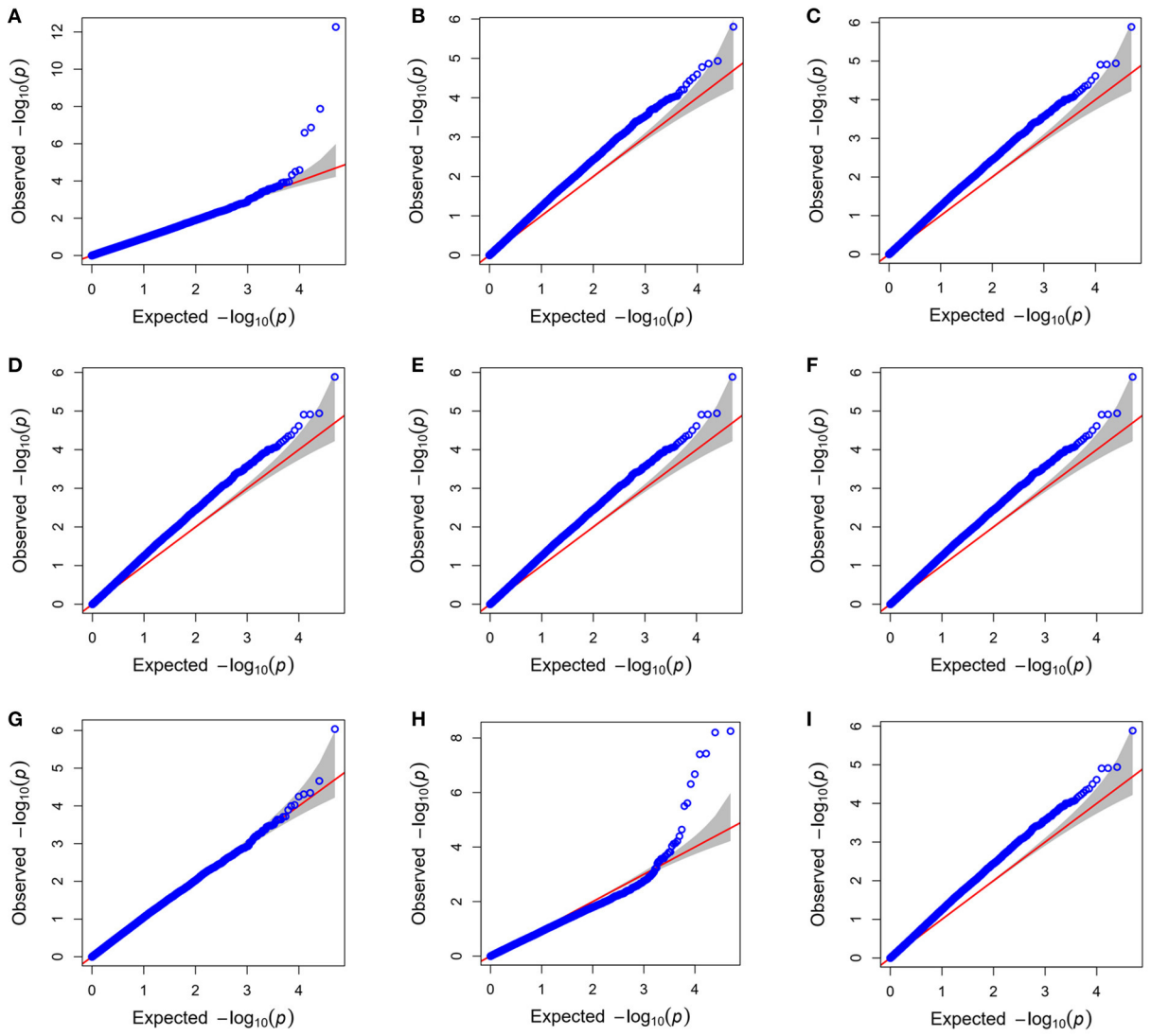


FIGURE 3
The Q–Q plot for litter size at birth (LSB) fitting Bayesian-information and Linkage-disequilibrium iteratively nested keyway (BLINK) **(A)**; general linear model (GLM) **(B)**; mixed linear model (MLM) and multi locus mixed model (MLMM) **(C)**; compressed mixed linear model (CMLM) **(D)**; enriched compressed mixed linear model (ECMLM) **(E)**, factored spectrally transformed linear mixed model (Fast-LMM) **(F)**; settlement of MLM under progressively exclusive relationship (SUPER) **(G)**; fixed and random model circulating probability unification (FarmCPU) **(H)**; and efficient mixed-model association (EMMA) and efficient mixed model association eXpedited (EMMAX) **(I)** GWAS models.

accession number: GCF_001704415.1). The detected genes harboring SNPs, their distance from SNPs and the roles of each gene are presented in Table 5.

Discussion

In Iran, the Markhoz goat is the only breed that can produce mohair in black, gray, white and varying shades of brown color. The fiber obtained from the Markhoz goat is both culturally and socially important for Kurdish people costumes (30). Thus, the decreasing population size of this valuable breed could be mitigated by using breeding programs incorporating genetic markers affecting reproductive traits such as litter size at birth (LSB). This is the first study on litter size traits in the Iranian goats at a genome-wide scale which can be employed as a practical tool to identify novel genetic markers that may influence litter size in the Markhoz and other goat breeds.

For both traits, the first three principal components (PCs) were included in the model to correct for population structure. A total of four SNPs, found on chromosomes 2, 20, and 21, were significantly associated with the PBV of LSB (Figure 2A). The Q–Q plots showed that observed versus expected data was well aligned, indicating minimal population stratification affecting

TABLE 2 Genome-wide association study identifies four SNPs significantly associated with the predicted breeding value of litter size at birth (LSB) in Markhoz goats.

Chr	SNP name	SNP id	Pos (bp)	Alleles	Unadj-P	FDR
2	snp35221-scaffold422-519219	rs268267345	121791230	A/G	1.35e−07	0.002
20	snp25995-scaffold269-1405685	rs268258357	55321621	A/G	1.33e−08	0.0003
21	snp57162-scaffold91-2213872	rs268288690	3499131	A/G	5.41e−13	2.69e−08
21	snp23799-scaffold240-558706	rs268256209	15409777	A/G	2.56e−07	0.003

Chr, chromosome number; Pos, chromosome position; Unadj-P, unadjusted *p*-value; FDR, false discovery rate.

the model ($\lambda_{GC} = 1.01$). In contrast, the results of the GWAS failed to reveal significant association with the PBV of LSW (Figure 2B).

The most significantly associated variant (rs268288690) with the PBV of LSB is located within the *GABRA5* gene. *GABRA5* encodes a 462 amino acid protein of the *GABA* receptors family known as *GABA-A*. It has been suggested that *GABA* acts at *GABA-A* receptors in the central and peripheral nervous systems as the major inhibitory neurotransmitter (31). Watanabe et al. (32) investigated the role of *GABA* in the regulation of GnRH neurons. They reported that in the course of fetal development, *GABA* is involved in the regulation of GnRH neuron migration from the olfactory placodes into the forebrain. They also stated that negative and positive feedback of estradiol are mediated by *GABA*, and there is a significant correlation between these feedbacks and frequency of *GABA* transmission to GnRH neurons. In a recent study, Di Giorgio et al. (33) demonstrated that *GABA-A* and *GABABRs* interact with kisspeptin (a protein encoded by the *Kiss1* gene) in the regulation of reproductive processes. For instance, *GABA* increases *Kiss1* expression by affecting *GABA-A* receptors in early embryo development. In addition, at the time of ovulation in adults, a main double excitatory input to (GnRH) neurons is provided by the AVPV/PeN neuron population, leading to the expression of *GABA* and kisspeptin.

Another significantly associated SNP identified in the present study (rs268256209) is located 99,644 bp upstream of the *AKAP13* gene. *AKAP13* is a member of the *AKAP* family which is a structurally diverse protein and is involved in the binding process to the regulatory subunit of protein kinase A (PKA) and confining the holoenzyme to discrete locations within the cell. Luconi et al. (34) reported that *AKAP* proteins are expressed in both female and male reproductive systems, especially during gametogenesis. It has been suggested that *AKAP–PKA* interactions control the maturation of oocytes (35).

Based on the functional role of *PPP1R1C*, *SV2B, TRNAS-GCU* and *SSFA2* genes, there is no evidence to suspect a causative association with the LSB trait in our study. Whereas, both *GABRA5* and *AKAP13*, are the more likely genes potentially influencing LSB based on their influence of reproductive processes.

TABLE 3 Genotypic frequency and least-square mean ± standard error of litter size at birth for identified SNPs in the Markhoz goats.

SNP id	Genotype	*N*	LSMean[1] ± SE	*p*-value
rs268267345	AA	65	$1.19^{b} \pm 0.04$	0.048
	AG	58	$1.25^{ab} \pm 0.04$	
	GG	14	$1.35^{a} \pm 0.07$	
rs268258357	AA	115	$1.20^{c} \pm 0.03$	<0.0001
	AG	21	$1.31^{b} \pm 0.05$	
	GG	1	$1.82^{a} \pm 0.16$	
rs268288690	AA	115	$1.17^{b} \pm 0.04$	<0.0001
	AG	20	$1.37^{a} \pm 0.05$	
	GG	2	$1.48^{a} \pm 0.15$	
rs268256209	AA	33	$1.30^{a} \pm 0.05$	0.001
	AG	74	$1.24^{a} \pm 0.04$	
	GG	30	$1.10^{b} \pm 0.05$	

[1]Different letters indicate a statistically significant difference between genotypic groups. Three genotypes of each locus are compared to each other.

The results of least-square analyses showed that rs268288690 SNP leads to a significant increase in litter size in Markhoz goats ($p < 0.01$) so that goats having two copies of the mutated allele (GG genotype) had more kids within the litter than those carrying one or no copy of the mutated allele (AG and AA genotypes). The GG genotype of rs268258357 SNP indicated the highest litter size (1.822 ± 0.161) among all identified genotypes. Similarly, rs268267345 SNP also had a positive impact on litter size. These findings revealed that only rs268256209 SNP negatively affected litter size, while the other three SNPs identified from GWAS positively influenced the number of kids in Markhoz goats.

Estimated heritability for LSB and LSW was 0.018 and 0.32e−6, respectively. These values are lower than the estimated genomic heritability of 0.011 for LSB and 0.21e−7 for LSW. One of the main reasons for the observed differences could be different sample sizes used for estimating heritability *via* BLUP and LDAK models. In BLUP, we used all available data (3,410) and pedigree (5,396) records, while only 136 genotyped individuals were used in the LDAK method to predict genomic heritability. Besides, much of the heritability of traits may not be

TABLE 4 Estimation of variance components and genetic parameters for litter size at birth (LSB) and litter size at weaning (LSW) in the Markhoz goats.

Trait	σ_a^2	σ_{pe}^2	σ_e^2	h^2(±SE)	h_G^2	R
LSB	0.0024	0.49e−7	0.1278	0.019 (±0.028)	0.011	0.018
LSW	0.82e−7	0.0140	0.2421	0.32e−6 (±0.14e−6)	0.21e−7	0.055

σ_a^2, additive genetic variance; σ_{pe}^2, permanent environmental variance; σ_e^2, residual variance; h^2, whole population heritability; h_G^2, genomic heritability; R, repeatability.

TABLE 5 Genes within 100 kb distance from identified significant SNPs and their description.

SNP ID	Candidate gene	Distance*	Gene description
rs268267345	*PPP1R1C*	Intron 2	Protein phosphatase 1 regulatory inhibitor subunit 1C
	SSFA2	−98,584	Sperm-specific Antigen 2
rs268258357	*TRNAS-GCU*	−71,861	Transfer RNA Serine (Anticodon GCU)
rs268288690	*GABRA5*	Intron 6	Gamma-aminobutyric acid type a receptor subunit Alpha5
rs268256209	*SV2B*	−41,296	Synaptic vesicle glycoprotein 2B
	AKAP13	+99,644	A-Kinase anchoring protein 13

*The distance from identified SNP.

accounted for by rare, low-frequency genetic variants, known as the missing heritability problem (36).

There are many genes that have been identified as associated with litter size in goat and sheep. Among them, *GDF9, BMP15, BMPR1B, and IGF1* genes have been widely discussed in literature acknowledging their effects on litter size. However, to the best of our knowledge, the associations between candidate genes detected in the present study and litter size have not been reported previously. Therefore, according to their vital functions in reproductive processes, the *GABRA5* and *AKAP13* genes could be important novel candidate genes for litter size in small ruminants. However, our study had some limitations, including small sample size and relatively low heritability of studied traits. Thus, more genotyped animals are required to validate the impact of these potential candidate genes on litter size in goats.

Additionally, due to the high genomic correlation between LSB and LSW traits, we expected to find some common regions for the two traits, but the GWAS for the PBV for LSW failed to detect any significantly associated genomic regions. One possible reason could be that LSW is affected by environmental conditions or the pattern of effects may be altered by environment and genetic interactions. Furthermore, LSW may be influenced by rare causal variants that are not included in the current goat medium-density SNP-chip and not captured by linkage disequilibrium. Additionally, the complexity of gene interactions on this trait, such as epistasis, was not considered in this study and may play a more prominent role in regulation. The next possible reason could be the low frequency of the causal variants, which require a larger sample size to capture effect.

Despite the significant markers found for LSB, the present study has some limitations regarding LSW trait. It should be noted that LSB and LSW had extremely low heritabilities, suggesting a low possibility to achieve rapid genetic progress through phenotypic selection for LSB. Furthermore, LSW is generally connected to the mothering ability of the doe and environmental factors such as farm management. In addition, the sample size used in this study was limited, due to the low population size of Markhoz goats. Thus, caution must be taken when interpreting the results of the present GWAS, especially for LSW trait.

Conclusion

To conclude, we found plausible candidate genes based on SNPs associated with the EBV of LSB in the Markhoz goats using the 50K Caprine SNP-chip for the first time. The significant SNPs and genes identified in the present study can be beneficial for future molecular-based breeding for increased litter size at birth in goats; however, due to the low sample size used in this study, the results should be interpreted with caution. It is noteworthy that a breeding program focused on the major variations for LSB would not necessarily increase the number of surviving progenies due to the extremely low heritability of LSW. There may be no net impact on LSW from the slight increase in litter size caused by the substantial variations due to reduced viability.

Data availability statement

The SNP genotype data and EBVs for LSB and LSW are available in the Zenodo repository (https://zenodo.org/record/5824843).

Ethics statement

The animal study was reviewed and approved by Cornell University Institutional Animal Care and Use Committee (protocol #2014-0121). Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

PM performed data curation, formal analysis, visualization, methodology, and writing—original draft. AR and JR managed the project and contributed to writing—review and editing manuscript. AN-G conducted data curation, interpreted the results and contributed to review and editing manuscript. MR was responsible for data visualization. HH contributed to review and editing manuscript and provided SNP data and financial support. All authors contributed to the article and approved the submitted version.

Funding

Funding for the Markhoz goats' genotyping and part of the analysis has been supported by the laboratory of HH at Cornell University.

Acknowledgments

We would like to thank the Markhoz Goat Performance Testing Station Management for their cooperation by providing data used in this study.

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. Mahmoudi P, Rashidi A, Razmkabir M. Inbreeding effects on some reproductive traits in Markhoz goats. *Anim Prod Sci.* (2018) 58:2178–83. doi: 10.1071/AN17043

2. Bahmani HR, Tahmoorespur M, Aslaminejad AS, Vatankhah M, Rashidi A. Simulating past dynamics and assessing current status of Markhoz Goat population on its habitat. *Iran J Appl Anim Sci.* (2015) 5:347–54.

3. Ahlawat S, Sharma R, Maitra A, Tantia MS. Current status of molecular genetics research of goat fecundity. *Small Rumin Res.* (2015) 125:34–42. doi: 10.1016/j.smallrumres.2015.01.027

4. Feng T, Geng CX, Lang XZ, Chu MX, Cao GL, Di R, et al. Polymorphisms of caprine GDF9 gene and their association with litter size in Jining Grey goats. *Mol Biol Rep.* (2011) 38:5189–97. doi: 10.1007/s11033-010-0669-y

5. An XP, Hou JX, Zhao HB, Li G, Bai L, Peng JY, et al. Polymorphism identification in goat GNRH1 and GDF9 genes and their association analysis with litter size. *Anim Genet.* (2013) 44:234–8. doi: 10.1111/j.1365-2052.2012.02394.x

6. Maitra A, Sharma R, Ahlawat S, Tantia MS, Roy M, Prakash V, et al. Association analysis of polymorphisms in caprine KiSS1 gene with reproductive traits. *Anim Reprod Sci.* (2014) 151:71–7. doi: 10.1016/j.anireprosci.2014.09.013

7. An, X. P., Hou, J. X., Gao, T. Y., Lei, Y. N., Song, Y. X., Wang, J. G., et al. Cao, and B.Y. (2015). Association analysis between variants in KITLG gene and litter size in goats. *Gene*. 558, 126–130. doi: 10.1016/j.gene.2014.12.058

8. Hou JX, An XP, Han P, Peng JY, Cao BY. Two missense mutations in exon 9 of caprine PRLR gene were associated with litter size. *Anim Genet.* (2015) 46:87–90. doi: 10.1111/age.12223

9. Ghoreishi H, Fathi-Yosefabad S, Shayegh J, Barzegari A. Identification of mutations in BMP15 and GDF9 genes associated with prolificacy of Markhoz goats. *Arch Anim Breed.* (2019) 62:565–70. doi: 10.5194/aab-62-565-2019

10. Mahmoudi P, Rashidi A, Rostamzadeh J, Razmkabir M. Association between c.1189G>A single nucleotide polymorphism of GDF9 gene and litter size in goats: a meta-analysis. *Anim Reprod Sci.* (2019) 209, 106140. doi: 10.1016/j.anireprosci.2019.106140

11. Mahmoudi P, Rashidi A, Rostamzadeh J, Razmkabir M. A novel variant in the promoter region of miR-9 gene strongly affects litter size in Markhoz goats. *Theriogenology.* (2020) 158:50–7. doi: 10.1016/j.theriogenology.2020.09.008

12. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. *PLoS ONE.* (2014) 9:e86227. doi: 10.1371/journal.pone.0086227

13. Nazari-Ghadikolaei A, Mehrabani-Yeganeh H, Miarei-Aashtiani R, Staiger EA, Rashidi A, Huson HJ, et al. Genome-wide association studies identify candidate genes for coat color and mohair traits in the Iranian Markhoz goat. *Front Genet.* (2018) 9:105. doi: 10.3389/fgene.2018.00105

14. Rahmatalla SA, Arends D, Reissmann M, Wimmers K, Reyer H, Brockmann GA, et al. Genome-wide association study of body morphological traits in Sudanese goats. *Anim Genet.* (2018) 49:478–82. doi: 10.1111/age.12686

15. Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, Conington J, et al. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. *J Dairy Sci.* (2018) 101:2213–25. doi: 10.3168/jds.2017-12919

16. Silva FF, Bambou JC, Oliveira JA, Barbier C, Fleury J, Machado T, et al. Genome wide association study reveals new candidate genes for

resistance to nematodes in Creole goat. *Small Rumin Res.* (2018) 166:109–14. doi: 10.1016/j.smallrumres.2018.06.004

17. Gholizadeh M, Rahimi-Mianji G, Nejati-Javaremi A, Koning DJD, Jonas E. Genomewide association study to detect QTL for twinning rate in Baluchi sheep. *J Genet.* (2014) 93:489–93. doi: 10.1007/s12041-014-0372-1

18. Xu S, Gao L, Xie X, Ren Y, Shen Z, Wang F, et al. Genome-wide association analyses highlight the potential for different genetic mechanisms for litter size among sheep breeds. *Front Genet.* (2018) 9:118. doi: 10.3389/fgene.2018.00118

19. Hernández-Montiel W, Martínez-Núñez MA, Ramón-Ugalde JP, Román-Ponce SI, Calderón-Chagoya R, Zamora-Bustillos R, et al. Genome-wide association study reveals candidate genes for litter size traits in Pelibuey sheep. *Animals.* (2020) 10:434. doi: 10.3390/ani10030434

20. Islam R, Liu X, Gebreselassie G, Abied A, Ma Q, Ma Y, et al. Genome-wide association analysis reveals the genetic locus for high reproduction trait in Chinese Arbas Cashmere goat. *Genes Genom.* (2020) 42:893–9. doi: 10.1007/s13258-020-00937-5

21. Gilmour AR, Gogel BJ, Cullis BR, Thompson R. *ASReml User Guide Release 2, 0.* Hemel Hempstead: VSN International Ltd (2006).

22. Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protoc.* (2006). doi: 10.1101/pdb.prot4455

23. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigasci Giy.* (2019) 154. doi: 10.1093/gigascience/giy154

24. 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

25. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. (2016). doi: 10.1007/978-3-319-24277-4_9

26. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet.* (2019) 51:277–84. doi: 10.1038/s41588-018-0279-5

27. SAS Institute. *Users Guide, Version 8, 2. Statistics*. Cary, NC: SAS Institute (2001).

28. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* (2011) 88:76–82. doi: 10.1016/j.ajhg.2010.11.011

29. Bickhart D, Rosen B, Koren S. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet.* (2017) 49:643–50. doi: 10.1038/ng.3802

30. Rashidi A, Bishop SC, Matika O. Genetic parameter estimates for pre-weaning performance and reproduction traits in Markhoz goats. *Small Rumin Res.* (2011) 100:100–6. doi: 10.1016/j.smallrumres.2011.05.013

31. Lee C, de Silva AJ. Interaction of neuromuscular blocking effects of neomycin and polymyxin B. *Anesthesiology.* (1979) 50:218–20. doi: 10.1097/00000542-197903000-00010

32. Watanabe M, Fukuda A, Nabekura J. The role of GABA in the regulation of GnRH neurons. *Front Neurosci.* (2014) 8:387. doi: 10.3389/fnins.2014.00387

33. Giorgio DI, Bizzozzero-Hiriart NP, Libertun MC, Lux-Lantos V. Unraveling the connection between GABA and kisspeptin in the control of reproduction. *Reproduction*. (2019) 157, 1741–7899. doi: 10.1530/REP-18-0527

34. Luconi M, Cantini G, Baldi E, Forti G. Role of a-kinase anchoring proteins (AKAPs) in reproduction. *Front Biol.* (2011) 16:1315–30. doi: 10.2741/3791

35. Newhall KJ, Criniti AR, Cheah CS, Smith KC, Kafer KE, Burkart AD, et al. Dynamic anchoring of PKA is essential during oocyte maturation. *Curr Biol.* (2006) 16:321–7. doi: 10.1016/j.cub.2005.12.031

36. Young AI. Solving the missing heritability problem. *PLoS Genet.* (2019) 15:e1008222. doi: 10.1371/journal.pgen.1008222

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 December 2022
DOI 10.3389/fgene.2022.960234

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Ali Esmailizadeh,
Shahid Bahonar University of Kerman, Iran
Fredrick Kabi,
National Livestock Resources Research Institute, Uganda

*CORRESPONDENCE
Endashaw Terefe,
endashawt@arsiun.edu.et
Abdulfatai Tijjani,
abdulfatai.tijjani@gmail.com

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

RECEIVED 02 June 2022
ACCEPTED 22 November 2022
PUBLISHED 09 December 2022

CITATION
Terefe E, Belay G, Han J, Hanotte O and Tijjani A (2022), Genomic adaptation of Ethiopian indigenous cattle to high altitude.
Front. Genet. 13:960234.
doi: 10.3389/fgene.2022.960234

Genomic adaptation of Ethiopian indigenous cattle to high altitude

Endashaw Terefe[1,2,3]*, Gurja Belay[1], Jianlin Han[4,5], Olivier Hanotte[2,6,7] and Abdulfatai Tijjani[2,6]*

[1]Department of Microbial Cellular and Molecular Biology (MCMB), College of Natural and Computational Science, Addis Ababa University, Addis Ababa, Ethiopia, [2]International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia, [3]Department of Animal Science, College of Agriculture and Environmental Science, Arsi University, Asella, Ethiopia, [4]Livestock Genetics Program, International Livestock Research Institute (ILRI), Nairobi, Kenya, [5]CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China, [6]Centre for Tropical Livestock Genetics and Health (CTLGH), The Roslin Institute, The University of Edinburgh, Midlothian, United Kingdom, [7]School of Life Sciences, University of Nottingham, Nottingham, United Kingdom

The mountainous areas of Ethiopia represent one of the most extreme environmental challenges in Africa faced by humans and other inhabitants. Selection for high-altitude adaptation is expected to have imprinted the genomes of livestock living in these areas. Here we assess the genomic signatures of positive selection for high altitude adaptation in three cattle populations from the Ethiopian mountainous areas (Semien, Choke, and Bale mountains) compared to three Ethiopian lowland cattle populations (Afar, Ogaden, and Boran), using whole-genome resequencing and three genome scan approaches for signature of selection (iHS, XP-CLR, and PBS). We identified several candidate selection signature regions and several high-altitude adaptation genes. These include genes such as *ITPR2, MB,* and *ARNT* previously reported in the human population inhabiting the Ethiopian highlands. Furthermore, we present evidence of strong selection and high divergence between Ethiopian high- and low-altitude cattle populations at three new candidate genes (*CLCA2, SLC26A2,* and *CBFA2T3*), putatively linked to high-altitude adaptation in cattle. Our findings provide possible examples of convergent selection between cattle and humans as well as unique African cattle signature to the challenges of living in the Ethiopian mountainous regions.

KEYWORDS

high altitude, hypoxia, Ethiopian cattle, adaptation, candidate gene, convergent evolution

Introduction

Ethiopia is endowed with diverse agro-climatic regions and altitudes that range from the lowest Afar depression (−160 m above sea level, masl) to the highest Semien Mountain (4,600 masl). The Ethiopian highlands are commonly found in the central part of the country, on both sides of the Rift Valley, extending from the Semien Mountain in the North to the Bale Mountain in the Southeast. Cold temperatures and humid weather are characteristics of the high-altitude plateaus in Ethiopia. Mixed livestock farming and crop

cultivation are major agricultural activities for livelihoods. These environments are characterized by some unique agricultural production activities and food sources, providing cash income to the local communities (Asresie and Zemedu, 2015). Likewise, they have contributed to the diversity of Ethiopian livestock.

Human populations started to occupy the Ethiopian high plateau by migrating from the Rift Valley to the Bale Mountain in the Middle Stone Age around 50–30,000 years ago (Ossendorf et al., 2019). Overtime, people living in such environments have become adapted to high-altitude stressors, including hypobaric hypoxia, ultraviolet light, cold temperature, and oxidative stress (Beall et al., 2002; San et al., 2013; Debevec et al., 2017; Yang et al., 2017; Storz, 2021). The possible genetic basis of human adaptation to high altitudes in Ethiopia has been previously reported (Beall et al., 2002; Alkorta-Aranburu et al., 2012; Scheinfeldt et al., 2012; Huerta-Sánchez et al., 2013; Edea et al., 2019; Wiener et al., 2021). Likewise, evolutionary adaptations of livestock exposed to high altitudes are expected to be associated with major changes in the anatomy and physiological functions following a long period of acclimatization. For example, the yak *Bos grunniens* possesses a larger heart and lungs as compared to cattle (Wang et al., 2016), leading to a high amount of inhaled air to supply sufficient oxygen to the respiratory cell system. Uteroplacental oxygen flow at the fetal stage (Simonson, 2015), higher hemoglobin concentration in blood (Alkorta-Aranburu et al., 2012; Scheinfeldt et al., 2012), pulmonary vasoconstriction (Wang et al., 2016), ability to avoid altitude sickness (Dolt et al., 2007), calcium metabolism (Wang et al., 2015), and better foraging ability and energy metabolism (Qiu et al., 2012; Edea et al., 2014) may all contribute to the high-altitude adaptation in cattle and other livestock species.

High-altitude adaptation in animals relies on their genetic background attained through natural selection. Hypoxia induced factors such as HIF-1a and its paralogs of HIF-2a and HIF-3a are oxygen regulating factors in a hypobaric hypoxia environment and are thus considered candidate genes for high-altitude adaptation. The HIF-1a gene is over-expressed in cattle, yak, humans, and the Tibetan gray wolf living at high altitudes (Newman et al., 2011; Bigham and Lee, 2014; Zhang et al., 2014; Wang et al., 2016; Verma et al., 2018; Werhahn et al., 2018). The HIF-1a pathways include the endothelial PAS domain 1 (*EPAS1*), vascular endothelial growth factor-A (*VEGF-A*), endothelial converting enzyme-1 (*ECE1*), glucose transport members 1 (*GLUT-1*), hexokinase 2 (*HK2*), and nitric oxide synthesis (*NOS2*) genes. These are all expressed in cattle adapted to high altitudes (Verma et al., 2018), and they play an important role in maintaining oxygen homeostasis and glucose metabolism in mammals (Hu et al., 2006; Majmundar et al., 2010). Hypoxia-related genes, including *EPAS1*, *RYR2*, and *ANGPT1*, were identified in high-altitude Tibetan gray wolves, and they were associated with the HIF signaling pathway, ATP binding, and response to oxygen-containing compounds (Zhang et al., 2014). Using the Illumina bovine low-density 50K SNP array, a study on the Ethiopian cattle population living at an altitude of 2,400 masl identified energy metabolism (*ATP2A3*, *CA2*, *MYO18B*, *SIK3*, *INPP4A*, and *IREB2*) and response to hypoxia (*BDNF*, *TFRC*, and *PML*) genes as candidate genes to the adaptation of cattle to the high-altitude environments (Edea et al., 2014).

Physiological and genomic landscape studies have revealed convergent evolution in several species to independently adapt to high altitudes in different geographic locations across the world (Scheinfeldt et al., 2012; Huerta-Sánchez et al., 2013; Azad et al., 2017; Witt and Huerta-Sánchez, 2019; Friedrich and Wiener, 2020). For example, human populations in the Tibetan, Andean, and Ethiopian highlands shared common candidates selected regions and genes linked to high-altitude adaptation, such as *PPARA* and *EDNRA* (Scheinfeldt et al., 2012; Simonson, 2015; Witt and Huerta-Sánchez, 2019). However, *ARNT2* and *THRB* were uniquely identified in the Ethiopian population (Scheinfeldt et al., 2012).

Only a few studies have reported the environmental adaptations of Ethiopian cattle at the full autosomal genome level (Kim et al., 2017; Kim et al., 2020). Though the bovine low-density SNP array (Edea et al., 2014) was the first to investigate Ethiopian cattle adaptation to different environments including hypoxia. However, it did not include high altitude adaptation of cattle population living at an altitude of >3,000 masl. Therefore, this study aimed to identify signatures of positive selection for high altitude adaptation in Ethiopian cattle using whole-genome resequencing. For this purpose, we selected three cattle populations from the highest mountainous areas (>3,000 masl) of the country (Bale, Choke, and Semien Mountain areas) (Table 1), and three Ethiopian cattle populations living at low altitudes.

Materials and methods

Cattle populations and whole-genome resequencing

The study involved a comparative analysis of indigenous cattle distributed at high altitudes (>3,000 masl) and low altitudes (<1,500 masl) in Ethiopia (Table 1). The high-altitude populations included Bale (n = 10) sampled in the Bale district (Bale Mountain, ~3,586 masl), Semien (n = 10) in the Gondar district (Semien Mountain, ~3,732 masl), and Choke (n = 10) in the Gojam district (Choke Mountain, ~3,410 masl). The low-land populations included Afar (n = 11) sampled in the Afar district (~729 masl), Ethiopian Boran (n = 10) from the Borana district (~1,368 masl), and Ogaden (n = 9) from the Ogaden district (~1,200 masl) (Figure 1).

Whole blood samples were collected aseptically from the jugular veins of unrelated individuals into 10 ml vacuum tubes containing EDTA. Genomic DNA was extracted using the

TABLE 1 Sampling location and cattle population description in the high and low altitudes.

Category	Breed	Region (district)	Location	Altitude (masl)	GPS		Climate
					Latitude (degree)	Longitude (degree)	
High altitude	Bale	Oromia (Bale)	Bale Mountain	3,586	6.77	39.75	Cold humid, highland
	Choke	Amhara (East Gojam)	Choke Mountain	3,410	10.60	37.84	Cold humid, highland
	Semien	Amhara (North Goder)	Semien Mountain	3,732	13.23	38.13	Cold humid, highland
Low altitude	Afar	Afar	Melka Were/ Asayta	729	9.34	40.17	Hot and dry lowland
	Boran	Oromia (Borena)	Dubulk	1,368	4.55	38.10	Hot and dry lowland
	Ogaden	Ogaden	Jigjiga	1,200	09.58	41.85	Hot and dry lowland

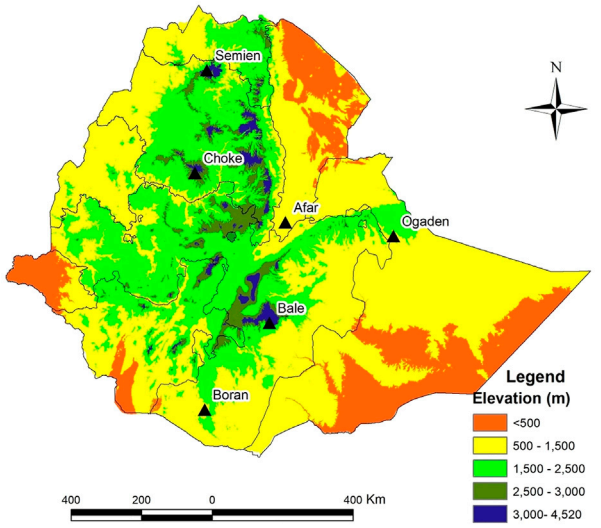


FIGURE 1
The elevation map of Ethiopia and sampling locations for high and low-altitude cattle populations.

QIAGEN DNeasy Blood and Tissue Kit (https://www.qiagen.com/us/) following the manufacturer's protocol. The DNA integrity was checked by a 1% agarose gel electrophoresis and observed under a UV light-based gel viewer. The concentration and quality of DNA for each sample were checked by using a DeNovix DS-11 FX Series Spectrophotometer/Fluorometer (DeNovix Inc., Wilmington, DE, United States). DNA samples (>50 μg/μl) were then shipped to Novogene, China (https://en.novogene.com/services/research-services/genome-sequencing/whole-genome-sequencing/animal-plant-whole-genome-sequencing-wgs/), where whole-genome sequencing was performed on an Illumina NovaSeq 6000 Platform (Illumina, San Diego, CA, United States) to generate 150 bp of paired-end reads. We included Gir (GenBank accession no. PRJNA343262), Angus (PRJNA318087), Muturu (PRJNA386202), and Butana (PRJNA574857) cattle for comparative analyses, following the same sequence quality control and variant calling procedures.

Read mapping and variant calling

The quality control of raw sequencing reads was performed using the *FastQC v0.11.9* program (https://github.com/s-andrews/FastQC/releases/tag/v0.11.9). Qualified raw reads were processed for initial trimming and filtering of the low-quality reads by removing adapters, short reads (sequence length <35 bp), and reads with low sequence base quality (quality score <20) using the *Trimmomatic v0.38* tool (Bolger et al., 2014). Clean reads were mapped to the latest taurine cattle reference genome of ARS-UCD1.2 (Shamimuzzaman et al., 2020) using the *BWA-MEM* algorithm of Burrows-Wheeler Aligner (*bwa v0.7.17*) (Li and Durbin, 2010).

Mapped reads were sorted and indexed using the *samtools v1.8* (Li et al., 2009) to produce a BAM file. Alignment sorting by coordinate and marking of potential PCR and optical duplicates were carried out using the *MarkDuplicates* tool in *Picard v2.18.2* package (http://picard.sourceforge.net). Base quality score recalibration (BQSR) and haplotype caller analysis were performed using the *GATK v3.8-1-0-gf15c1c3ef* according to its best practice workflows (McKenna et al., 2010). The known variants of ARS1.2PlusY_BQSR_v3.vcf.gz provided by the 1,000 Bull Genomes project were used for masking known sites for all cattle samples. The *GATK PrintReads* was run to adjust the base quality scores in the data based on information from the table and to produce a recalibrated bam file.

Then, the genomic variant call format (*gVCF*) file for each sample was created using the *GATK HaplotypeCaller* command

from the recalibrated bam file, and all samples were combined to obtain a joint genotype file using the *GATK CombineGVCFs*. Finally, the variants were processed for variant recalibration with a 99.9 truth sensitivity filter level using the *GATK* to reduce false discovery rates that minimize the noise created by low standard variants. After all quality checking, approximately 36.6 million biallelic autosomal SNPs were identified and used for downstream analyses.

Population genetic structure

Principal component analysis (PCA) and admixture modeling were performed, based on the SNP genotypes, to determine the population genetic structure of indigenous Ethiopian cattle living in high and low altitudes. We used Angus, Gir, Muturu, and Butana cattle as reference breeds (European taurine, Asian zebu, African taurine, and non-Ethiopian zebu cattle). The dataset in the vcf file was first converted to a plink format (map, ped, and fam) after pruning the SNPs in linkage disequilibrium (LD) ($r^2 \geq 0.5$), minor allele frequency (MAF) (< 0.05), and missing genotype (call rate > 10%) based on a step-wise procedure using 50 SNPs windows and 10 SNPs steps. After the stringent variant quality check, 5.1 million autosomal SNPs with an average of 98.1% genotyping rate were used for admixture modeling with the *Admixture v1.3.0* software (Alexander et al., 2009) to estimate the ancestry proportion of individual samples. The ancestral proportions in the hierarchical clustering of individual samples were optimized at the *K* values ranging from 2 to 10 and the admixture plot was visualized using the R package. For PCA, we removed SNPs with MAF < 0.01, SNPs with missing genotypes > 10%, and SNP calling rate < 90%. After this filtering, 25.5 million SNPs were used for PCA. The eigenvectors of each sample were calculated using *PLINK 1.9* (Purcell et al., 2007) and the result was plotted with the ggplot2 in the R package.

Integrated haplotype homozygosity analysis

The phasing and imputation of missed genotypes were estimated per chromosome for each population using the *Beagle v5.1* software (Browning and Browning, 2007). The length of homozygous haplotypes along a chromosome was used to estimate the LD decay. The extended haplotype homozygosity (EHH) from each SNP, which is the probability that two randomly chosen homologous chromosomes carrying the core haplotype of interest are identical by descent (Sabeti et al., 2002), was then calculated. The integrated haplotype score (iHS) compares the integrated EHH between the ancestral allele relative to the derived allele in a population (Voight et al., 2006). It detects selective sweeps when alleles are near fixation. The iHS analysis was done using the *REHH v2.0* R package (Gautier et al., 2017) for alleles with MAF within a population > 0.05. A genomic window of 100 kb and a step size of 50 kb were used to identify candidate regions of selection signatures.

Cross-population composite likelihood ratio (XP-CLR)

XP-CLR test was done between cattle living at high and low altitudes using the haplotype phased data of each chromosomal window of 100 kb and a step size of 50 kb. The test is based on local allele frequency changes in a genomic region between the two groups. The method is most sensitive to recent selection and detects departures from neutrality that could be compatible with hard or soft selection sweeps (Chen et al., 2010).

Population branch statistic (PBS)

PBS analysis was employed to detect genomic regions under selection with highly divergent haplotypes (Yi et al., 2010). We run population differentiation (F_{ST}) analysis using the *vcftools v0.1.15* (Danecek et al., 2011) between the Semien population for the highest altitude representative compared to Afar, Boran, and Ogaden cattle for the low altitude one, and the Sudanese Butana cattle as an outgroup. The top 0.5% of the candidate regions detected by the iHS and XP-CLR tests were used for PBS analysis. We estimated divergence time using a log-transformation of one minus the F_{ST} value for each comparison and calculated the PBS value using the method described in Huerta-Sánchez et al. (2013).

Functional annotation and haplotype structure of candidate genes

The candidate genomic regions were annotated using the taurine cattle reference genome ARS-UCD1.2 (Shamimuzzaman et al., 2020) in the Ensembl database (http://www.ensembl.org/index.html). The Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6.8, https://david.ncifcrf.gov/home.jsp) was used to understand the biological functions and molecular pathways of the candidate genes (Huang et al., 2009), according to the minimum similarity thresholds for enrichment scores at 1.0 and p values ≤ 0.05. Further functional annotations were done from the literature published for humans and other vertebrates. Additional structural and functional analyses of the genomic regions of the candidate genes were evaluated using haplotype structure, LD, F_{ST}, nucleotide diversity, and STRING

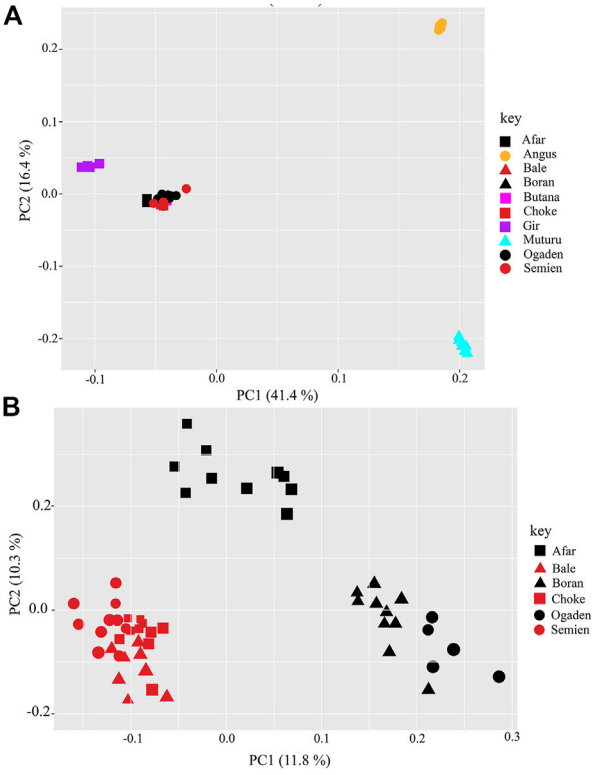


FIGURE 2
A plot of the first and second principal components (PC1 and PC2) of **(A)** the whole population dataset and **(B)** Ethiopian high altitude (brown color: Bale, Choke, and Semien) and low altitude (black color: Afar, Boran, and Ogaden) cattle populations.

protein-protein interaction network database. The haplotype structure was estimated based on pairwise LD heatmap of the SNPs using the *LDBlockShow* (Dong et al., 2020). The F_{ST} and nucleotide diversity of the candidate gene regions were estimated using the *vcftools v 0.1.15* (Danecek et al., 2011) in a 10 kb window and 5 kb step size to determine the strength and pattern of selection signatures between the high- and low-altitude cattle. Protein-protein interaction analyses were done using the STRING online platform for the cattle genome reference database (https://string-db.org/).

Results

Population genetic structure

The PCA and admixture plots describe the population genetic structure of the indigenous Ethiopian cattle living in the high and low altitudes compared with European taurine cattle (Angus), African taurine (Muturu), Asian zebu cattle (Gir), and African zebu (Butana) (Figure 2A). The first and second principal components (PC) represent 57.8% of the total variation. The first PC (PC1, 41.4%) separates the zebu (Gir, Butana, and Ethiopian cattle) from the taurine (Angus and Muturu), while the second PC (PC2, 16.4%) differentiate the African zebu (Ethiopian cattle Butana) and Muturu from all non-African cattle (Gir and Angus) (Figure 2A). Next, a second PCA was conducted, excluding the reference cattle. The PC1 (11.8%) and PC2 (10.3%) differentiate the Ethiopian high-altitude (HA) from low-altitude (LA) cattle (Figure 2B). The population genetic admixture plot supports the PCA result, which separated cattle populations in the whole dataset into four ancestry clusters (Figure 3A). At K = 4, the genetic ancestry of the Ethiopian cattle was inferred to be 97.0% of African zebu, 1.6% Asian zebu (represented here by the Gir), 1.0% African taurine (Muturu), and 0.4% European taurine (Angus) (Figure 3B). A small shared European taurine component is observed in Bale, Choke, Semien, and Boran cattle. It is, however, higher (2.2%) in Ogaden cattle. The Afar and Boran populations have very little African taurine ancestry proportion. Butana cattle share a similar genetic background to other Ethiopian cattle. To explore potential genome-wide selection signatures for high-altitude adaptation, we analyzed the HA and LA cattle populations separately, following the Ethiopian cattle PCA results (Figure 2B).

Selection signatures within Ethiopian high- and low-altitude cattle populations

We performed a genomic scan combining the three HA, as well as combining the three LA populations, using the within-population iHS test to identify recent and/or ongoing footprints of natural selection (Vatsiou et al., 2016). Using the *REHH v2.0* R package, we calculated genome-wide iHS for each focal SNP from the phased data (Gautier et al., 2017). Subsequently, we summarized the selection statistics across a sliding 100 kb genomic window with a 50 kb step size. From the empirical distribution of iHS statistics, we applied a *p*-value threshold < 1.0E-6, equivalent to -log10 iHS ≥ 6, to select the candidate regions under selection (Figure 4A, Supplementary Table S1).

There are 144 candidate selected regions across 29 autosomes within the Ethiopian zebu populations living in high altitudes (Supplementary Table S1). These regions vary in size from 150 to 750 kb. They overlap with 264 protein-coding genes based on the Ensembl taurine cattle assembly (ARS-UCD1.2) (Supplementary Table S1). Of these, 28 protein-coding genes were identified within the top 10 *iHS* regions. Most of these genes remain uncharacterized with the exception of *ITPR2* on BTA5 (5:83.45–83.60 Mb, iHS –log10 p = 8.19), *DUSP10* on BTA16 (16:25.05–25.20 Mb, iHS –log10 p = 8.24), and GTPase IMAP family members 4–7 genes (*GIMAP4–7*) on BTA4 (4:112.95–113.35 Mb, *iHS* –log10 p = 8.60). These annotated genes are possibly involved in high-altitude adaptation, especially the former two genes with

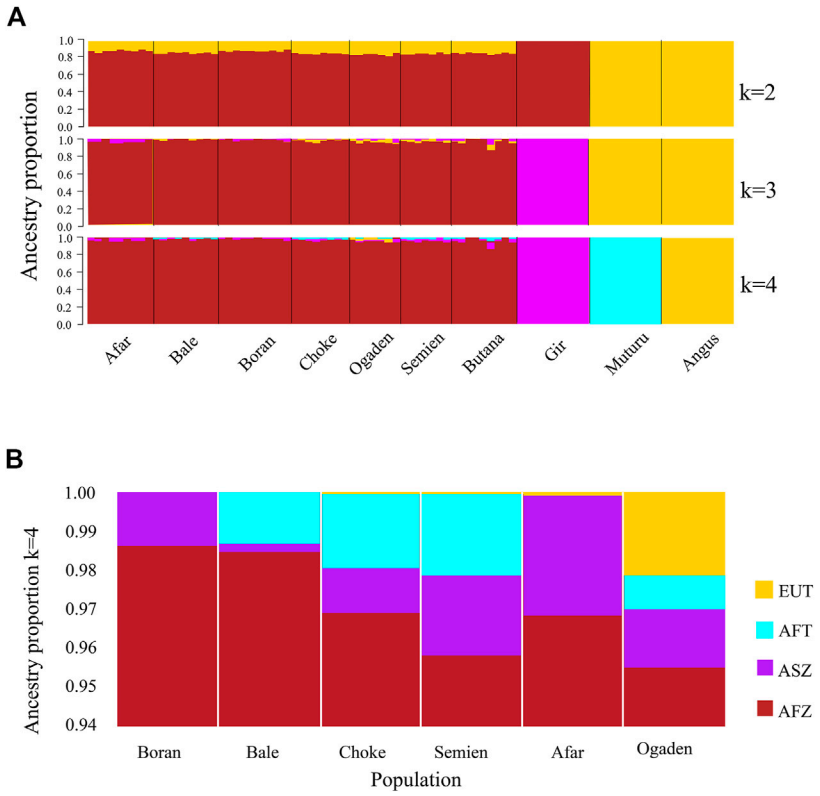


FIGURE 3
The plot of population genetic admixture analysis. **(A)** Cattle population in the whole dataset and **(B)** Ethiopian cattle ancestry proportion.

functions linked to the response to hypoxia (*ITPR2*) and oxygen-containing compounds (*DUSP10*), respectively. *GIMAP4*, *GIMAP5*, and *GIMAP7* functions are related to the immune response and hematopoiesis (Chen et al., 2011; Schwefel et al., 2013). They play a significant role in modulating immune functions by controlling cell death and the activation of T cells (Ho and Tsai, 2017).

To contrast the selection signatures of Ethiopian cattle living in high altitudes with the ones living at low altitudes, we performed an additional genomic scan based on the iHS in the three LA cattle populations at the same threshold (p-value < 1.0E-6). We identified only 20 candidate regions under selection (Figure 4B, Supplementary Table S2). These regions vary in size from 150 kb to 650 kb. They overlap with 50 protein-coding genes. Twenty-three (~45%) were common with those identified in the HA populations. The common genes include 14 genes, mostly uncharacterized, present in the top 10 iHS regions, except for the three GIMAP family members, GIMAP 4, 5, and 7 genes. Among the remaining nine common genes, *VEGFC* and *EP300* are possibly linked to the adaptation to high altitudes due to their functions in the vascular system (Herbert and Stainier, 2012; Huerta-Sánchez et al., 2013; Bigham and Lee, 2014; Azad et al., 2017; Zheng et al., 2017). However, due to the fewer candidate regions identified in the LA populations, we decided to increase the iHS threshold to p value < 1.0E-7, equivalent to -log10 iHS ≥ 5, which added 113 protein-coding genes, of which 63 were common in both HA and LA populations (Figure 4B, Supplementary Table S2).

Comparative genomic signatures between Ethiopian high- and low-altitude cattle populations

We further investigate the genomic footprints of natural selection in indigenous Ethiopian cattle by contrasting the allele frequency profiles between the HA and LA populations using the XP-CLR test. The top 0.5% XP-CLR scores (XP-CLR > 10) include 216 candidate windows of 100 kb size regions, from which 251 protein-coding genes were annotated (Supplementary Table S3). Unlike many uncharacterized genes within the iHS regions, the top 10 signals identified by the XP-CLR test include seven annotated genes (*MSRB3, LEMD3,* and *WIF1* on BTA5, *SLC26A2, HMGXB3,* and *CSF1R* on BTA7, and *RXFP2* on BTA12) (Figure 5). These are not found within the top high altitude iHS signals. Some have functions that may be related to high-altitude adaptation. For instance, *MSRB3* is involved in cold tolerance in Arabidopsis and high-altitude adaptation in Tibetan

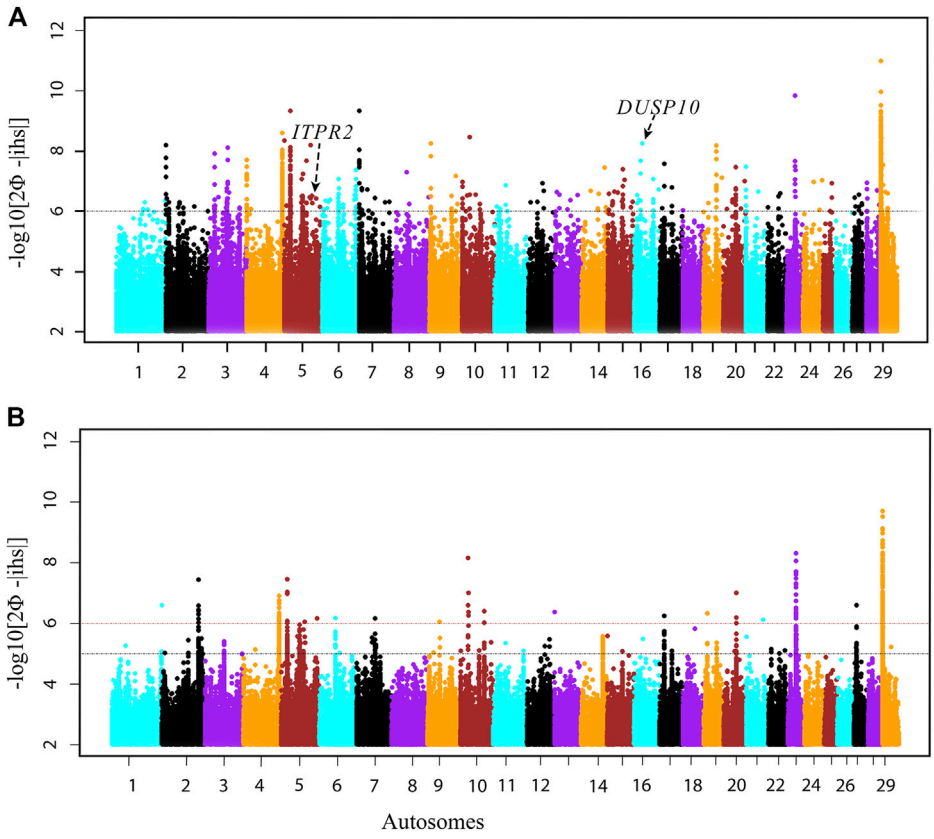


FIGURE 4
Manhattan plots of genome-wide scans based on the iHS test. **(A)** HA, high-altitude and **(B)** LA, low-altitude Ethiopian cattle populations.

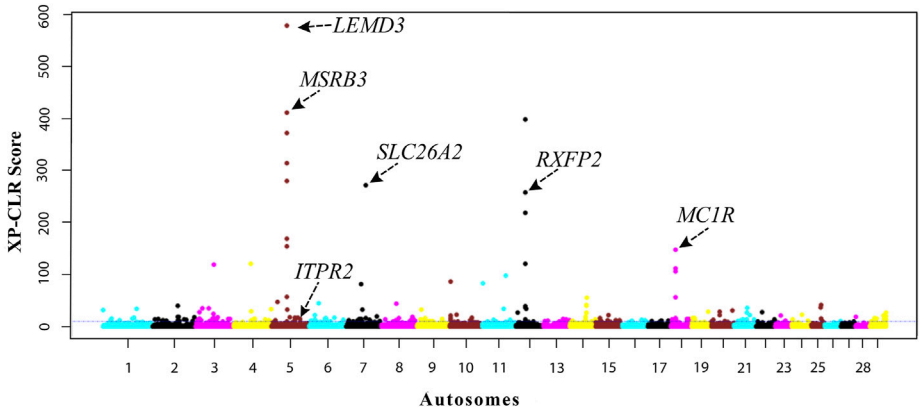


FIGURE 5
Manhattan plot of genome-wide XP-CLR scores by contrasting the high- and low-altitude cattle populations.

dogs and sheep (Kwon et al., 2007; Vaysse et al., 2011; Wei et al., 2016; Witt and Huerta-Sánchez, 2019). This gene was also reported to protect cells from oxidative stress caused by hypoxia (Hansel et al., 2005) as well as from cold and heat stress (Lim et al., 2012). *RXFP2* was reported to control horn type, development, and morphology (Pan et al., 2018; Ahbara

TABLE 2 List of overlapping regions and candidate genes identified using the iHS and XPCLR selection scan methods including, gene functions in reported species.

BAT	Region start	Region end	XPCLR score	iHS value	Gene name	Gene function/phenotype	Species	References
2	42.25	42.45	10.99	6.28	GALNT13			
3	17.55	17.6	11.54	6.15	ENSBTAG0 0000050002	Novel gene/uncharacterized protein	—	—
3	57.5	57.6	16.01	6.75	CLCA4	Rennin secretion, ion channel activity		Palubiski et al. (2020); Weir and Olschewski, (2006)
					CLCA1	"		"
3	57.55	57.65	12.65	6.75	CLCA2	"		"
5	83.45	83.55	16.65	8.19	ITPR2	Response to hypoxia	Human	Huerta-Sánchez et al. (2013); Jurkovicova et al. (2008); Manalo et al. (2005); Qu et al. (2015)
8	48.4	48.5	43.83	6.23	GDA			
12	29.6	29.7	11.17	6.31	B3GLCT	Carbohydrate metabolic process, protein glycosylation, horn development, environment adaptation	Cattle, sheep	Ahbara et al. (2019); Flori et al. (2019); Pan et al. (2018)
14	81.45	81.55	13.63	6.04	COL14A1	Protein binding, angiogenesis	Mice, rats, sheep, human	Chai et al. (2004); Copple et al. (2011); Wiener et al. (2021); Zhang et al. (2018)
16	60.65	60.75	12.94	7.05	SOAT1	Cholesterol metabolic process	Insects, mice	Guan et al. (2020); Miron and Tirosh, (2019); Zuniga-Hertz and Patel, (2019)
					AXDND1	Response to bone fracture/bone synthesis	human	Pettersson-Kymmer et al. (2013)
28	14.5	15.5	18.81	6.94	NUP133	Regulate mitochondrial function and oxidative stress response	Mouse	Sunny et al. (2020)
					ABCB10	ATPase-coupled transmembrane transporter activity; regulates heme synthesis; Iron metabolism; reactive oxygen species	Mouse, zebrafish, human cell culture	Bayeva et al. (2013); Liesa et al. (2012); Seguin and Ward, (2018); Valverde et al. (2015); Yamamoto et al. (2014)
					ENSBTAG0 0000034225	Basal transcription, coactivators, and promotor recognition.		

et al., 2019; Liu et al., 2020) and linked to high-altitude adaptation to hypoxia in sheep (Guo et al., 2021).

There are 14 genes in common to both XP-CLR and iHS (HA) tests, (*CLCA4*, *CLCA1*, *CLCA2*, *ITPR2*, *ABCB10*, *NUP133*, *GDA*, *GALNT13*, *ENSBTAG00000050002*, *COL14A1*, *AXDND1*, *B3GLCT*, *SOAT1*, and *ENSBTAG00000034225*) (Table 2; Figure 7). These genes could be regarded as promising candidates subjected to natural selection for high-altitude adaptation in Ethiopian HA cattle. On the other hand, no shared candidate gene was found between the LA iHS cattle and the XP-CLR test.

Functional annotation of genes under selection in Ethiopian high-altitude cattle populations

We conducted a functional annotation using the DAVID visualization tools, based on the Ensembl taurine cattle assembly (ARS-UCD1.2, to identify GO terms and KEGG pathways for the candidate genes that we detected in the Ethiopian HA cattle populations following iHS and XP-CLR analyses. Genes with fold enrichment >1.2 and p-value ≤ 0.05 were considered to be significant (Table 3). Several top candidate genes related to environmental stress such as hypobaric hypoxia, temperature, and UV radiation were clustered into important GO terms, including response to hypoxia (GO:0001666; p-value: 9.3E-06), response to oxygen-containing compound (GO:1901700; p-value: 1.0E-04), ion channel activity (GO: 0005216; p-value: 4.8E-06), glucose homeostasis (GO:0042593; p-value: 5.1E-03), and ATPase activity (GO:0016887; p-value: 1.3E-03), which are biological processes potentially relevant to high altitude adaptation. These findings are in line with previous reports on cattle and other species adapted to high altitudes (Remillard and Yuan, 2006; Edwards et al., 2007; Shimoda and Polak, 2011; Ge et al., 2013; Veith et al., 2016; Moore, 2017; Hu et al., 2019; Friedrich and Wiener, 2020).

TABLE 3 Gene ontology (GO) clustering and enrichment analyses of candidate genes identified by genome-wide iHS and XP-CLR scans in the high-altitude cattle populations.

GO term	Count	*p* value	Fold change	Gene
GO: 0005216~ion channel activity	15	4.8E-06	4.5	*ITPR2, KCNJ8, GPR89A, TRPM3, CLCA1, NOX5, GRIK3, CACNG2, SLC26A, CLCA4, GABRB2, ABCC9, KCNQ3, CLCA2, TPC3*
GO: 0001666~response to hypoxia	7	9.3E-04	6.1	*ITPR2, MB, CBFA2T3, SRF, VEGFC, HMOX1, ARNT*
GO: 0034101~erythrocyte homeostasis	6	1.0E-03	7.7	*MAFB, MB, FOXO3, SRF, HMOX1, ARID4A*
bta04924: Renin secretion	6	2.9E-04	9.9	*ITPR2, CLCA1, PRKACB, CLCA4, CLCA2, GUCY1A2*
GO: 1901700~response to oxygen-containing compound	19	1.0E-04	2.8	*DUSP10, ITPR2, KCNJ8, IMPACT, NDUFS4, SESN3, AVPR1A, STAT1, PDX1, PTK7, SLC11A1, ADH5, EFNA5, SSTR2, NOX4, HRH4, CRY2, ENPP1, MZB1*
GO: 0008217~regulation of blood pressure	6	2.1E-03	6.6	*AVPR1A, POMC, GRIP2, ERAP1, ADH5, MYH6*
GO: 0020037~heme binding	5	2.1E-02	4.7	*MB, HMOX1, CYP4F2, GUCY1A2, ENSBTAG00000048257*
GO: 0016887~ATPase activity	8	1.3E-02	3.1	*ABCB10, ATP6V0A1, YTHDC2, MYO10, ABCC9, ATP6V1G1, MYH7, MYH6*
GO: 0042593~glucose homeostasis	6	5.1E-03	5.3	*SESN3, POMC, PDX1, FOXO3, CRY2, NOX4*

Low atmospheric oxygen concentration in the inhaled air causes low oxygen levels in the arterial blood reducing cellular energy production, which leads to cellular stress and then induces several factors to increase oxygen availability to cell mitochondria for energy homeostasis. Physiological homeostasis is established through increasing tissue oxygen supply by mounting vascular smooth muscle tone to withstand fast blood flow pressure by inducing the formation of additional blood vessels (angiogenesis), increasing the number of erythrocytes, and improving heme-binding affinity. Supporting these adaptive mechanisms, we identified candidate genes in the biological processes of erythrocyte homeostasis (GO:0034101; *p*-value: 1.0E-03), heme-binding (GO:0020037; *p*-value: 2.1E-02), and the regulation of blood pressure (GO:0008217; *p*-value: 2.1E-03) enhancement. The increases in erythrocyte, hemoglobin concentration, and heme-binding affinity enable more oxygen transportation to tissues in hypoxia-adapted animals (Storz, 2007; Zhang et al., 2007; Storz and Moriyama, 2008; Storz et al., 2010; Yalcin and Cabrales, 2012; Storz, 2016).

Identification of candidate genes associated with high-altitude adaptation

We further analyzed the candidate genes clustered into biological processes relevant to high altitude adaptation (Table 3) using the population branch statistics (PBS). We compared Semien cattle from the highest Ethiopian mountain area to each of the LA cattle populations (Afar, Boran, and Ogaden) using Butana cattle from the Sudanese arid region as an outgroup (see *Materials and methods*). The 10 kb window outliers from the PBS analysis represent the most differentiated genomic regions (PBS value ≥ 0.2). They overlap with *SLC26A2, CLCA1, CLCA2, KCNJ8, GUCY1A2,* and *CBFA2T3* (Figure 6, Supplementary Table S4). These genes have possible roles in ion channel activity, renin secretion, response to hypoxia, response to oxygen-containing compounds, and heme-binding (Table 3). The genomic region within the *SLC26A2* gene was the most differentiated in the PBS scans (Figure 6). *SLC26A2* is a ubiquitously expressed $SO4^{2-}$ transporter with high expression levels in cartilage and several epithelia (Ohana et al., 2012; Park et al., 2014). This gene is involved in body size and male fertility in humans (Kujala et al., 2007; Touré, 2019), and its mutations have been implicated in dwarfism (Yang and Liang, 2021) and dysplasia (Pineda et al., 2013; Zheng et al., 2019; Heidari et al., 2021).

In addition, *CLCA2* and two other paralogs, *CLCA1* and *CLCA4*, and *ITPR2* were the only four candidate genes detected by the three genomic scans (Figure 7, Supplementary Table S5). Moreover, the variants within the *CLCA2* in the HA populations showed a higher level of linkage disequilibrium (LD) compared to the LA populations (Figure 8). Similarly, the nucleotide diversity and population differentiation plot show the *CLCA2* gene region with significant variation compared to regions of the two paralog genes (Figure 9). Therefore, we considered *CLCA2, ITPR2*, *SCL26A2,* and *CBFA2T3* as strong candidate genes putatively linked to high-altitude adaptation in Ethiopian cattle.

Other genes of interest include GO terms linked to the response to hypoxia (Table 3). These include *MB* (BTA5: 73.81–73.82 Mb), *CBFA2T3* (BTA18: 14.05–14.1 Mb), and *SRF* (BTA23: 16.77–16.78 Mb) from XP-CLR scans results, *ARNT* (BTA3: 19.8–19.9 Mb) and *VEGFC* (BTA27: 8.0–8.1 Mb) from iHS scans results, and *ITPR2* from both XP-CLR and iHS scan results. *ARNT* is involved in the positive regulation of vascular endothelial growth factor (*VEGF*) activation. *VEGFC*, a VEGF homolog, is involved in regulating endothelial cell proliferation and angiogenesis in response to the low oxygen concentration in the arterial

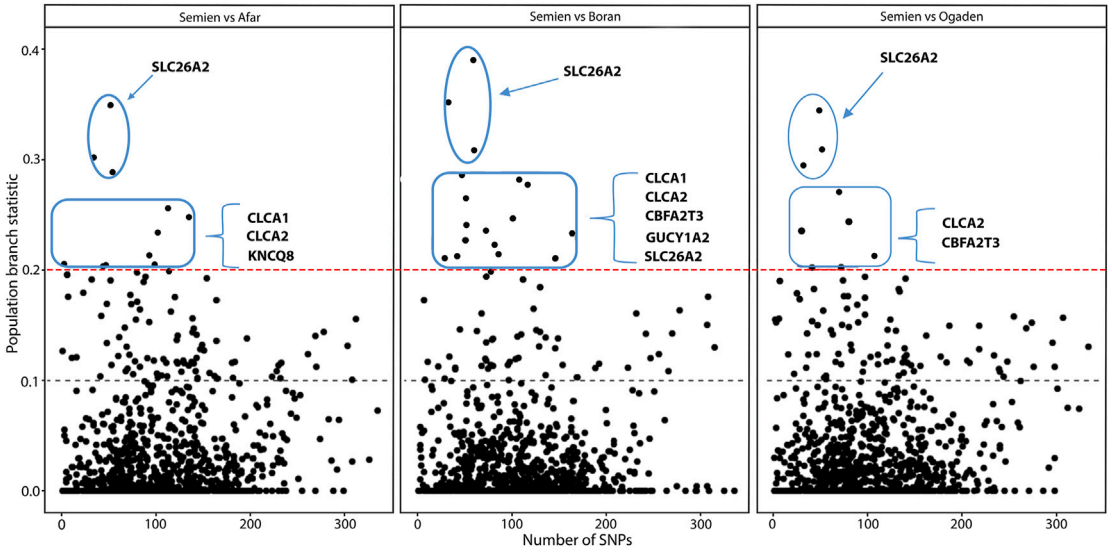


FIGURE 6
The distribution of the population branch statistic (PBS) values in 10 kb genomic regions as a function of the number of SNPs.

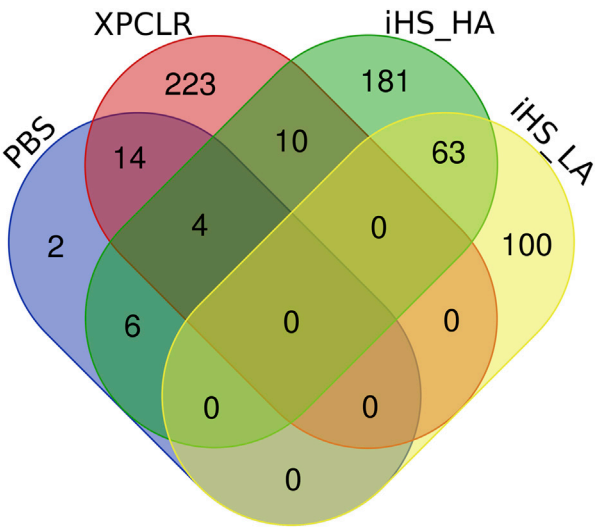


FIGURE 7
Candidate genes supported by the iHS, XP-CLR, and PBS analyse in high-altitude (HA) and low-altitude (LA) cattle populations.

blood (Kumar et al., 2011; Herbert and Stainier, 2012; Ramakrishnan et al., 2014).

ARNT also enhances endothelial cell growth in the vascular line and it is expressed during the early phase of the growth of new blood vessels (angiogenesis) (Scheinfeldt et al., 2012; Geng et al., 2014; Graham and Presnell, 2017). Protein-protein interaction network analysis shows that ARNT interacts with hypoxia-inducible factors such as HIF1a, EPAS1, EP300, and its paralog CREBBP (Figure 10). *CBFA2T3* is clustered in response to hypoxia and functions as a transcription regulator of *HIF1a* through interaction with *EGLN1* and promoting the *HIF1a* prolyl hydroxylation-dependent ubiquitination and proteasomal degradation pathways (Kumar et al., 2015). It also contributes to the inhibition of glycolysis and the stimulation of mitochondrial respiration by down-regulating the expression of glycolytic genes as direct targets of *HIF1a* (Kumar et al., 2013).

Discussion

This study aimed to unravel at the autosomal genome level the adaptation of Ethiopian indigenous cattle to the extreme environmental conditions of its mountainous areas. We studied specifically three cattle populations of Semien, Bale, and Choke living in a mountainous area of more than 3,000 masl by contrasting them with the indigenous cattle population from the Ethiopian lowlands. Population genetic structure validated the African zebu admixture of indicine and taurine status of all the studied indigenous Ethiopian cattle (Figure 1), while the PCA result shows some level of genetic differentiation between the high-altitude Ethiopian cattle populations from those originating from the low altitude locations (Figure 2B). We then applied the iHS, XP-CLR, and PBS methods to detect selection signatures within and between the Ethiopian cattle populations living at

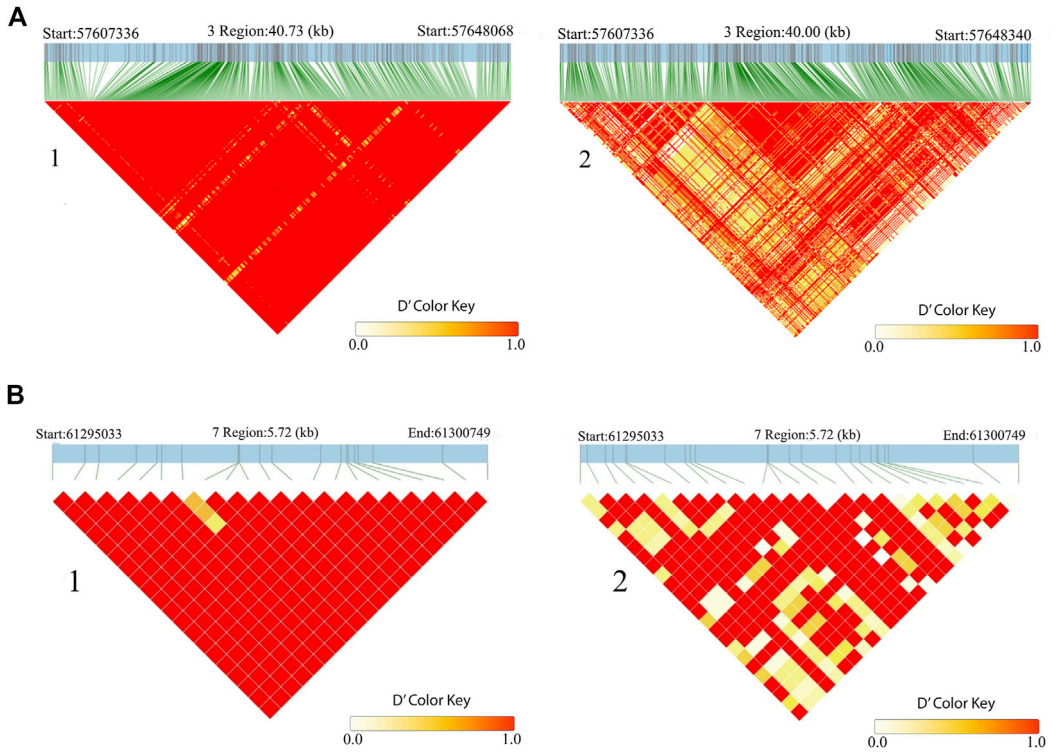


FIGURE 8
LD block heatmap of the candidate genes of *CLCA2* **(A)** and *SLC26A2* **(B)** in the high-altitude (1) and low-altitude (2) cattle populations.

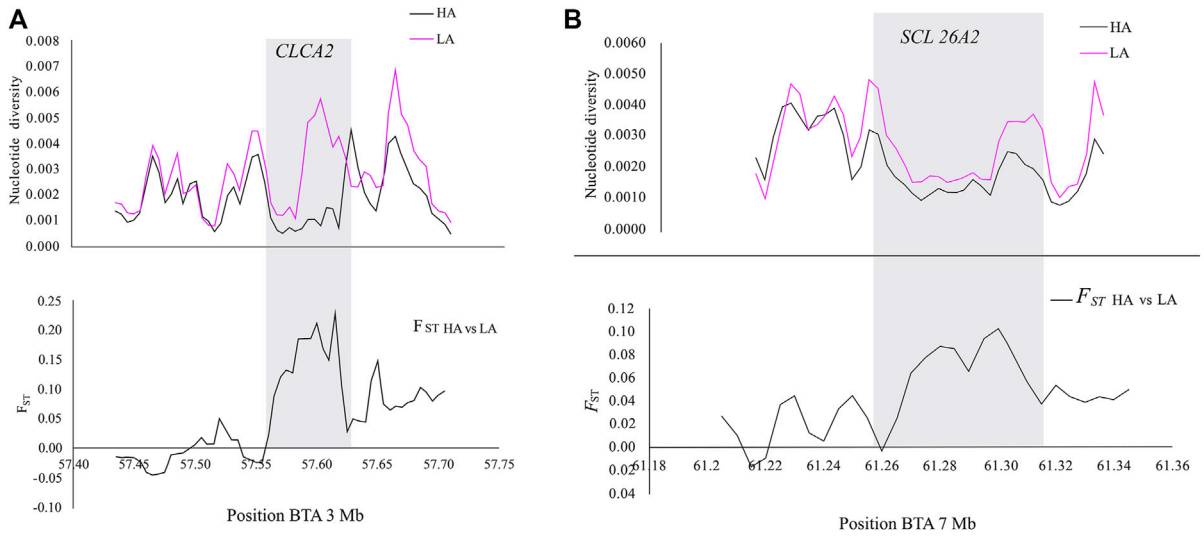


FIGURE 9
Plots of the nucleotide diversity within and F_{ST} values between the genomic regions of *CLCA2* **(A)** and *SCL26A2* **(B)** in high-altitude (HA) and low-altitude (LA) cattle populations.

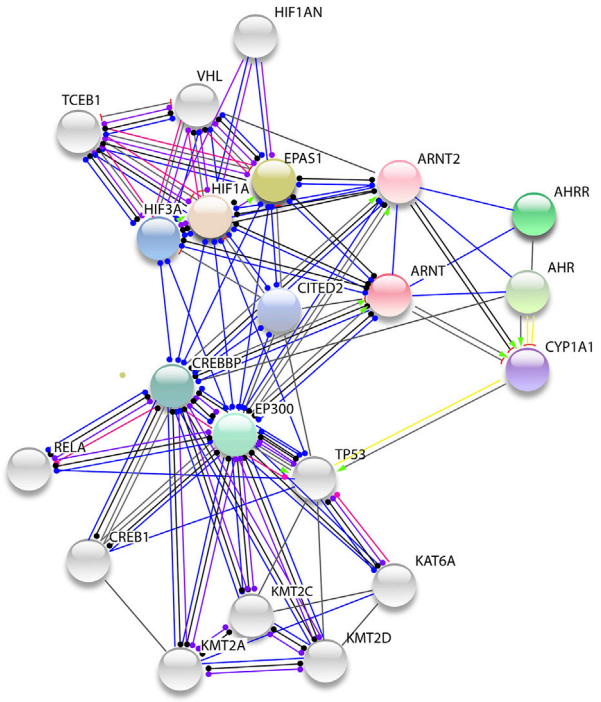


FIGURE 10
Protein-protein network of ARNT. The genes show several protein interaction networks interacting with hypoxia induced factors (HIF1a, EPAS1, CREBBP, and EP300).

high and low altitudes. Finally, additional comparisons of the candidate genomic regions between the high- and low-altitude cattle populations were carried out based on their nucleotide diversity, population differentiation, and haplotype LD heatmap differences for a detailed exploration of the high altitude adaptation.

Novel candidate gene identified in this study

We identified three novel strong candidate genes (*CLCA2, SLC26A2,* and *CBFA2T3*) (Figure 6) for high altitude adaptation along with the previously reported *ITPR2* gene. The functional analysis clustered the candidate genes into ion channel activity (*CLCA2*, *SLC26A2*, and *ITPR2*), response to hypoxia (*ITPR2* and *CBFA2T3*), and renin secretion KEGG pathway (*CLCA2*) (Table 3). The renin pathway and ion channel activity regulate smooth muscle tone and epithelial secretion in response to hypoxia (Al-Hashem et al., 2012; Palubiski et al., 2020) by controlling arterial blood flow pressure (Shimoda and Polak, 2011). Hypoxia induces the expression of *CLCA2* in the pulmonary artery smooth muscle of rats and controls cell proliferation and apoptosis in the ERK1/2-MAPK signaling pathway (Huang et al., 2017; Zhao et al., 2017). Similarly, the renin secretion pathways maintain the amount of plasma renin and aldosterone concentration by modulating the normal relationship between plasma osmolality and plasma vasopressin concentration in humans (Bestle et al., 2002; Savoia et al., 2011). The renin-angiotensin and vasopressin function is stimulated by increased blood pressure caused by vesicular smooth muscle tone (Chassagne et al., 2000) to regulate high blood flow to balance cellular oxygen demand.

SLC26A2 showed the highest PBS value in the high-altitude cattle populations (Figure 6). The haplotype LD heatmap, nucleotide diversity, and population differentiation index all supporting positive selection at the genome region overlapping with the gene (Figures 8, 9). The function of this gene is related to ion transport, and it plays a role in chondrocyte proliferation, differentiation, and growth in endochondral bone formation (Park et al., 2014). In humans, it regulates body size, and its recessive allele contributes to the dwarfism phenotype (Yang and Liang, 2021) and dysplasia (Pineda et al., 2013; Zheng et al., 2019; Heidari et al., 2021). A previous study reported a dominant allele at *SLC26A2* linked to higher heels and stronger claws in dairy cattle, while mutation at the gene causes dysplasia (Brenig et al., 2003). Considering the rugged and rocky terrain of the Ethiopian highlands, strong claws and high heels may prove advantageous. Further phenotypic characterization of the Ethiopian highland cattle may support this interpretation. The short stature and small body size of cattle observed in the Ethiopian high-altitude cattle confer the evidence. Though confirmatory analysis is required to differentiate the nature of short stature and small body size for HA adaptation in Ethiopian cattle, it could be a possible mechanism of the cold and high-altitude adaptation as it was reported in humans adapted to high altitude (Ma et al., 2019).

Candidate convergent genome evolution between cattle and humans living in the Ethiopian highlands

The human population in Ethiopia occupied the high altitudes thousands of years ago, expanding from the lower Rift Valley in the early Pleistocene age (Aldenderfer, 2006). Archaeological evidence suggests that humans inhabited Bale Mountain approximately 50–30 thousand years ago (Ossendorf et al., 2019). Today, the human communities occupying the high altitude areas where the cattle samples were collected are the Oromo (Bale Mountain) and the Amhara (Semien and Choke mountains). The beginning of the settlement of the Amhara to these high-altitude regions is thought to have started around 5,000 years (Alkorta-Aranburu et al., 2012),

while the settlement of the Oromo people was since early 1500s as reported by Hassen (1990) (cited in Alkorta-Aranburu et al., 2012; Huerta-Sánchez et al., 2013). The settlers in these territories were agrarian and had close interaction with their animals as sources of food and means of food production. Both humans and cattle living in the Ethiopian high altitudes share similar environmental challenges. Humans and ruminants living at high altitudes can be exposed to extended hypoxia stress and develop high-altitude sicknesses that may lead to high-altitude pulmonary hypertension (Friedrich and Wiener, 2020). For example, reports have indicated that cattle exposed to high altitudes may develop brisket disease caused by hypoxia (Newman et al., 2011; Wuletaw et al., 2011). Besides hypoxia, UV light and cold temperatures have been reported as major risk factors that challenge the survival of humans and other species in high altitude environments. Through a long-term evolutionary process, these risk factors may have induced positive selection pressures for physiological and morphological features that contribute to the evolutionary adaptation to high altitude environments (Witt and Huerta-Sánchez, 2019). Candidate genes detected in Ethiopian people living at high altitudes (Huerta-Sánchez et al., 2013), including *ITPR2*, *ARNT*, *EP300*, *MB*, and *HMOX1*, were also detected in Ethiopian cattle living in similar environments, supporting a convergent evolution between these two mammalian species.

Previous studies on the Ethiopian human population adapted to high altitudes have reported the *ITPR2* gene (Huerta-Sánchez et al., 2013) as a candidate gene. The *ITPR2* is also one of the candidate genes detected in HA cattle populations. It regulates vascular endothelial cells and intracellular calcium ion channel activity (Manalo et al., 2005; Jurkovicova et al., 2008). Following hypoxia, the cardiovascular system will increase blood flow by increasing pressure through vasoconstriction, increased heart rate, and myocardial contractility (Parati et al., 2015). These adaptive physiological mechanisms will enhance the supply of blood oxygen to tissues. *ITPR2* increases intracellular calcium concentration in vascular smooth muscle and it controls vasoconstriction avoiding pulmonary hypertension (Remillard and Yuan, 2006; Newman et al., 2011; Lai et al., 2015). *ITPR2*, as part of the calcium gated channel activities, also enhances endothelial cell proliferation lining and it triggers the vasculature and remodeling of the arterial tone to control the high blood pressure following hypoxic exposure (Makino et al., 2011; Hübner et al., 2015).

High altitude adaptation also depends on the concentration of hemoglobin in red blood cells and its affinity to oxygen in tissues (Alkorta-Aranburu et al., 2012). Also, increasing the number of erythrocytes will lead to higher hemoglobin concentration at the tissue level (Siebenmann et al., 2015). The candidate *MB* gene (Table 3) has been reported to play a role in increasing the hemoglobin concentration in muscle (Fraser et al., 2006; Jaspers et al., 2014) and increasing oxygen storage and binding affinity in hypoxic conditions (Hoppeler and Vogt, 2001; Li et al., 2018). This myoglobin gene was also reported under selection in the Ethiopian and Tibetan human populations living in highlands (Beall et al., 2002; Moore et al., 2002; Alkorta-Aranburu et al., 2012; Scheinfeldt et al., 2012). The gene is involved in erythrocyte homeostasis and regulates the level of hemoglobin (Avivi et al., 2010) in response to high altitude adaptation.

Selection signatures overlap between Ethiopia cattle and other species adapted to high altitudes

Several studies have reported candidate positive selection signatures for high-altitude adaptation in different species. Here, besides the overlap with human candidate selected regions, we identified several candidate regions which aligned with genes reported under selection in other species adapted to high altitudes. They include *MSRB3* with the highest XP-CLR score in our study and *MC1R* previously reported under selection in Ethiopian highland sheep (Edea et al., 2019). The *MSRB3* gene was also reported in Tibetan dogs and sheep (Vaysse et al., 2011; Wei et al., 2016; Witt and Huerta-Sánchez, 2019). It has a pleiotropic effect in being involved in the ossification and adipose tissue development in cattle (Saatchi et al., 2014). It has also been linked to ear size in Tibetan sheep (Wei et al., 2016), pigs (Zhang et al., 2015), and dogs (Vaysse et al., 2011). The *MSRB3* gene also protects cells from oxidative stress caused in mammals by hypoxia (Hansel et al., 2005), while it is linked to cold and heat tolerance in Drosophila (Lim et al., 2012) and cold tolerance in Arabidopsis (Kwon et al., 2007). Cold temperature is one of the environmental stressors that trigger animal cells to transduce energy to adapt to cold temperatures.

Last but not least, among the genes present within candidate genomic regions detected by both XP-CLR and iHS analyses, we do have *ABCB10* (Table 3). This gene was previously reported in candidate selected regions in humans, and several other species, including cattle (Bayeva et al., 2013; Martinez et al., 2020). Its function is related to iron metabolism and heme biosynthesis (Haase, 2010; Shah et al., 2013; Yamamoto et al., 2014; Seguin and Ward, 2018). *ABCB10* is also involved in the transport of heme out of the mitochondria, before hemoglobinization of erythropoietic cells (Liesa et al., 2012; Bayeva et al., 2013).

Conclusion

Despite the particularly challenging environmental conditions of the high-altitude Ethiopian highlands and the relatively recent arrival of African indicine cattle in these areas, we identified several genomic regions with evidence of positive selection for high-altitude environment adaptations at the autosomal level. These include genes previously reported in other mammalian species, including humans, living in high altitude areas in Ethiopia or other parts of the

world, as well as in Ethiopian-specific cattle genomic regions. Our results show that these indigenous livestock populations are locally adapted, and they have developed a physiological mechanism to cope with the environmental challenges of hypoxia, UV radiation, and cold temperature. It calls for the conservation of these indigenous cattle adaptations as well as for their utilization in breeding programs combining the improvement of productivity with adaptability.

Data availability statement

The original contributions presented in the study are publicly available. The WGS data of Bale and Semien cattle samples are available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA698721, while the Choke cattle samples WGS data is available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA841948/.

Ethics statement

Ethical review and approval were not required for the animal study because they were not applicable in the country. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

ET collected samples and carried out laboratory analysis. ET and AT performed analysis and data interpretation. OH reviewed critically the draft manuscript input from JH who assisted in the acquisition of the grant. OH and JH sequenced the data. ET wrote the draft manuscript, which was critically reviewed by AT, GB, OH, and JH. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to acknowledge the following institutions and personnel for funding and facilitating the research. The International Livestock Research Institute (ILRI) LiveGne program, supported by the CGIAR Research Program on Livestock (CRP livestock project) sponsored by the CGIAR funding contributors to the Trust Fund (http://www.cgiar.org/about-us/our-funders/). The Bill and Melinda Gates Foundation and UK aid from the UK Foreign, Commonwealth, and Development Office (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College), and The Chinese Government contribution to CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources in Beijing (2018-GJHZ-01). ET is a post-graduate fellow registered at Addis Ababa University (Ethiopia) and he was granted an ILRI post-graduate fellowship for his Ph.D. study.

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.2022.960234/full#supplementary-material

References

Ahbara, A., Bahbahani, H., Almathen, F., Abri, M. A., Agoub, M. O., Abeba, A., et al. (2019). Genome-wide variation, candidate regions and genes associated with fat deposition and tail morphology in Ethiopian indigenous sheep. *Front. Genet.* 10, 699–721. doi:10.3389/fgene.2018.00699

Al-Hashem, F. H., Alkhateeb, M. A., Shatoor, A. S., Khalil, M. A., and Sakr, H. F. (2012). Chronic exposure of rats to native high altitude increases in blood pressure via activation of the renin-angiotensin-aldosterone system. *Saudi Med. J.* 33, 1169–1176.

Aldenderfer, M. (2006). Modelling plateau peoples: The early human use of the world's high plateaux. *World Archaeol.* 38, 357–370. doi:10.1080/00438240600813285

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

Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K., and Di Rienzo, A. (2012). The genetic architecture of adaptations to high altitude in Ethiopia. *PLoS Genet.* 8, e1003110. doi:10.1371/journal.pgen.1003110

Asresie, A., and Zemedu, L. (2015). Contribution of livestock sector in Ethiopian economy : A review. *Adv. Life Sci. Technol.* 29, 79–91.

Avivi, A., Gerlach, F., Joel, A., Reuss, S., Burmester, T., Nevo, E., et al. (2010). Neuroglobin, cytoglobin, and myoglobin contribute to hypoxia adaptation of the subterranean mole rat Spalax. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21570–21575. doi:10.1073/pnas.1015379107

Azad, P., Stobdan, T., Zhou, D., Hartley, I., Akbari, A., Bafna, V., et al. (2017). High-altitude adaptation in humans: From genomics to

integrative physiology. *J. Mol. Med.* 95, 1269–1282. doi:10.1007/s00109-017-1584-7

Bayeva, M., Khechaduri, A., Wu, R., Burke, M. A., Wasserstrom, J. A., Singh, N., et al. (2013). ATP-binding cassette B10 regulates early steps of heme synthesis. *Circ. Res.* 113, 279–287. doi:10.1161/CIRCRESAHA.113.301552

Beall, C. M., Decker, M. J., Brittenham, G. M., Kushner, I., Gebremedhin, A., and Strohl, K. P. (2002). An Ethiopian pattern of human adaptation to high-altitude hypoxia. *Proc. Natl. Acad. Sci. U. S. A.* 99, 17215–17218. doi:10.1073/pnas.252649199

Bestle, M. H., Olsen, N. V., Poulsen, T. D., Roach, R., Fogh-Andersen, N., and Bie, P. (2002). Prolonged hypobaric hypoxemia attenuates vasopressin secretion and renal response to osmostimulation in men. *J. Appl. Physiol.* 92, 1911–1922. doi:10.1152/japplphysiol.00936.2001

Bigham, A. W., and Lee, F. S. (2014). Human high-altitude adaptation: Forward genetics meets the HIF pathway. *Genes Dev.* 28, 2189–2204. doi:10.1101/gad.250167.114

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Brenig, B., Baumgartner, B. G., Kriegesmann, B., Habermann, F., Fries, R., and Swalve, H. H. (2003). Molecular cloning, mapping, and functional analysis of the bovine sulfate transporter SLC26a2 gene. *Gene* 319, 161–166. doi:10.1016/S0378-1119(03)00806-0

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi:10.1086/521987

Chai, J., Jones, M. K., and Tarnawski, A. S. (2004). Serum response factor is a critical requirement for VEGF signaling in endothelial cells and VEGF-induced angiogenesis. *FASEB J.* 18, 1264–1266. doi:10.1096/fj.03-1232fje

Chassagne, C., Eddahibi, S., Adamy, C., Rideau, D., Marotte, F., Dubois-Rande, J.-L., et al. (2000). Modulation of angiotensin II receptor expression during development and regression of hypoxic pulmonary hypertension. *Am. J. Respir. Cell Mol. Biol.* 22, 323–332. doi:10.1165/ajrcmb.22.3.3701

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi:10.1101/gr.100545.109

Chen, Y., Yu, M., Dai, X., Zogg, M., Wen, R., Weiler, H., et al. (2011). Critical role for Gimap5 in the survival of mouse hematopoietic stem and progenitor cells. *J. Exp. Med.* 208, 923–935. doi:10.1084/jem.20101192

Copple, B. L., Bai, S., Burgoon, L. D., and Moon, J. O. (2011). Hypoxia-inducible factor-1α regulates the expression of genes in hypoxic hepatic stellate cells important for collagen deposition and angiogenesis. *Liver Int.* 31, 230–244. doi:10.1111/j.1478-3231.2010.02347.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Debevec, T., Millet, G. P., and Pialoux, V. (2017). Hypoxia-induced oxidative stress modulation with physical activity. *Front. Physiol.* 8, 84–89. doi:10.3389/fphys.2017.00084

Dolt, K. S., Mishra, M. K., Karar, J., Baig, M. A., Ahmed, Z., and Pasha, M. A. Q. (2007). cDNA cloning, gene organization and variant specific expression of HIF-1α in high altitude yak (Bos grunniens). *Gene* 386, 73–80. doi:10.1016/j.gene.2006.08.004

Dong, S.-S., He, W.-M., Ji, J.-J., Zhang, C., Guo, Y., and Yang, T.-L. (2020). LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *bioRxiv*, 151332. 06.14. doi:10.1101/2020.06.14.151332

Edea, Z., Dadi, H., Dessie, T., and Kim, K.-S. (2019). Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes Genomics* 41, 973–981. doi:10.1007/s13258-019-00820-y

Edea, Z., Dadi, H., Kim, S. W., Park, J. H., Shin, G. H., Dessie, T., et al. (2014). Linkage disequilibrium and genomic scan to detect selective loci in cattle populations adapted to different ecological conditions in Ethiopia. *J. Anim. Breed. Genet.* 131, 358–366. doi:10.1111/jbg.12083

Edwards, C. J., Bollongino, R., Scheu, A., Chamberlain, A., Tresset, A., Vigne, J. D., et al. (2007). Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proc. Biol. Sci.* 274, 1377–1385. doi:10.1098/rspb.2007.0020

Flori, L., Moazami-Goudarzi, K., Alary, V., Araba, A., Boujenane, I., Boushaba, N., et al. (2019). A genomic map of climate adaptation in Mediterranean cattle breeds. *Mol. Ecol.* 28, 1009–1029. doi:10.1111/mec.15004

Fraser, J., De Mello, L. V., Ward, D., Rees, H. H., Williams, D. R., Fang, Y., et al. (2006). Hypoxia-inducible myoglobin expression in nonmuscle tissues. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2977–2981. doi:10.1073/pnas.0508270103

Friedrich, J., and Wiener, P. (2020). Selection signatures for high-altitude adaptation in ruminants. *Anim. Genet.* 51, 157–165. doi:10.1111/age.12900

Gautier, M., Klassmann, A., and Vitalis, R. (2017). Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17, 78–90. doi:10.1111/1755-0998.12634

Ge, R. L., Cai, Q., Shen, Y. Y., San, A., Ma, L., Zhang, Y., et al. (2013). Draft genome sequence of the Tibetan antelope. *Nat. Commun.* 4, 1858–1867. doi:10.1038/ncomms2860

Geng, X., Feng, J., Liu, S., Wang, Y., Arias, C., and Liu, Z. (2014). Transcriptional regulation of hypoxia inducible factors alpha (HIF-α) and their inhibiting factor (FIH-1) of channel catfish (ictalurus punctatus) under hypoxia. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 169, 38–50. doi:10.1016/j.cbpb.2013.12.007

Graham, A. M., and Presnell, J. S. (2017). Hypoxia Inducible Factor (HIF) transcription factor family expansion, diversification, divergence and selection in eukaryotes. *PLoS One* 12, e0179545. doi:10.1371/journal.pone.0179545

Guan, C., Niu, Y., Chen, S. C., Kang, Y., Wu, J. X., Nishi, K., et al. (2020). Structural insights into the inhibition mechanism of human sterol O-acyltransferase 1 by a competitive inhibitor. *Nat. Commun.* 11, 2478–2511. doi:10.1038/s41467-020-16288-4

Guo, T., Zhao, H., Yuan, C., Huang, S., Zhou, S., Lu, Z., et al. (2021). Selective sweeps uncovering the genetic basis of horn and adaptability traits on fine-wool sheep in China. *Front. Genet.* 12, 604235. doi:10.3389/fgene.2021.604235

Haase, V. H. (2010). Hypoxic regulation of erythropoiesis and iron metabolism. *Am. J. Physiol. Ren. Physiol.* 299, F1–F13. doi:10.1152/ajprenal.00174.2010

Hansel, A., Heinemann, S. H., and Hoshi, T. (2005). Heterogeneity and function of mammalian MSRs: Enzymes for repair, protection and regulation. *Biochim. Biophys. Acta* 1703, 239–247. doi:10.1016/j.bbapap.2004.09.010

Heidari, M., Soleyman-Nejad, M., Isazadeh, A., Taskiri, M. H., Bolhassani, M., Sadighi, N., et al. (2021). Identification of a novel homozygous mutation in the DDR2 gene from a patient with spondylo-meta-epiphyseal dysplasia by whole exome sequencing. *Iran. J. Basic Med. Sci.* 24, 191–195. doi:10.22038/IJBMS.2020.44487.10405

Herbert, S. P., and Stainier, D. Y. R. (2012). Molecular control of endothelial cell behaviour during blood vessel morphogenesis. *Nat. Rev. Mol. Cell Biol.* 12, 551–564. doi:10.1038/nrm3176

Ho, C. H., and Tsai, S. F. (2017). Functional and biochemical characterization of a T cell-associated anti-apoptotic protein, GIMAP6. *J. Biol. Chem.* 292, 9305–9319. doi:10.1074/jbc.M116.768689

Hoppeler, H., and Vogt, M. (2001). Muscle tissue adaptations to hypoxia. *J. Exp. Biol.* 204, 3133–3139. doi:10.1242/jeb.204.18.3133

Hu, C.-J., Iyer, S., Sataur, A., Covello, K. L., Chodosh, L. A., and Simon, M. C. (2006). Differential regulation of the transcriptional activities of hypoxia-inducible factor 1 alpha (HIF-1alpha) and HIF-2alpha in stem cells. *Mol. Cell. Biol.* 26, 3514–3526. doi:10.1128/MCB.26.9.3514-3526.2006

Hu, X. J., Yang, J., Xie, X. L., Lv, F. H., Cao, Y. H., Li, W. R., et al. (2019). The genome landscape of Tibetan sheep reveals adaptive introgression from Argali and the history of early human settlements on the Qinghai-Tibetan Plateau. *Mol. Biol. Evol.* 36, 283–303. doi:10.1093/molbev/msy208

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Huang, L. J., Zhang, C. C., Zhao, M. P., Zheng, M. X., Ying, L., Chen, X. W., et al. (2017). The regulation of MAPK signaling pathway on cell proliferation and apoptosis in hypoxic PASMCs of rats. *Chin. J. Appl. Physiol.* 33, 226–230. doi:10.12047/j.cjap.5422.2017.056

Hübner, C. A., Schroeder, B. C., and Ehmke, H. (2015). Regulation of vascular tone and arterial blood pressure: Role of chloride transport in vascular smooth muscle. *Pflugers Arch.* 467, 605–614. doi:10.1007/s00424-014-1684-y

Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., et al. (2013). Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol. Biol. Evol.* 30, 1877–1888. doi:10.1093/molbev/mst089

Jaspers, R. T., Testerink, J., Gaspera, B. D., Chanoine, C., Bagowski, C. P., and Van Der Laarse, W. J. (2014). Increased oxidative metabolism and myoglobin expression in zebrafish muscle during chronic hypoxia. *Biol. Open* 3, 718–727. doi:10.1242/bio.20149167

Jurkovicova, D., Sedlakova, B., Lacinova, L., Kopacek, J., Sulova, Z., Sedlak, J., et al. (2008). Hypoxia differently modulates gene expression of inositol 1, 4, 5-trisphosphate receptors in mouse kidney and HEK 293 cell line. *Ann. N. Y. Acad. Sci.* 1148, 421–427. doi:10.1196/ANNALS.1410.034

Kim, J., Hanotte, O., Mwai, O. A., Dessie, T., Bashir, S., Diallo, B., et al. (2017). The genome landscape of indigenous African cattle. *Genome Biol.* 18, 34–14. doi:10.1186/s13059-017-1153-y

Kim, K., Kwon, T., Dessie, T., Yoo, D. A., Mwai, O. A., Jang, J., et al. (2020). The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism. *Nat. Genet.* 52, 1099–1110. doi:10.1038/s41588-020-0694-2

Kujala, M., Hihnala, S., Tienari, J., Kaunisto, K., Hästbacka, J., Holmberg, C., et al. (2007). Expression of ion transport-associated proteins in human efferent and epididymal ducts. *Reproduction* 133, 775–784. doi:10.1530/rep.1.00964

Kumar, B., Chile, S. A., Ray, K. B., Vidyadhar Reddy, G. E. C., Addepalli, M. K., Manoj Kumar, A. S., et al. (2011). VEGF-C differentially regulates VEGF-A expression in ocular and cancer cells; Promotes angiogenesis via RhoA mediated pathway. *Angiogenesis* 14, 371–380. doi:10.1007/s10456-011-9221-5

Kumar, P., Gullberg, U., Olsson, I., and Ajore, R. (2015). Myeloid translocation gene-16 co-repressor promotes degradation of hypoxia-inducible factor 1. *PLoS One* 10, e0123725. doi:10.1371/journal.pone.0123725

Kumar, P., Sharoyko, V. V., Spégel, P., Gullberg, U., Mulder, H., Olsson, I., et al. (2013). The transcriptional Co-repressor myeloid translocation gene 16 inhibits glycolysis and stimulates mitochondrial respiration. *PLoS One* 8, e68502. doi:10.1371/journal.pone.0068502

Kwon, S. J., Kwon, S. Il, Bae, M. S., Cho, E. J., and Park, O. K. (2007). Role of the methionine sulfoxide reductase MsrB3 in cold acclimation in Arabidopsis. *Plant Cell Physiol.* 48, 1713–1723. doi:10.1093/pcp/pcm143

Lai, N., Lu, W., and Wang, J. (2015). Ca2+ and ion channels in hypoxia-mediated pulmonary hypertension. *Int. J. Clin. Exp. Pathol.* 8, 1081–1092.

Li, C., Li, X., Liu, J., Fan, X., You, G., Zhao, L., et al. (2018). Investigation of the differences between the Tibetan and Han populations in the hemoglobin–oxygen affinity of red blood cells and in the adaptation to high-altitude environments. *Hematology* 23, 309–313. doi:10.1080/10245332.2017.1396046

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., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liesa, M., Qiu, W., and Shirihai, O. S. (2012). Mitochondrial ABC transporters function: The role of ABCB10 (ABC-me) as a novel player in cellular handling of reactive oxygen species. *Biochim. Biophys. Acta* 1823, 1945–1957. doi:10.1016/j.bbamcr.2012.07.013

Lim, D. H., Han, J. Y., Kim, J. R., Lee, Y. S., and Kim, H. Y. (2012). Methionine sulfoxide reductase B in the endoplasmic reticulum is critical for stress resistance and aging in Drosophila. *Biochem. Biophys. Res. Commun.* 419, 20–26. doi:10.1016/j.bbrc.2012.01.099

Liu, J., Yuan, C., Guo, T., Wang, F., Zeng, Y., Ding, X., et al. (2020). Genetic signatures of high-altitude adaptation and geographic distribution in Tibetan sheep. *Sci. Rep.* 10, 18332–18413. doi:10.1038/s41598-020-75428-4

Ma, Jia, Zhang, Z., Niu, W., Chen, J., Guo, S., Liu, S., et al. (2019). Education, altitude, and humidity can interactively explain spatial discrepancy and predict short stature in 213, 795 Chinese school children. *Front. Pediatr.* 7, 425–510. doi:10.3389/fped.2019.00425

Majmundar, A. J., Wong, W. J., and Simon, M. C. (2010). Hypoxia-inducible factors and the response to hypoxic stress. *Mol. Cell* 40, 294–309. doi:10.1016/j.molcel.2010.09.022

Makino, A., Firth, A. L., and Yuan, J. X.-J. (2011). "Endothelial and smooth muscle cell ion channels in pulmonary vasoconstriction and vascular remodeling," in *Comprehensive physiology* (Hoboken, NJ, USA: John Wiley & Sons), 139–148. doi:10.1002/cphy.c100023

Manalo, D. J., Rowan, A., Lavoie, T., Natarajan, L., Kelly, B. D., Ye, S. Q., et al. (2005). Transcriptional regulation of vascular endothelial cell responses to hypoxia by HIF-1. *Blood* 105, 659–669. doi:10.1182/blood-2004-07-2958

Martinez, M., Fendley, G. A., Saxberg, A. D., and Zoghbi, M. E. (2020). Stimulation of the human mitochondrial transporter ABCB10 by zinc-mesoporphrin. *PLoS One* 15, e0238754. doi:10.1371/journal.pone.0238754

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

Miron, N., and Tirosh, O. (2019). Cholesterol prevents hypoxia-induced hypoglycemia by regulation of a metabolic ketogenic shift. *Oxid. Med. Cell. Longev.* 2019, 5829357. doi:10.1155/2019/5829357

Moore, L. G. (2017). Measuring high-altitude adaptation. *J. Appl. Physiol.* 123, 1371–1385. doi:10.1152/japplphysiol.00321.2017

Moore, L. G., Zamudio, S., Zhuang, J., Droma, T., and Shohet, R. V. (2002). Analysis of the myoglobin gene in tibetans living at high altitude. *High. Alt. Med. Biol.* 3, 39–47. doi:10.1089/152702902753639531

Newman, J. H., Holt, T. N., Hedges, L. K., Womack, B., Memon, S. S., Willers, E. D., et al. (2011). High-altitude pulmonary hypertension in cattle (brisket disease): Candidate genes and gene expression profiling of peripheral blood mononuclear cells. *Pulm. Circ.* 1, 462–469. doi:10.4103/2045-8932.93545

Ohana, E., Shcheynikov, N., Park, M., and Muallem, S. (2012). Solute carrier family 26 member a2 (Slc26a2) protein functions as an electroneutral SOFormula/OH-/Cl-exchanger regulated by extracellular Cl-. *J. Biol. Chem.* 287, 5122–5132. doi:10.1074/jbc.M111.297192

Ossendorf, G., Groos, A. R., Bromm, T., Tekelemariam, M. G., Glaser, B., Lesur, J., et al. (2019). Middle Stone Age foragers resided in high elevations of the glaciated Bale Mountains, Ethiopia. *Science* 365, 583–587. doi:10.1126/science.aaw8942

Palubiski, L. M., O'Halloran, K. D., and O'Neill, J. (2020). Renal physiological adaptation to high altitude: A systematic review. *Front. Physiol.* 11, 756. doi:10.3389/fphys.2020.00756

Pan, Z., Li, S., Liu, Q., Wang, Z., Zhou, Z., Di, R., et al. (2018). Whole-genome sequences of 89 Chinese sheep suggest role of RXFP2 in the development of unique horn phenotype as response to semi-feralization. *Gigascience* 7, giy019–15. doi:10.1093/gigascience/giy019

Parati, G., Ochoa, J. E., Torlasco, C., Salvi, P., Lombardi, C., and Bilo, G. (2015). Aging, high altitude, and blood pressure: A complex relationship. *High. Alt. Med. Biol.* 16, 97–109. doi:10.1089/ham.2015.0010

Park, M., Ohana, E., Choi, S. Y., Lee, M. S., Park, J. H., and Muallem, S. (2014). Multiple roles of the SO4(2-)/Cl-/OH- exchanger protein Slc26a2 in chondrocyte functions. *J. Biol. Chem.* 289, 1993–2001. doi:10.1074/jbc.M113.503466

Pettersson-Kymmer, U., Lacroix, A., Eriksson, J., Bergstrom, U., Melin, B., Wibom, C., et al. (2013). Genome-wide association study meta-analysis identifies the SOAT1/AXDND1 locus to be associated with hip and forearm fracture risk. *Bone Abstr.* doi:10.1530/boneabs.2.1

Pineda, T., Rossi, A., Bonafè, L., Superti-Furga, A., and Velasco, H. M. (2013). Report of a novel mutation in the SLC26A2 gene found in a Colombian adult patient with diastrophic dysplasia. *Rev. Fac. Med.* 61, 255–259.

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

Qiu, Q., Zhang, G., Ma, T., Qian, W., Ye, Z., Cao, C., et al. (2012). The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44, 946–949. doi:10.1038/ng.2343

Qu, Y., Tian, S., Han, N., Zhao, H., Gao, B., Fu, J., et al. (2015). Genetic responses to seasonal variation in altitudinal stress : Whole-genome resequencing of great tit in eastern himalayas. *Sci. Rep.* 5, 14256–14310. doi:10.1038/srep14256

Ramakrishnan, S., Anand, V., and Roy, S. (2014). Vascular endothelial growth factor signaling in hypoxia and inflammation. *J. Neuroimmune Pharmacol.* 9, 142–160. doi:10.1007/s11481-014-9531-7

Remillard, C. V., and Yuan, J. X.-J. (2006). High altitude pulmonary hypertension: Role of K+ and Ca2+ channels. *High. Alt. Med. Biol.* 6, 133–146. doi:10.1089/ham.2005.6.133

Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* 15, 442. doi:10.1186/1471-2164-15-442

Sabeti, P. C. C., Reich, D. E. E., Higgins, J. M. M., Levine, H. Z. P. Z. P., Richter, D. J. J., Schaffner, S. F. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi:10.1038/nature01140

San, T., Polat, S., Cingi, C., Eskiizmir, G., Oghan, F., and Cakir, B. (2013). Effects of high altitude on sleep and respiratory system and theirs adaptations. *ScientificWorldJournal.* 2013, 241569. doi:10.1155/2013/241569

Savoia, C., Burger, D., Nishigaki, N., Montezano, A., and Touyz, R. M. (2011). Angiotensin II and the vascular phenotype in hypertension. *Expert Rev. Mol. Med.* 13, e11–e25. doi:10.1017/S1462399411001815

Scheinfeldt, L. B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., et al. (2012). Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.* 13, R1. doi:10.1186/gb-2012-13-1-r1

Schwefel, D., Arasu, B. S., Marino, S. F., Lamprecht, B., Köchert, K., Rosenbaum, E., et al. (2013). Structural insights into the mechanism of GTPase activation in the GIMAP family. *Structure* 21, 550–559. doi:10.1016/J.STR.2013.01.014

Seguin, A., and Ward, D. M. (2018). Mitochondrial ABC transporters and iron metabolism. *J. Clin. Exp. Pathol.* 08, 6–10. doi:10.4172/2161-0681.1000338

Shah, D. I., Takahashi-makise, N., Cooney, J. D., Li, L., Iman, J., Pierce, E. L., et al. (2013). Mitochondrial Atpif1 regulates heme synthesis in developing erythroblasts. *Nature*. 491, 608–612. doi:10.1038/nature11536.Mitochondrial

Shamimuzzaman, M., Le Tourneau, J. J., Unni, D. R., Diesh, C. M., Triant, D. A., Walsh, A. T., et al. (2020). Bovine genome database: New annotation tools for a new reference genome. *Nucleic Acids Res.* 48, D676–D681. doi:10.1093/nar/gkz944

Shimoda, L. A., and Polak, J. (2011). Hypoxia. 4. Hypoxia and ion channel function. *Am. J. Physiol. Cell Physiol.* 300, 951–967. doi:10.1152/ajpcell.00512.2010

Siebenmann, C., Cathomen, A., Hug, M., Keiser, S., Lundby, A. K., Hilty, M. P., et al. (2015). Hemoglobin mass and intravascular volume kinetics during and after exposure to 3, 454-m altitude. *J. Appl. Physiol.* 119, 1194–1201. doi:10.1152/japplphysiol.01121.2014

Simonson, T. S. (2015). Altitude adaptation: A glimpse through various lenses. *High. Alt. Med. Biol.* 16, 125–137. doi:10.1089/ham.2015.0033

Storz, J. F. (2007). Hemoglobin function and physiological adaptation to hypoxia in high-altitude mammals. *J. Mammal.* 88, 24–31. doi:10.1644/06-MAMM-S-199R1.1

Storz, J. F. (2016). Hemoglobin–oxygen affinity in high-altitude vertebrates: Is there evidence for an adaptive trend? *J. Exp. Biol.* 219, 3190–3203. doi:10.1242/jeb.127134

Storz, J. F. (2021). High-altitude adaptation: Mechanistic insights from integrated genomics and physiology. *Mol. Biol. Evol.* 38, 2677–2691. doi:10.1093/molbev/msab064

Storz, J. F., and Moriyama, H. (2008). Mechanisms of hemoglobin adaptation to high altitude hypoxia. *High. Alt. Med. Biol.* 9, 148–157. doi:10.1089/ham.2007.1079

Storz, J. F., Scott, G. R., and Cheviron, Z. A. (2010). Phenotypic plasticity and genetic adaptation to high-altitude hypoxia in vertebrates. *J. Exp. Biol.* 213, 4125–4136. doi:10.1242/jeb.048181

Sunny, D. E., Hammer, E., Strempel, S., Joseph, C., Manchanda, H., Ittermann, T., et al. (2020). Nup133 and ERα mediate the differential effects of hyperoxia-induced damage in male and female OPCs. *Mol. Cell. Pediatr.* 7, 10. doi:10.1186/s40348-020-00102-8

Touré, A. (2019). Importance of slc26 transmembrane anion exchangers in sperm post-testicular maturation and fertilization potential. *Front. Cell Dev. Biol.* 7, 230. doi:10.3389/fcell.2019.00230

Valverde, G., Zhou, H., Lippold, S., De Filippo, C., Tang, K., Herráez, D. L., et al. (2015). A novel candidate region for genetic adaptation to high altitude in Andean populations. *PLoS One* 10, 01254444–e125522. doi:10.1371/journal.pone.0125444

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Mol. Ecol.* 25, 89–103. doi:10.1111/mec.13360

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Pielberg, G. R., Sigurdsson, S., et al. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.* 7, e1002316. doi:10.1371/journal.pgen.1002316

Veith, C., Schermuly, R. T., Brandes, R. P., and Weissmann, N. (2016). Molecular mechanisms of hypoxia-inducible factor-induced pulmonary arterial smooth muscle cell alterations in pulmonary hypertension. *J. Physiol.* 594, 1167–1177. doi:10.1113/JP270689

Verma, P., Sharma, A., Sodhi, M., Thakur, K., Bharti, V. K., Kumar, P., et al. (2018). Overexpression of genes associated with hypoxia in cattle adapted to Trans Himalayan region of Ladakh. *Cell Biol. Int.* 9999, 1141–1148. doi:10.1002/cbin.10981

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72–e0458. doi:10.1371/journal.pbio.0040072

Wang, K., Yang, Y., Wang, L., Ma, T., Shang, H., Ding, L., et al. (2016). Different gene expressions between cattle and yak provide insights into high-altitude adaptation. *Anim. Genet.* 47, 28–35. doi:10.1111/age.12377

Wang, M. D., Dzama, K., Rees, D. J. G., and Muchadeyi, F. C. (2015). Tropically adapted cattle of Africa: Perspectives on potential role of copy number variations. *Anim. Genet.* 47, 154–164. doi:10.1111/age.12391

Wei, C., Wang, H., Liu, G., Zhao, F., Kijas, J. W., Ma, Y., et al. (2016). Genome-wide analysis reveals adaptation to high altitudes in Tibetan sheep. *Sci. Rep.* 6, 26770–26811. doi:10.1038/srep26770

Weir, E. K., and Olschewski, A. (2006). Role of ion channels in acute and chronic responses of the pulmonary vasculature to hypoxia. *Cardiovasc. Res.* 71, 630–641. doi:10.1016/j.cardiores.2006.04.014

Werhahn, G., Senn, H., Ghazali, M., Karmacharya, D., Sherchan, A. M., Joshi, J., et al. (2018). The unique genetic adaptation of the Himalayan wolf to high-altitudes and consequences for conservation. *Glob. Ecol. Conserv.* 16, e00455. doi:10.1016/J.GECCO.2018.E00455

Wiener, P., Robert, C., Ahbara, A., Salavati, M., Abebe, A., Kebede, A., et al. (2021). Whole-genome sequence data suggest environmental adaptation of Ethiopian sheep populations. *Genome Biol. Evol.* 13, 1–18. doi:10.1093/gbe/evab014

Witt, K. E., and Huerta-Sánchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374, 20180235. doi:10.1098/rstb.2018.0235

Wuletaw, Z., Wurzinger, M., Holt, T., Dessie, T., and Sölkner, J. (2011). Assessment of physiological adaptation of indigenous and crossbred cattle to hypoxic environment in Ethiopia. *Livest. Sci.* 138, 96–104. doi:10.1016/j.livsci.2010.12.005

Yalcin, O., and Cabrales, P. (2012). Increased hemoglobin O2 affinity protects during acute hypoxia. *Am. J. Physiol. Heart Circ. Physiol.* 303, H271–H281. doi:10.1152/ajpheart.00078.2012

Yamamoto, M., Arimura, H., Fukushige, T., Minami, K., Nishizawa, Y., Tanimoto, A., et al. (2014). Abcb10 role in heme biosynthesis *in vivo* : Abcb10 knockout in mice causes anemia with protoporphyrin IX and iron accumulation. *Mol. Cell. Biol.* 34, 1077–1084. doi:10.1128/MCB.00865-13

Yang, J., Jin, Z.-B., Chen, J., Huang, X.-F., Li, X.-M., Liang, Y.-B., et al. (2017). Genetic signatures of high-altitude adaptation in Tibetans. *Proc. Natl. Acad. Sci. U. S. A.* 114, 4189–4194. doi:10.1073/pnas.1617042114

Yang, L. L., and Liang, S. S. (2021). Study on pathogenic genes of dwarfism disease by next-generation sequencing. *World J. Clin. Cases* 9, 1600–1609. doi:10.12998/wjcc.v9.i7.1600

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78. doi:10.1126/science.1190371

Zhang, B., Niu, W., Dong, H. Y., Liu, M. L., Luo, Y., and Li, Z. C. (2018). Hypoxia induces endothelial-mesenchymal transition in pulmonary vascular remodeling. *Int. J. Mol. Med.* 42, 270–278. doi:10.3892/ijmm.2018.3584

Zhang, H., Wu, C. X., Chamba, Y., and Ling, Y. (2007). Blood characteristics for high altitude adaptation in Tibetan chickens. *Poult. Sci.* 86, 1384–1389. doi:10.1093/ps/86.7.1384

Zhang, W., Fan, Z., Han, E., Hou, R., Zhang, L., Galaverni, M., et al. (2014). Hypoxia adaptations in the grey wolf (*Canis lupus* chanco) from qinghai-tibet plateau. *PLoS Genet.* 10, e1004466. doi:10.1371/journal.pgen.1004466

Zhang, Y., Liang, J., Zhang, L., Wang, Ligang, Liu, X., Yan, H., et al. (2015). Porcine methionine sulfoxide reductase B3: Molecular cloning, tissue-specific expression profiles, and polymorphisms associated with ear size in *Sus scrofa*. *J. Anim. Sci. Biotechnol.* 6, 60–69. doi:10.1186/s40104-015-0060-x

Zhao, M. P., Ma, Y. C., Zhang, C. C., Huang, L. J., Zheng, M. X., Li, G. L., et al. (2017). The effects of ERK1/2 pathway on the expression of calcium activated chloride channel in hypoxia in PASMCs rat model. *Zhongguo Ying Yong Sheng Li Xue Za Zhi* 33, 47–50. doi:10.12047/j.cjap.5448.2017.011

Zheng, C., Lin, X., Xu, X., Wang, C., Zhou, J., Gao, B., et al. (2019). Suppressing UPR-dependent overactivation of FGFR3 signaling ameliorates SLC26A2-deficient chondrodysplasias. *EBioMedicine* 40, 695–709. doi:10.1016/j.ebiom.2019.01.010

Zheng, W. S., He, Y. X., Cui, C. Y., Ouzhu, L., Deji, Q., Peng, Y., et al. (2017). EP300 contributes to high-altitude adaptation in Tibetans by regulating nitric oxide production. *Zool. Res.* 38, 163–170. doi:10.24272/j.issn.2095-8137.2017.036

Zuniga-Hertz, J. P., and Patel, H. H. (2019). The evolution of cholesterol-rich membrane in oxygen adaption: The respiratory system as a model. *Front. Physiol.* 10, 1340–1418. doi:10.3389/fphys.2019.01340

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fgene.2022.1084824

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Seyed Abbas Rafat,
University of Tabriz, Iran
Xiaolong Kang,
Ningxia University, China

*CORRESPONDENCE
Junli Sun,
sjn313@126.com

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

RECEIVED 31 October 2022
ACCEPTED 27 December 2022
PUBLISHED 09 January 2023

CITATION
Chen Z, Zhu M, Wu Q, Lu H, Lei C, Ahmed Z and Sun J (2023), Analysis of genetic diversity and selection characteristics using the whole genome sequencing data of five buffaloes, including Xilin buffalo, in Guangxi, China.
Front. Genet. 13:1084824.
doi: 10.3389/fgene.2022.1084824

Analysis of genetic diversity and selection characteristics using the whole genome sequencing data of five buffaloes, including Xilin buffalo, in Guangxi, China

Zhefu Chen[1,2], Min Zhu[1], Qiang Wu[1], Huilin Lu[1], Chuzhao Lei[2], Zulfiqar Ahmed[3] and Junli Sun[1]*

[1]Guangxi Key Laboratory of Livestock Genetic Improvement, Animal Husbandry Research Institute of Guangxi Zhuang Autonomous Region, Nanning, China, [2]College of Animal Science and Technology, Northwest A&F University, Xianyang, China, [3]Faculty of Veterinary and Animal Sciences, University of Poonch Rawalakot, Rawalakot, China

Buffalo is an economically important livestock that renders useful services to manhood in terms of meat, milk, leather, and draught. The Xilin buffalo is among the native buffalo breeds of China. In the present study, the genetic architecture and selection signature signals of Xilin buffalo have been explored. Correlation analysis of the population structure of Xilin buffalo was conducted by constructing NJ tree, PCA, ADMIXTURE and other methods. A total of twenty-five (n = 25) Xilin buffalo whole genome data and data of forty-six (n = 46) buffaloes published data were used. The population structure analysis showed that the Xilin buffalo belong to the Middle-Lower Yangtze. The genome diversity of Xilin buffalo was relatively high. The CLR, π ratio, $F_{ST,}$ and XP-EHH were used to detect the candidate genes characteristics of positive selection in Xilin buffalo. Among the identified genes, most of the enriched signal pathways were related to the nervous system and metabolism. The mainly reported genes were related to the nervous system (*GRM5, GRIK2, GRIA4*), reproductive genes (*CSNK1G2, KCNIP4*), and lactation (*TP63*). The results of this study are of great significance for understanding the molecular basis of phenotypic variation of related traits of Xilin buffalo. We provide a comprehensive overview of sequence variations in Xilin buffalo genomes. Selection signatures were detected in genomic regions that are possibly related to economically important traits in Xilin buffalo and help in future breeding and conservation programs of this important livestock genetic resource.

KEYWORDS

Xilin buffalo, genomic diversity, population structure, genetic signatures, whole genome sequencing

1 Introduction

Domestic buffaloes are predominantly distributed in Asian countries. According to behavior and chromosome karyotype, domestic buffaloes are divided into two types: riverine buffalo (*Bubalus bubalis*, 2 n = 50) and swamp buffalo (*Bubalus bubalis* carabanensis, 2n = 48) (Fischer and Ulbrich, 1967). As an important economic livestock species in the world, the important traits e.g., milk production, growth, reproduction, hair color, etc. Have been focused previously as important indicators for selection (Liu et al., 2018). From the year 1999–2019, the number of buffalo increased by about 25.9% which in turn increased

milk production by 106% and buffalo beef production by 45% (Di Stasio and Brugiapaglia, 2021). In addition, buffalo is a good drought animal that compensate for about 20%–30% of the agricultural labor force (Michelizzi et al., 2010). Although with the popularization of mechanization, the role of buffalo as a servant has been gradually replaced, it is still the most important source of labor in some remote mountainous areas in southern China. The buffaloes are used for plowing the agricultural land, particularly paddy rice fields. In addition, buffaloes are used in a cart for transporting heavier goods as compared with cattle (Michelizzi et al., 2010).

The Xilin buffalo is mainly produced in the plateau and mountainous areas of Xilin Longlin and Tianlin County of Guangxi. It is one of the local buffalo varieties in Guangxi. Due to the influence of natural ecological and environmental conditions and local socioeconomic activities, it is gradually formed after long -term natural selection and artificial selection. The Xilin buffalo is characterized by a large body size, gentle temperament, strong farming ability, good growth and development, efficient consumption of roughages, good mountain climbing, strong adaptability, and disease resistance (He et al., 2011). At the present, there is a single germplasm conservation farm in the main producing area which is primarily used for the production of hybrid females by crossing the local buffaloes with foreign excellent varieties (such as Murrah buffalo and Italian buffalo, mainly Murrah buffalo) for improvement of milk and meat production.

With the development of whole genome resequencing (WGS) technology, the reduction of sequencing cost, the genetic structure, evolutionary history, origin, and domestication of domestic animals such as pigs, cattle and sheep, etc. Have been widely and systematically studied become possible as an effective cost tool (Stothard et al., 2011).

Many WGS-based buffalo studies initially concentrated on the economically relevant characteristics of commercial breeds (Li et al., 2020). The genetic characteristics and selection pressure signals of Xilin buffalo have not been deeply studied by using WGS data earlier. The study on the genetic structure and population history of Xilin buffalo is helpful to analyze the genetic basis of adaptability and other traits and provides a theoretical basis for the improvement and conservation of Xilin buffalo varieties.

2 Materials and methods

2.1 Sample collection and sequencing

Blood and ear tissue samples were collected from the native home tract (Xilin County of Guangxi Province, China) of pure Xilin buffaloes (n = 25). The genomic DNA was extracted by the standard phenol-chloroform method (Green and Sambrook, 2012) and subjected to Illumina NovaSeq sequencing at Novogene Bioinformatics Institute, Beijing, China. By using pair-end sequencing technology an average insertion size of 500 bp was constructed for each sample and the average reading length was 150 bp. In addition, 46 published whole-genome sequences data of swamp buffalo including Guizhou white (n = 10), Binhu (n = 3), Fuzhong (n = 11), and Mediterranean (n = 22) were downloaded from NCBI(PRJNA547460) which fully described the characteristics of population structure, genetic diversity, single nucleotide polymorphisms (SNPs), and natural or artificial selection. The details of the five varieties are listed in Table 1.

TABLE 1 Sample information of 71 buffaloes from 5 buffalo breeds.

Varieties	Abbreviation	Sample size	Type
Xilin Buffalo	XL	25	Swamp
Guizhou White Buffalo	GZB	10	Swamp
Binhu buffalo	BH	3	Swamp
Fuzhong Buffalo	FZ	11	Swamp
Mediterranean Buffalo	MD	22	River

2.2 Construction of buffalo pseudo chromosome

In the present study, the published buffalo data were obtained from the reference genome assembly of buffalo (GCF_000471725.1) from NCBI. However, due to the complexity of the data, it is only assembled to the scaffold level. If it is directly used for comparison and subsequent analysis, it will lead to a double increase in computing and storage resources. Therefore, this study used the method of artificial connection of pseudo chromosomes. The reference genome is connected to 24 + X + unplaced chromosomes which can reflect the authenticity of chromosomes to the greatest extent (Amaral et al., 2008).

2.3 Genome wide alignment and variation detection

The sequenced reads after quality control were compared to the constructed buffalo pseudo chromosome by BWA-MEM (Li and Durbin, 2009), and repeated reads introduced by PCR were removed by Picard. The genome-wide high-quality genetic variation was detected by GATK (version 3.6-0-g89b7209) (Nekrutenko and Taylor, 2012) where the filtering conditions of SNP were as follows: (1) QD (Quality by Depth) < 2; (2) variants with FS (Phred-scaled p-value using Fisher's exact test to detect strand bias) > 60 were filtered; MQRankSum (Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities) < 12.5; (4) ReadPosRankSum (evaluate the reliability of variation by the position of variation in read) < −8; (5) MQ (RMS Mapping Quality) < 40.0; (6) Mean sequencing depth > 3x or < 1/3x (7) SOR (StrandOddsRatio) > 3.0; (8) maximum missing rate < .1; (9) SNP is strictly limited to double alleles. The Annovar software was used to annotate the variant information.

2.4 Analysis of buffalo population structure

First, VCF files of SNPs of 71 buffalo were converted into corresponding Plink files (bed. bim. fam by using vcftools) and PLINK (version 1.9) (Purcell et al., 2007) software were used to filter out the linkage disequilibrium sites with R2 greater than .2. The parameter is set as: -- indep pairwise 50 50.2. The filtered data were used to construct NJ tree, PCA, ADMIXTURE and other population structure-related analyses. In order to clarify, the

phylogenetic relationship of 71 Buffalo, adjacency tree (NJ phylogenetic tree) (Yu et al., 2021) is constructed in this study. The genetic distance matrix is calculated by using the parameter "-- distance matrix" of PLINK 1.9 and then the matrix is transformed into. meg format, which will get imported .meg format into the MEGA6.0 software. Build the NJ phylogenetic tree and set the bootstrap value to 1000. Finally, using online iTOL (https://itol.embl.de/), the tool displays the obtained phylogenetic tree and beautifies it. The software package EIGENSOFT V5.0 and SmartPCA (Patterson et al., 2013) were used for PCA analysis of filtered buffalo autosomal SNP data sets. The significance of each eigenvector is calculated by the Tracy-Widom test. Admixture v. 1.3.0 (Alexander et al., 2009) was used to analyse the ancestral components of 71 buffalo autosomal SNP data sets. This study simulates that from k = 2 to k = 5. The bootstrap value of each k value was set to 20 and the optimal value was finally obtained according to the Cross-Validation (CV) value.

2.5 Genetic diversity, linkage disequilibrium and ROH detection

We used VCFtools to estimate the nucleotide diversity of each breed in window sizes of 50 kb with 50 kb increments. The Linkage disequilibrium (LD) decay with the physical distance between SNPs was calculated and visualized by using PopLDdecay software with default parameters (Rahimmadar et al., 2021). The run of homozygosity (ROH) was identified using the--homozyg option implemented in PLINK which slides a window of 50 SNPs (-homozyg-window-snp 50) across the genome estimating homozygosity (Makanjuola et al., 2021). The following settings were performed for ROH identification: (1) required minimum density (–homozyg-density 50); (2) number of heterozygotes allowed in a window (–homozyg-window-het 3); (3) the number of missing calls allowed in window (–homozyg-window-missing 5). The number and length of ROH for each breed were estimated and length of ROH was divided into three categories: .5–1 Mb, 1–2 Mb, 2–4 Mb. (Forutan et al., 2018). F_{ROH} is calculated by calculating the ratio of the total length of ROH fragments in the genome to the total length (L_{ROH}) of the genome (L_{auto}). The formula is as follows: $F_{ROH} = \sum L_{ROH}/L_{auto}$

2.6 Selective scanning recognition

We adopted the following strategies for genome scanning of Xilin buffalo. First, we utilized nucleotide diversity (θπ) (Hudson, 1992) and the composite likelihood ratio test (CLR) (Nielsen et al., 2005) to detect the selection characteristics of Xilin buffalo. By using VCFtools, the nucleotide diversity was estimated using a sliding window of 50 kb and a step size of 20 kb. We used SweepFinder to calculate the CLR test of the sites in the non-overlapping 50 kb window in order to calculate the empirical *p*-value of π and CLR window and take the overlapping part of the first 1% window of each method as the candidate mark for selection.

Second, we performed comparisons between Xilin buffalo and Mediterranean buffalo using fixation index (F_{ST}) (Hudson, 1992) and cross-group extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007). FST analysis was calculated in 50 kb windows with a 20 kb step using VCFtools (Danecek et al., 2011). XP-EHH statistics based on the extended haplotype was calculated for each population pair using selscan v1.1 (Szpiech and Hernandez, 2014). For XP-EHH selective scanning, our test statistic is the average normalized XP-EHH score of each 50 kb region. An XP-EHH score is directional: a positive score suggests that selection is likely to have happened in Xilin buffalo, whereas a negative score suggests the same about reference population. Significant genomic regions were identified by *p*-value < .01. Genomic regions identified by at least two methods were considered to be candidate regions of positive selection.

To better understand, the gene function and signaling pathways of the identified candidate genes, KOBAS 3.0 was used for GO and KEGG pathway enrichment analysis (Shen et al., 2019). Only when the corrected *p*-value < .05, were the GO and KEGG pathways considered significantly enriched.

3 Results

3.1 Identification of single nucleotide polymorphisms

In this study, individual genomes of 25 Xilin buffaloes were generated to ~ 12.1 × coverage each and were jointly genotyped with publicly available genomes of three buffalo populations from different regions of China and Mediterranean Buffalo (Italy), and the average mapping rate was 99.37% (Supplementary Table S1). In total, ~ 5.0 billion reads of sequences were generated. Using BWA-MEM, reads were aligned to the buffalo reference genome sequence (GCA_000471725.1) with an average of 10.6 × coverage. We annotated 28,347,965 biallelic SNPs found in 71 buffaloes. genomic annotation for showing the location of those SNPs that most of the SNPs existed in the intron region (65.532%) or intergenic region (19.514%). The exon contains merely 2.15% of the total SNPs with 529,920 synonymous SNPs (Supplementary Table S3).

3.2 Population genetic structure and genetic relationship

At present, there are ~202 million buffaloes in the world, mainly distributed in Asia (196 million, accounting for 97.0%), Africa (3.4 million) and South America (~2 million). (Zhang et al., 2007). According to the previous research of Sun et al. (2020), Asian buffaloes are divided into five regions according to their geographical distribution: the upper reaches of the Yangtze River, the middle and lower reaches of the Yangtze River, Southwest China, Southeast Asia and South Asia, and added Italy (Mediterranean buffalo). NJ phylogenetic tree was constructed from the whole genome data of 71 buffalo. As shown in (Figure 1A), the different colors represent buffaloes in different regions. These 71 buffalo are mainly divided into two branches: swamp and river buffalo. As for swamp buffalo is concerned, the buffalo in the same geographical area gather together. Some individuals are in the middle of the two types of buffalo in the phylogenetic tree which represents the hybrid individuals produced by the hybridization of the two types of buffalo. Principal component analysis (PCA) was used to further explore the genetic relationship between different buffalo populations. The

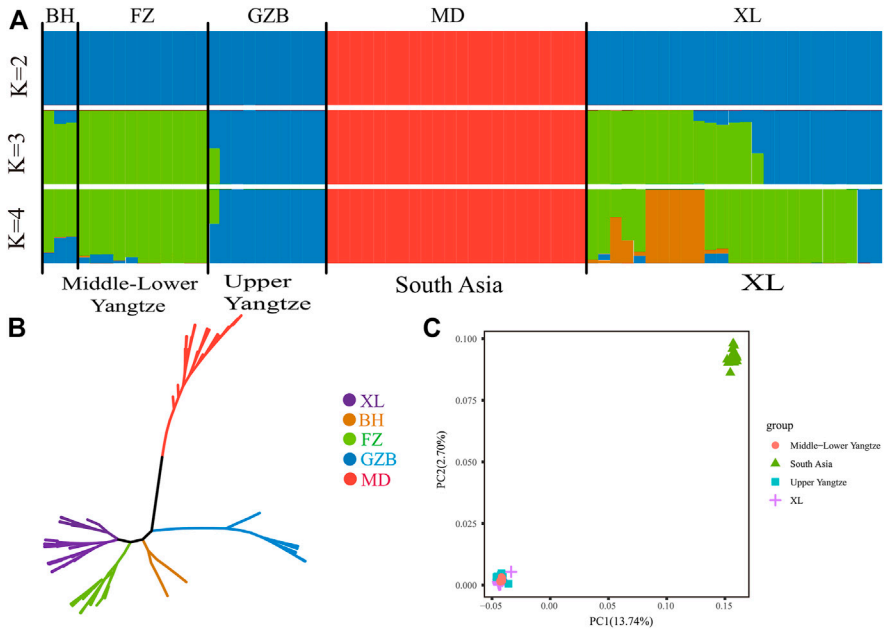


FIGURE 1
Population structure and relationships of Xinlin Buffalo. **(A)** Model-based clustering of buffalo using the ADMIXTURE program with K = 2 to 4(X). **(B)** Neighbour-joining tree of buffaloes constructed using whole-genome autosomal SNP data. **(C)** Principal component analysis (PCA) showing PC1 against PC2. The X axis represents PC1, and the Y axis represents PC2.

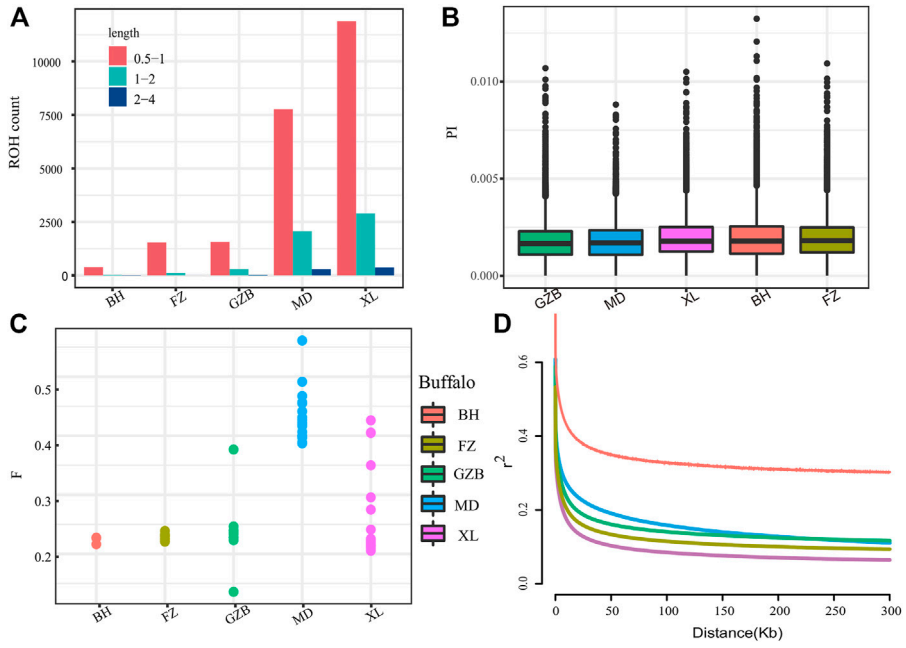


FIGURE 2
Summary statistics for genomic variation: **(A)** The distribution of the total number of ROH across chromosomes. **(B)**Genome-wide distribution of nucleotide diversity of each breed in 50 kb windows with 20 kb steps. **(C)** Inbreeding coefficient from each breed. **(D)** Genome-wide average LD decay estimated from each breed. The X axis is the physical distance (kb), and the Y axis is the LD coefficient (r^2).

results of PCA show that PC1(13.74%) and PC2(2.70%) distinguish riverine from swamp buffalo and the results of PC3(2.48%) show that Xilin buffalo is more similar to Middle lower Yangtze buffalo which is consistent with the literature (Figures 1B, C).

The whole genome data of 71 buffalo were analyzed by ADMIXTURE in order to perform ancestral component analysis (Figure 1A). When there is k = 2 it indicates the buffaloes of riverine and swamp origin. When there is k = 3, it

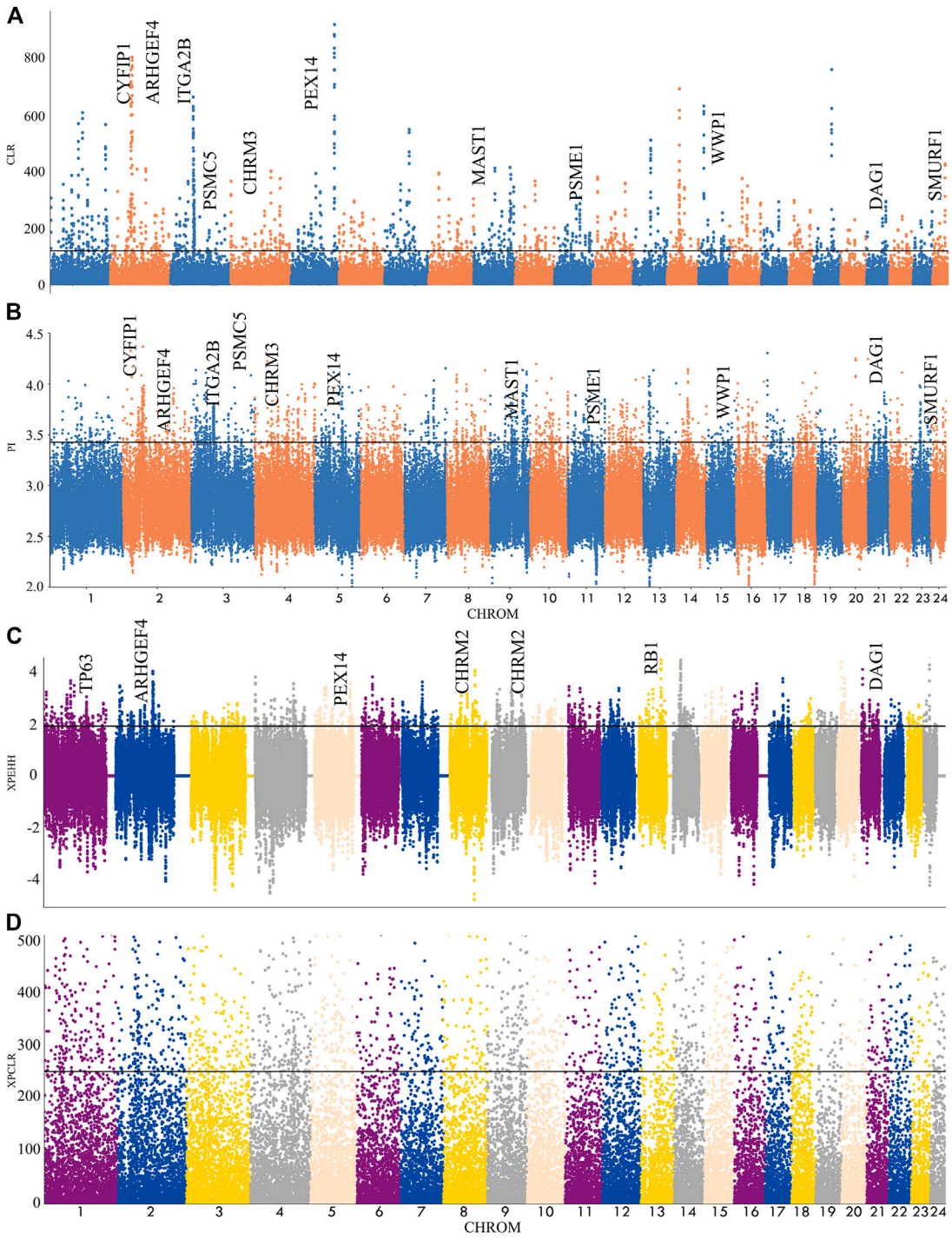


FIGURE 3
Genome-wide selection scan in Xilin buffalo using sliding window analysis (50 kb window size, 20 kb step size, 99th percentile cut-off). **(A)**. Selection signatures in Xilin buffalo for CLR (Mediterranean and Xilin). **(B)**. Selection signatures in Xilin buffalo for π-ratio (Mediterranean and Xilin). **(C)** Selection signatures in Xilin buffalo for XPEHH (Mediterranean and Xilin). **(D)** Selection signatures in Xilin buffalo for XPCLR (Mediterranean and Xilin). The thresholds (top 1%) of F_{ST}, π-ratio, XPEHH and XPCLR are marked with a horizontal black line.

shows that swamp buffaloes can be divided into two groups: Middle-Lower Yangtze buffalo (green) and Upper Yangtze buffalo (blue). When k = 4, it represented that the Xilin buffalo is classified as Yangtze River buffalo.

3.3 Genomic variation pattern

The runs of homozygosity (ROH) are a continuous homozygous region in the DNA sequence of diploid organisms.

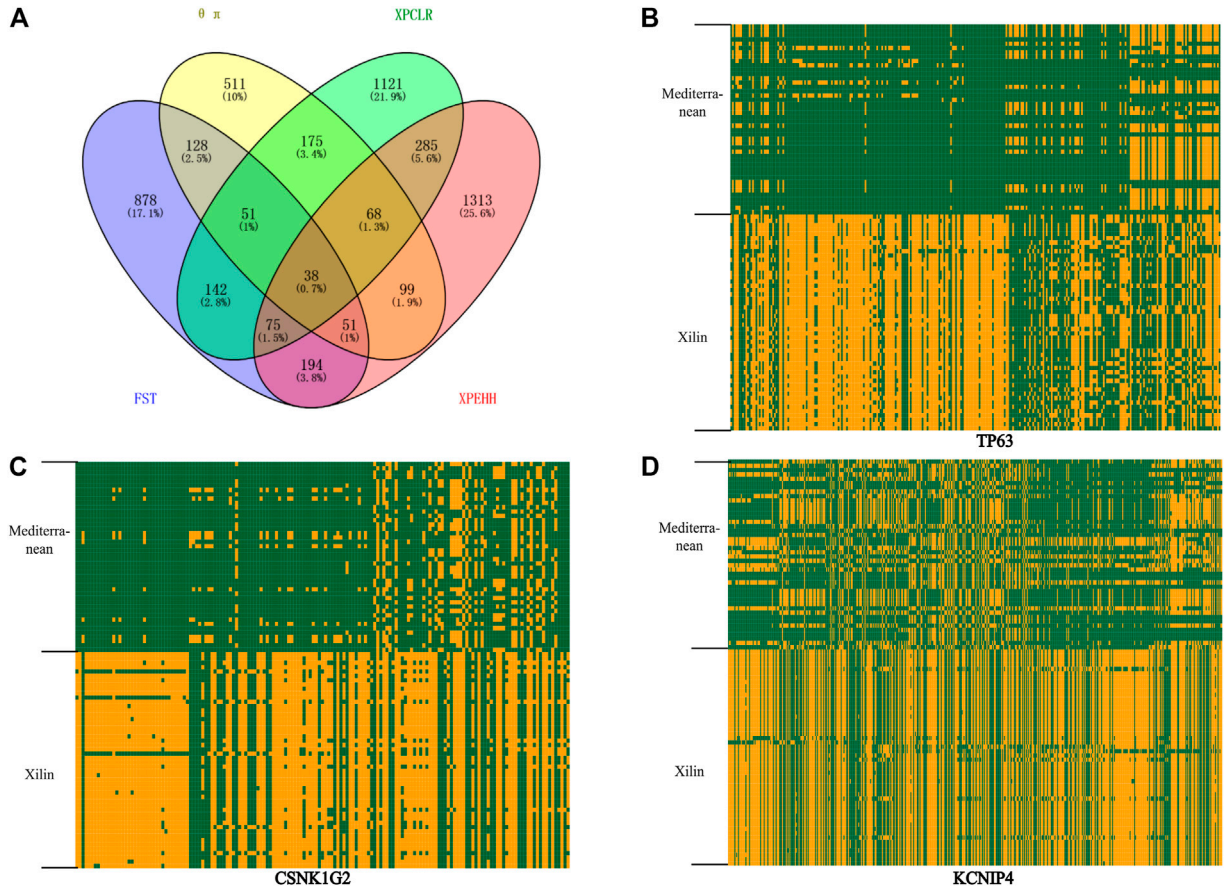


FIGURE 4
Analysis of the signatures of positive selection in the genome of Xilin Buffalo **(A)** Venn diagram showing the gene overlap among θπ, XPCLR, FST and XP-EHH. **(B)** Haplotype pattern heatmaps of the *TP63* gene region. **(C)** Haplotype pattern heatmaps of the *CSNK1G2* gene region. **(D)** Haplotype pattern heatmaps of the *KCNIP4* gene region. The major allele at each SNP position in Xilin buffalo is colored in yellow, the minor one in green.

We used ROH to evaluate the homozygosity of each individual. To evaluate the ROH patterns of Xilin buffalo and other buffalo breeds, we divided the length of ROH into three categories: .5–1 Mb, 1–2 Mb, and 2–4 Mb. A long ROH is the result of blood mating, while a shorter ROH reflects the influence of distant ancestors. The identified ROH length was mostly between .5 and 1 Mb (ROH diagram) (Figure 2A). The π map showed that the nucleotide diversity of the Xilin buffalo was the highest, followed by that of the Binhu buffalo, Fuzhong buffalo, Mediterranean buffalo, and Guizhou white buffalo (Figure 2B). The inbreeding degree of the inbreeding population is usually measured by the average inbreeding coefficient of the population. The inbreeding coefficient refers to the degree of gene purification expressed as a percentage according to the number of generations of inbreeding. According to the results in the figure, the average locus of Mediterranean buffalo is the highest, indicating that the population was first and most stable through artificial breeding, and the inbreeding coefficient of other breeds is close (Figure 2C). The whole genome average linkage disequilibrium (LD) of the Xilin buffalo is the lowest, and the LD value of the Binhu buffalo is the highest. Due to the different genetic backgrounds of different populations with the same population type and species, the decay rate of LD is also very different. Domestication selection will reduce the genetic diversity of the population and strengthen the correlation (linkage degree) between loci. Therefore, in general, the higher the degree of domestication, the greater the selection intensity, and the slowest rate of LD attenuation (El et al., 2021) (Figure 2D).

3.4 Functional enrichment analysis of specific SNP in Xilin buffalo

In this experiment, four methods (F_{ST}, π ratio, XP-CLR, XP-EHH) were used to detect the selection signal of Xilin buffalo by comparing the Xilin buffalo population with the Mediterranean buffalo population (Figure 3). Among the four methods, if a gene was significantly detected by at least two methods ($p < .005$), the gene was regarded as a real candidate gene.

A total of 113 genes were screened and many KEGG pathways and Gene Ontology (GO) related to nerves and exercise endurance were significantly enriched (corrected p-value<.05). The KEGG pathway is significantly related to the nervous system with the glutamatergic synapse. The GO enrichment analysis detected many nerves and muscle-related GO entries, including 'Nervous system development, GO:007399', 'Neuronal projection, GO:

0043005', 'actin binding, GO:003779', which reflect the nervous system and endurance played an extremely important role in the domestication and breeding of the Xilin buffalo.

3.5 Genome wide selective scanning test

Nucleotide diversity analysis (θ π) and complex likelihood ratio (CLR) were used to detect the selection-related genomic regions in the Xilin buffalo population. A total of 1121 (θ π) and 677 (CLR) (Figure 4A) genes were identified in Xilin buffalo with 357 overlapped. One of the most significant pathways (p-value < .05) was the Regulation of actin cytoskeleton which contained five genes (*CYFIP1, ITGA2B, ARHGEF4, CHRM3, CHRM2*) related to beef tenderness, feed efficiency and compensatory gain (Xia et al., 2021). Based on the gene ontology analysis of Xilin Buffalo, it is found that Xilin Buffalo has increased the GO category, including 'microtubule anchoring' (*MAST1, DAG1, PEX14*), 'Proteasome mediated ubiquitin-dependent protein catabolism' (*SMURF1, WWP1, PSME2, PSME1, DCAF11, HERC2, PSMC5*).

F_{ST} and XP-EHH tests were used to detect the positive selection characteristics of Xilin and riverine (Mediterranean) buffaloes Through the analysis, 1557 and 2123 hypothetical favorable positive selection genes were obtained from F_{ST} and XP-EHH methods, respectively, and 358 genes were obtained from both methods.

38 overlapping genes were detected in the above four selection methods which indicates that these genes have strong selection ability in the Xilin buffalo (Figure 4A). It is worth noting that (*CSNK1G2, TP63*), *CSNK1G2* is related to spermatogenesis, *MFG-E8* is a sign of high milk production in dairy animals, *TP63* participates in breast secretion by activating *MFG-E8*, and *RB1* is related to the formation of bovine intramuscular fat (marbling) (Lim et al., 2013).

4 Discussion

Understandings the characteristics of population structure and genetic diversity is very important for genetic evaluation, environmental adaptation, utilization, and protection of genetic resources of cattle breeds. In the present study, the whole genome sequences of 25 Xilin buffaloes were analysed. According to the geographical distribution, the buffalo are divided into six geographical regions: Upper Yangtze, middle lower Yangtze, Southwest China, Southeast Asia, South Asia, and Italy. Through ADMIXTURE analysis, we proved that the Xilin buffalo belongs to the Yangtze River.

The nucleotide diversity level of Xilin buffalo was slightly higher than the other breeds (average θπ = .0017). The relatively high genomic diversity of Xilin buffalo might be related to its weak and short selection history. The Xilin buffalo showed a similar structural heritability to Fuzhong Buffalo which is related to its similar geographical location and genetic background. In addition, the LD attenuation pattern of each variety was basically consistent with the results of nucleotide diversity. The ROH distribution pattern of Xilin buffalo was analyzed by comparing it with other cattle breeds. The ROH is common in bovine autosomes but the observed varietal differences in ROH length and burden patterns indicate differences in varietal origin and recent management. Compared with the cattle breeds analyzed in this study, Xilin buffalo showed more short/medium ROH (.5–2 Mb) and the average number of ROH was the highest.

By comparing with Mediterranean buffalo, we found that Xilin buffalo has unique signaling pathways in the nervous system, reproductive system, and lactation. In this study, the KEGG pathway and GO related to the nervous system were significantly enriched and the most significantly enriched pathway was the Glutamatergic synapse. In addition, GO analysis is also significantly enriched by many GO items related to the development of the nervous system such as neurons, dendritic spines, and synapses, and positive regulation of dendrite morphogenesis. Previous studies have proved that dendritic spines and their structural and functional plasticity are the cellular basis of learning and memory (Kasai et al., 2003). Therefore, it is speculated that these neural-related KEGG pathways and GO entries also play an extremely important role in the domestication of swamp buffalo. It has been reported that the glutamatergic synaptic pathway is related to the adaptability of mice to stress and fear behavior (Kamprath et al., 2010). It contains three genes *GRM5*, *GRIK2,* and *GRIA4,* and *GluR6* is encoded by *GRIK2* which is highly expressed in the brain and is associated with autosomal recessive intellectual disability (Motazacker et al., 2007). The *GRIK2* knockout mice showed decreased fear and memory, anxiety, and despair (Shaltiel et al., 2008). In rabbits, *GRIK2* was identified as a candidate domestication gene (Carneiro et al., 2014). The *GRIK2* is highly expressed in the brain tissues of buffalo, goats, sheep, and cattle. Studies have reported that *GRM5* is related to social interaction and sports behavior (Xu et al., 2021). The swamp buffalo has a gentle temperament and is mainly used for servitude. It can be easily trained for rice farming, cart pulling, and other labor (Chantalakhana and Bunyavejchewin, 1994). These traits indicate that the identified pathways and candidate genes related to the nervous system were strongly artificially selected during the domestication of swamp buffalo.

Reproductive performance is an important index to measure the economic benefit of a variety. The Xilin Buffalo has good reproductive performance and its oestrus cycle is between 20 and 25 days, with an average of 21.04 days. We found that both *CSNK1G2* and *KCNIP4* genes showed universal strong positive selection signals in Xilin buffalo/Mediterranean buffalo and both were related to reproduction. The *CSNK1G2* gene is related to sperm surface modification, sperm maturation, and sperm-egg communication of bull sperm (Byrne et al., 2012). The *CSNK1G2* gene is associated to the ability of frozen-thawed sperm to respond appropriately to stress (Pini et al., 2018). It has also been observed that *CSNK1G2* knockout mice show premature aging of the testes (Li et al., 2020) whereas the *KCNIP4* gene is closely related to chicken reproductive traits (Fan et al., 2017).

The Xilin buffalo was mainly used for both meat and milk. After crossbreeding, the introduced milk variety Mora buffalo over the years, the average lactation yield of the three generations has increased significantly (2389 ± 700.2 kg) and now its value is progressing because of better milk and meat production. In this study, we also found a selection signal related to lactation (*TP63*) and *MFG-E8* as a marker of high milk production in dairy animals. The *TP63* participates in breast secretion by activating *MFG-E8*. Previous studies have confirmed that *TP63* plays a role in regulating the growth and differentiation of mammary epithelial cells (Verma et al., 2021). By considering the influence of natural ecological

environmental conditions and local social and economic activities, these genes may play an important role in the reproduction and lactation performance of the Xilin buffalo after long-term natural and artificial selection.

5 Conclusion

Utilizing WGS data, the present study described the Xilin buffalo's whole genome level. The direction for the genetic assessment and coherent breeding plan of the Xilin buffalo was identified by examining the characteristics of population structure and genomic diversity. In addition, we also identified a series of candidate genes involved in milk production, neural control, and fertility. Moreover, the results of this study enable breeders to better understand the genomic characteristics of Xilin buffalo for artificial selection or adaptation to the local environment.

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/, PRJNA573503.

Ethics statement

The animal study was reviewed and approved by This study was conducted according to the Faculty Animal Policy and Welfare Committee of Northwest A&F University (FAPWC-NWAFU).

Author contributions

ZC contributed to the construction and execution of this manuscript. ZC, and MZ performed the experiments. QW contributed analysis tools. HL conceived and designed the experiments. CL and ZA revised the manuscript and provided suggestions. CL and JS provided the laboratories for statistical analysis and the funding for the research.

Funding

The work was supported by the Guangxi special project for innovationdriven development (AA17204024).

Acknowledgments

We thank the High-Performance Computing of Northwest A&F University (NWAFU) for providing computing resources.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1084824/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Amaral, M. E. J., Grant, J. R., Riggs, P. K., Stafuzza, N. B., Womack, J. E., Goldammer, T., et al. (2008). A first generation whole genome RH map of the river buffalo with comparison to domestic cattle. *Bmc Genomics* 9, 631. doi:10.1186/1471-2164-9-631

Byrne, K., Leahy, T., McCulloch, R., Colgrave, M. L., and Holland, M. K. (2012). Comprehensive mapping of the bull sperm surface proteome. *Proteomics* 12 (23-24), 3559–3579. doi:10.1002/pmic.201200133

Carneiro, M., Rubin, C. J., Di Palma, F., Albert, F. W., Alfoldi, J., Martinez, B. A., et al. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345 (6200), 1074–1079. doi:10.1126/science.1253714

Chantalakhana, C., and Bunyavejchewin, P. (1994). Buffaloes and draught power. *Outlook Agric.* 23 (2), 91–95. doi:10.1177/003072709402300204

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Di Stasio, L., and Brugiapaglia, A. (2021). Current knowledge on River buffalo meat: A critical analysis. *Anim. (Basel)* 11 (7), 2111. doi:10.3390/ani11072111

El, H. A., Rocha, D., Venot, E., Blanquet, V., and Philippe, R. (2021). Long-range linkage disequilibrium in French beef cattle breeds. *Genet. Sel. Evol.* 53 (1), 63. doi:10.1186/s12711-021-00657-8

Fan, Q. C., Wu, P. F., Dai, G. J., Zhang, G. X., Wang, J. Y., Xue, Q., et al. (2017). Identification of 19 loci for reproductive traits in a local Chinese chicken by genome-wide study. *Genet. Mol. Res. Gmr* 16 (1). doi:10.4238/gmr16019431

Fischer, H., and Ulbrich, F. (1967). Chromosomes of the Murrah buffalo and its crossbreds with the asiatic swamp buffalo (Bubalus bubalis). *J. Animal Breed. Genet.* 84 (1-4), 110–114. doi:10.1111/j.1439-0388.1967.tb01102.x

Forutan, M., Ansari, M. S., Baes, C., Melzer, N., Schenkel, F. S., and Sargolzaei, M. (2018). Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genomics* 19 (1), 98. doi:10.1186/s12864-018-4453-z

Green, M. R., and Sambrook, J. (2012). "Molecular cloning: A laboratory manual," in *Three-volume set*. Fourth Edition (United States: Cold Spring Harbor Laboratory Pr).

He, L., Su, J., Zhang, L., Deng, S., He, Y., Huang, X., et al. (2011). Survey on national local breeds - Fuzhong and Xilin buffalo. *Guangxi Animal Husb. Veterinary Med.* 27 (6), 3. (In Chinese).

Hudson, R. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* 132, 583. doi:10.1093/genetics/132.2.583

Kamprath, K., Plendl, W., Marsicano, G., Deussing, J. M., Wurst, W., Lutz, B., et al. (2010). Endocannabinoids mediate acute fear adaptation via glutamatergic neurons independently of corticotropin-releasing hormone signaling. *Genes Brain & Behav.* 8 (2), 203–211. doi:10.1111/j.1601-183X.2008.00463.x

Kasai, H., Matsuzaki, M., Noguchi, J., Yasumatsu, N., and Nakahara, H. (2003). Structure–stability–function relationships of dendritic spines. *Trends Neurosci.* 26 (7), 360–368. doi:10.1016/S0166-2236(03)00162-0

Li, D., Ai, Y., Guo, J., Dong, B., and Wang, X. (2020). Casein kinase 1G2 suppresses necroptosis-promoted testis aging by inhibiting receptor-interacting kinase 3. *eLife Sci.* 9, e61564. doi:10.7554/eLife.61564

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

Li, X., Yang, J., Shen, M., Xie, X. L., Liu, G. J., Xu, Y. X., et al. (2020). Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. *Nat. Commun.* 11 (1), 2815. doi:10.1038/s41467-020-16485-1

Lim, D., Lee, S. H., Kim, N. K., Cho, Y. M., Kim, H., Seong, H. H., et al. (2013). Gene Co-expression analysis to characterize genes related to marbling trait in hanwoo (Korean) cattle. *Asian-australasian J. Animal Sci.* 26 (1), 19–29. doi:10.5713/ajas.2012.12375

Liu, J. J., Liang, A. X., Campanile, G., Plastow, G., Zhang, C., Wang, Z., et al. (2018). Genome-wide association studies to identify quantitative trait loci affecting milk production traits in water buffalo. *J. Dairy Sci.* 101 (1), 433–444. doi:10.3168/jds.2017-13246

Makanjuola, B. O., Maltecca, C., Miglior, F., Marras, G., Abdalla, E. A., Schenkel, F. S., et al. (2021). Identification of unique ROH regions with unfavorable effects on production and fertility traits in Canadian Holsteins. *Genet. Sel. Evol.* 53 (1), 68. doi:10.1186/s12711-021-00660-z

Michelizzi, V. N., Dodson, M. V., Pan, Z., Amaral, M. E. J., Michal, J. J., Mclean, D. J., et al. (2010). Water buffalo genome science comes of age. *Int. J. Biol. Sci.* 6 (4), 333–349. doi:10.7150/ijbs.6.333

Motazacker, M. M., Rost, B. R., Hucho, T., Garshasbi, M., Kahrizi, K., Ullmann, R., et al. (2007). A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. *Am. J. Hum. Genet.* 81 (4), 792–798. doi:10.1086/521275

Nekrutenko, A., and Taylor, J. (2012). Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. *Nat. Rev. Genet.* 13 (9), 667–672. doi:10.1038/nrg3305

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15 (11), 1566–1575. doi:10.1101/gr.4252305

Patterson, N., Price, A. L., and Reich, D. (2013). Population structure and eigenanalysis. *Plos Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Pini, T., Rickard, J. P., Leahy, T., Crossett, B., Graaf, S., and de Graaf, S. P. (2018). Cryopreservation and egg yolk medium alter the proteome of ram spermatozoa. *J. Proteomics* 181, 73–82. doi:10.1016/j.jprot.2018.04.001

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Rahimmadar, S., Ghaffari, M., Mokhber, M., and Williams, J. L. (2021). Linkage disequilibrium and effective population size of buffalo populations of Iran, Turkey, Pakistan, and Egypt using a medium density SNP array. *Front. Genet.* 12, 12608186. doi:10.3389/fgene.2021.608186

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 449913–449918. doi:10.1038/nature06250

Shaltiel, G., Maeng, S., Malkesman, O., Pearson, B., Schloesser, R. J., Tragon, T., et al. (2008). Evidence for the involvement of the kainate receptor subunit GluR6 (GRIK2) in mediating behavioral displays related to behavioral symptoms of mania. *Mol. Psychiatry* 13 (9), 858–872. doi:10.1038/mp.2008.20

Shen, S., Kong, J., Qiu, Y., Yang, X., Wang, W., and Yan, L. (2019). Identification of core genes and outcomes in hepatocellular carcinoma by bioinformatics analysis. *J. Cell Biochem.* 120 (6), 10069–10081. doi:10.1002/jcb.28290

Stothard, P., Choi, J. W., Basu, U., Sumner-Thomson, J. M., Meng, Y., Liao, X., et al. (2011). Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics* 12, 559. doi:10.1186/1471-2164-12-559

Sun, T., Shen, J., Achilli, A., Chen, N., Chen, Q., Dang, R., et al. (2020). Genomic analyses reveal distinct genetic architectures and selective pressures in buffaloes. *Gigascience* 9 (2), giz166. doi:10.1093/gigascience/giz166

Szpiech, Z. A., and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31 (10), 2824–2827. doi:10.1093/molbev/msu211

Verma, A. K., Ali, S. A., Singh, P., Kumar, S., and Mohanty, A. K. (2021). Transcriptional repression of MFG-E8 causes disturbance in the homeostasis of cell cycle through DOCK/ZP4/STAT signaling in buffalo mammary epithelial cells. *Front. Cell Dev. Biol.* 9, 568660. doi:10.3389/fcell.2021.568660

Xia, X., Zhang, S., Zhang, H., Zhang, Z., Chen, N., Li, Z., et al. (2021). Assessing genomic diversity and signatures of selection in Jiaxian Red cattle using whole-genome sequencing data. *BMC Genomics* 22 (1), 43. doi:10.1186/s12864-020-07340-0

Xu, J., Marshall, J. J., Kraniotis, S., Nomura, T., Zhu, Y., and Contractor, A. (2021). Genetic disruption of Grm5 causes complex alterations in motor activity, anxiety and social behaviors. *Behav. Brain Res.* 411, 411113378. doi:10.1016/j.bbr.2021.113378

Yu, Z., Zhang, W., Gu, C., Chen, J., Zhao, M., Fu, L., et al. (2021). Genomic analysis of Ranavirus and exploring alternative genes for phylogenetics. *Transbound. Emerg. Dis.* 68 (4), 2161–2170. doi:10.1111/tbed.13864

Zhang, Y., Sun, D., and Yu, Y. (2007). Genetic diversity and differentiation of Chinese domestic buffalo based on 30 microsatellite markers. *Anim. Genet.* 38, 569. doi:10.1111/j.1365-2052.2007.01648.x

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fgene.2022.1034094

OPEN ACCESS

EDITED BY
Zexi Cai,
Aarhus University, Denmark

REVIEWED BY
Ran Li,
Northwest A&F University, China
Guangxin E,
Southwest University, China
Zhixin Chai,
Southwest Minzu University, China

*CORRESPONDENCE
Zhijie Ma,
zhijiema@126.com

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

RECEIVED 01 September 2022
ACCEPTED 02 November 2022
PUBLISHED 10 January 2023

CITATION
Li G, Luo J, Wang F, Xu D, Ahmed Z, Chen S, Li R and Ma Z (2023), Whole-genome resequencing reveals genetic diversity, differentiation, and selection signatures of yak breeds/populations in Qinghai, China. *Front. Genet.* 13:1034094.
doi: 10.3389/fgene.2022.1034094

Whole-genome resequencing reveals genetic diversity, differentiation, and selection signatures of yak breeds/populations in Qinghai, China

Guangzhen Li[1,2,3], Jing Luo[1,2,3], Fuwen Wang[4], Donghui Xu[1,2,3], Zulfiqar Ahmed[5], Shengmei Chen[1,2,3], Ruizhe Li[1,2,3] and Zhijie Ma[1,2,3]*

[1]Academy of Animal Science and Veterinary Medicine, Qinghai University, Xining, China, [2]Key Laboratory of Animal Genetics and Breeding on Tibetan Plateau, Ministry of Agriculture and Rural Affairs, Xining, China, [3]Plateau Livestock Genetic Resources Protection and Innovative Utilization Key Laboratory of Qinghai Province, Xining, China, [4]College of Animal Science and Technology, Northwest A&F University, Xianyang, China, [5]Faculty of Veterinary and Animal Sciences, University of Poonch Rawalakot, Rawalakot, Pakistan

The Qinghai Province of China is located in the northeast region of the Qinghai–Tibetan Plateau (QTP) and carries abundant yak genetic resources. Previous investigations of archaeological records, mitochondrial DNA, and Y chromosomal markers have suggested that Qinghai was the major center of yak domestication. In the present study, we examined the genomic diversity, differentiation, and selection signatures of 113 Qinghai yak, including 42 newly sequenced Qinghai yak and 71 publicly available individuals, from nine yak breeds/populations (wild, Datong, Huanhu, Xueduo, Yushu, Qilian, Geermu, Tongde, and Huzhu white) using high-depth whole-genome resequencing data. We observed that most of Qinghai yak breeds/populations have abundant genomic diversity based on four genomic parameters (nucleotide diversity, inbreeding coefficients, linkage disequilibrium decay, and runs of homozygosity). Population genetic structure analysis showed that Qinghai yak have two lineages with two ancestral origins and that nine yak breeds/populations are clustered into three distinct groups of wild yak, Geermu yak, and seven other domestic yak breeds/populations. F_{ST} values showed moderate genetic differentiation between wild yak, Geermu yak, and the other Qinghai yak breeds/populations. Positive selection signals were detected in candidate genes associated with disease resistance (*CDK2AP2*, *PLEC*, and *CYB5B*), heat stress (*NFAT5*, *HSF1*, and *SLC25A48*), pigmentation (*MCAM*, *RNF26*, and *BOP1*), vision (*C1QTNF5*, *MFRP*, and *TAX1BP3*), milk quality (*OPLAH* and *GRINA*), neurodevelopment (*SUSD4*, *INSYN1*, and *PPP1CA*), and meat quality (*ZRANB1*), using the integrated PI, composite likelihood ratio (CLR), and F_{ST} methods. These findings offer new insights into the genetic mechanisms underlying target traits in yak and provide important information for understanding the genomic characteristics of yak breeds/populations in Qinghai.

KEYWORDS

***Bos grunniens*, whole-genome resequencing, genomic diversity, population structure, selection signature**

Introduction

The yak is a large, unique ungulate that lives in the climatically challenging conditions (limited oxygen, extreme cold, highly variable daytime and nighttime temperatures, and scanty flora) of the Qinghai–Tibetan Plateau (QTP) and nearby high-altitude regions (Wiener et al., 2003; Zhang et al., 2020). Domestic yak (*Bos grunniens*) is descended from wild yak (*Bos mutus*) (Qiu et al., 2015) and is an indispensable part of the Tibetan culture, providing basic resources such as meat, milk, transportation, fuels, and hides to Tibetans and other nomadic peoples living in high-altitude environments (Wiener et al., 2003; Jia et al., 2019; Jia et al., 2020). The Qinghai Province of China, located in the northeast region of the QTP, possesses a variety of yak genetic resources and is regarded as a major center of yak domestication (Guo et al., 2006; Ma, 2019). Due to its special geographical location, complex plateau climate, and long breeding history, Qinghai is home to some unique yak populations/breeds. For instance, two improved breeds (Datong and Ashdan) and four indigenous breeds (Qinghai Plateau, Huanhu, Xueduo, and Yushu) are currently recognized in this region (National Committee of Animal Genetic Resources, 2021).

According to our previous reports on genetic variations in Y-chromosomal markers, both wild and domestic yak in Qinghai have relatively high paternal genetic diversity with weak phylogeographic structures and two paternal origins (Ma et al., 2018; Ma et al., 2022). In addition, maternal genetic diversity of the wild and domestic yak in Qinghai indicates that wild and domestic yak have high levels of genetic diversity and can be clustered into three lineages (Wang et al., 2021; Li et al., 2022). Since a female domestic yak genome was first assembled in 2012, the domestic yak reference genome has been improved twice at the chromosomal level (Qiu et al., 2012; Ji et al., 2021; Zhang et al., 2021). Obviously, the completeness and accuracy of the latest yak reference genome (BosGru3.0, GCA_005887515.2) are significantly higher than those of the previously reported genomes (Zhang et al., 2021). This information has laid a strong foundation for further exploration of the genomic diversity, population structure, and phylogenetic relationships of yak breeds/populations at the genome level. Also, the availability of the high-quality yak reference genome, which was built using long-read sequencing technology (Zhang et al., 2021), has enabled the identification of the genetic basis of complex traits. Following the wide application of whole-genome sequencing (WGS), the genomic diversity, origin, domestication, population structure, coat color, and high-altitude adaptation of yak have become research hotspots (Qiu et al., 2012; Wang et al., 2014; Qiu et al., 2015; Liang et al., 2016; Zhang et al., 2016; Medugorac et al., 2017; Lan et al., 2018; Xie et al., 2018; E et al., 2019a; 2019b; Wang et al., 2019; Chai et al., 2020; Lan et al., 2021; Bao et al., 2022; Gao et al., 2022). Although a few Qinghai yak breeds/populations, including Qilian, Datong, Qinghai Plateau, and Huanhu, had been previously analyzed using WGS, this strategy could not comprehensively reveal genomic information on Qinghai yak breeds/populations because of its low-depth WGS data (average sequencing depth around ×7.2) and small number of samples (Chai et al., 2020). Additionally, the distribution area of some yak breeds such as Qinghai Plateau and Yushu are adjacent to the habitats of wild yak. Hybridization between domestic and wild yak is ubiquitous. Therefore, it is particularly necessary to explore the relationship between Qinghai domestic yak breeds/populations and wild yak using WGS data.

It is essential to thoroughly ascertain the genomic diversity, population structure, and selection signature of Qinghai yak breeds/populations to increase their potential breeding value. In this study, we performed high-depth WGS analysis of 113 yak, including 42 newly sequenced yak from nine Qinghai yak breeds/populations, to identify single-nucleotide polymorphisms (SNPs) based on the latest yak reference genome (BosGru3.0, GCA_005887515.2) and to reveal their genomic diversity and population structure, as well as candidate signatures of positive selection. The gathered information provides a baseline for the exploration and evaluation of yak breeds/populations in Qinghai, China. Moreover, it contributes to the conservation and utilization of these precious yak genetic resources.

Materials and methods

Ethical approval

This study was conducted following animal welfare requirements. The procedures approved for experiments were based on the recommendations of the Regulations for the Administration of Affairs Concerning Experimental Animals of China. The Institutional Animal Care and Use Committee of the Academy of Animal Science and Veterinary Medicine, Qinghai University, approved all the animal experiments in this study.

Sample collection and whole-genome resequencing

Based on our previous report on Y-chromosome variations in Qinghai yak (Ma et al., 2018), 42 individuals (NCBI: PRJNA827919) representing breed/population-specific haplotypes were selected for resequencing in this study. Peripheral blood samples were collected from the primary producing area of each Qinghai yak breed/population (Figure 1A; Supplementary Table S1), including Datong (DT, n = 6), Yushu (YS, n = 6), Xueduo (XD, n = 5), Huanhu (HH, n = 5), Geermu (GEM, n = 5), Qilian (QL, n = 5), Tongde (TD, n = 7), and Huzhu white (HZ, n = 3). Genomic DNA was extracted using the Whole Blood DNA Extraction Kit (Aidlab Biotechnologies Co., Ltd, China) (Supplementary Table S2). Pair-end libraries were constructed for each individual (500 bp insert size), and the quantified DNA was subjected to the Illumina Nova 6000 sequencing platform using a 2 × 150 bp model at the Novogene Bioinformatics Institute (Beijing, China) (Supplementary Table S3). To systematically explore the genomic diversity, population structure, and selection signature of Qinghai yak breeds/populations, we conducted a comprehensive search for whole Qinghai yak genome sequences in the NCBI database, resulting in 71 whole-genome sequences of six Qinghai yak breeds/populations for combined analysis, including wild (WY, n = 19), Datong (DT, n = 23), Yushu (YS, n = 16), Huanhu (HH, n = 7), Geermu (GEM, n = 3), and Qilian (QL, n = 3). In total, this study analyzed 113 whole-genome sequences from nine Qinghai yak breeds/populations (Supplementary Table S1).

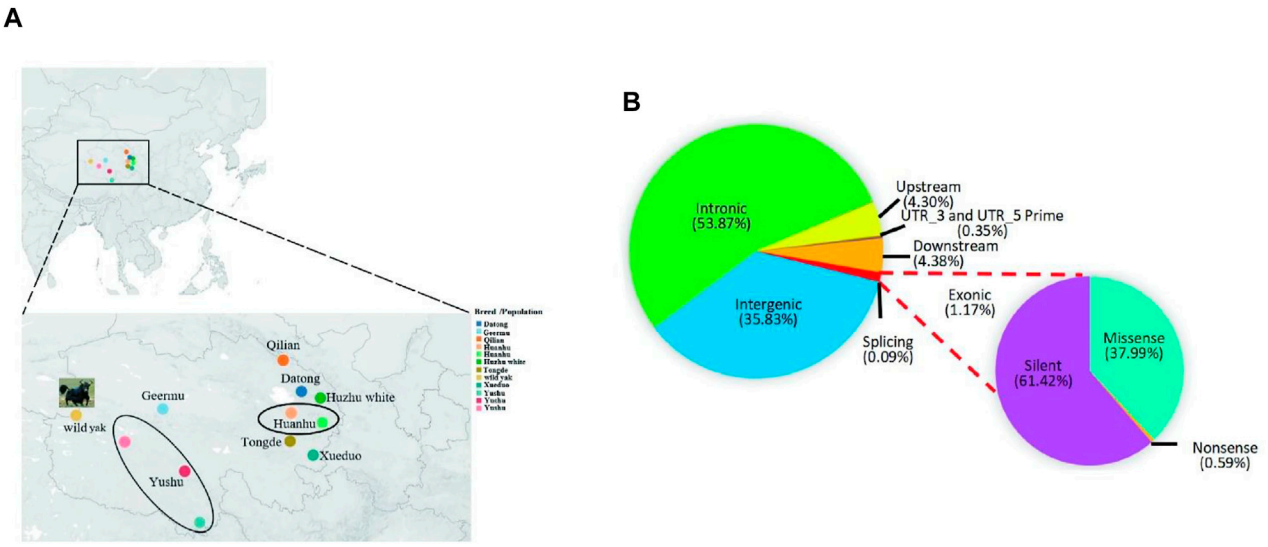


FIGURE 1
Distribution of samples of Qinghai yak breeds/populations and SNPs in yak genome. **(A)** Sampling sites and geographic distributions of nine Qinghai yak breeds/populations. **(B)** Distribution of detected SNPs in the genomes of nine yak breeds/populations. The circles represent the sampling ranges of Huanhu and Yushu yak.

Read mapping and single-nucleotide polymorphism calling

The raw reads were trimmed using Trimmomatic (sliding window: 3:15 minlen:35 trailing:20 leading:20 avgqual: 20 tophred33) (Bolger et al., 2014) to remove adapters and low-quality bases. The clean reads were aligned to the latest yak reference genome (BosGru3.0, GCA_005887515.2) using Burrows–Wheeler Aligner (v0.7.13-r1126) software (Li and Durbin, 2009) with default parameters. Picard tools (http://broadinstitute.github.io/picard) were used to filter potential duplicate reads. We used the "Haplotype Caller," "Genotype GVCFs," and "Select Variants" modules of the Genome Analysis Toolkit (GATK, v3.8-1-0-gf15c1c3ef) (McKenna et al., 2010) to call the SNPs. After SNP calling, we used the "Variant Filtration" module of GATK to obtain high-quality SNPs with the parameters ("DP <249 (1/3-fold total sequence depth for all individuals) || DP >2245 (three-fold of total sequence depth for all individuals) || QD <2.0 || FS >60.0 || MQ <40.0 || MQRankSum <−12.5 || ReadPosRankSum <−8.0 || SOR >3.0"). Finally, based on the yak reference genome (BosGru3.0, GCA_005887515.2), SNPs were functionally annotated by ANNOVAR software (Wang et al., 2010). In addition, the SNP densities of each Qinghai yak breed/population were calculated by VCFtools (v0.1.12) (window 100,000) software (Danecek et al., 2011).

Population genomic parameters

To reveal the genomic variations of Qinghai yak, we calculated the genomic parameters, including nucleotide diversity (Pi), expected heterozygosity (He), observed heterozygosity (Ho), inbreeding coefficient (F_{Hom}), linkage disequilibrium (LD) decay, and runs of homozygosity (ROH) for nine Qinghai yak breeds/populations (Datong, Geermu, Huanhu, Huzhu white, Qilian, Tongde, Xueduo, Yushu, and wild yak).

We estimated the genomic nucleotide diversity of each yak breed/population using VCFtools (v0.1.12) software (-window-pi 50,000 -window-pi-step 20,000). The output of the -het function by VCF tools was a summary for each individual of the observed number of homozygous sites [O(hom)] and the expected number of homozygous sites [E(hom)]. It also included the total number of sites for which the individual had data and the inbreeding coefficient (F_{Hom}), which was the canonical estimate of genomic F_{Hom} based on excess SNP homozygosity (Keller et al., 2011). The SNPs of 113 genomes of nine Qinghai yak breeds/populations were pruned at high levels of pairwise LD by PLINK (v1.9) software (Purcell et al., 2007), excluding SNPs in strong LD ($r^2 > 0.2$) within a sliding window of 50 SNPs advanced by five SNPs at a time (Zhang et al., 2019).

The ROHs were calculated by PLINK (v1.9) software. SNPs with minor allele frequencies (MAF < 0.05) were excluded due to instability. PLINK (v1.9) software was used for sliding windows of a minimum of 50 SNPs across the genomes to identify ROHs, allowing for two missing SNPs and one heterozygous site per window. The minimum number of continuous homozygous SNPs constituting an ROH was set to 100. The minimum SNP density coverage was set to at least 50 SNPs per kb, allowing centromeric and SNP-poor regions to be algorithmically excluded from the analysis. The maximum gap between two consecutive homozygous SNPs was set at 100 kb. The number and length of ROHs for each yak breed/population were estimated, and the length of ROHs was divided into three categories: 0.5–1 Mb, 1–2 Mb, and > 2 Mb, reflecting ancient, historical, and recent inbreeding, respectively (Kirin et al., 2010; Bhati et al., 2020).

Population genetic structure and clustering pattern

VCFtools (v0.1.12) software was used to convert the VCF files into the PLINK format. Linkage sites in the genomic data were removed with parameters (-indep-pair-wise 50 5 0.2) using PLINK (v1.9) software, and then the filtered data were used for principal component analysis (PCA) and admixture analysis. The population genetic structure was estimated using ADMIXTURE (v1.3.0) (Alexander and Lange, 2011), considering 2–4 clusters (*K*), and the results were visualized *via* R (v3.6.1) software. Based on the pairwise distance matrix among individuals, a neighbor-joining (NJ) tree was constructed by MEGA (v10.2.6) (Saitou and Nei, 1987; Kumar et al., 2016). The PCA of 113 individuals was performed by smartPCA in the EIGENSOFT (v5.0) package (Patterson et al., 2006) to estimate the eigenvectors. The Tracy–Widom distribution was used to assess the significance of each principal component, and the results of the first and second principal components were plotted using the ggplot2 package in R (v3.6.1) software. VCFtools (Danecek and McCarthy, 2017) software was used to calculate the fixation index (F_{ST}) between nine Qinghai yak breeds/populations, and the results were visualized *via* ImageGP (http://www.ehbio.com/ImageGP/). To further explore the relationships among nine Qinghai yak breeds/populations, we calculated the linearized R_{ST} values ($R_{ST} = F_{ST}/(1- F_{ST})$). Based on the R_{ST} values, the cluster pattern among nine Qinghai yak breeds/populations was revealed by multidimensional scale (MDS) analysis using SPSS (v18.0) software.

Genome-wide selective sweep

Only SNPs with less than 10% missing nucleotides were used for selective sweep scanning. To explore the selection pressure of domestic and wild yak in Qinghai, we performed a detection of positive selection signature. In this study, three methods, including composite likelihood ratio (CLR), nucleotide diversity (PI), and pairwise fixation index (F_{ST}), were used to detect the selection signatures in Qinghai yak. The CLR was computed using SWEEPFINDER2 software to detect SNPs within a non-overlapping 50 kb window (Nielsen et al., 2005; DeGiorgio et al., 2016). The PI was estimated using VCFtools (50 kb sliding window and 20 kb step) in Qinghai domestic yak. Consistent with previous methods, the top 1% windows were selected as the candidate region under selection (Chen et al., 2018; Xia et al., 2021). To find candidate genes related to adaptation, immunity, and economic traits, the F_{ST} was calculated between wild and domestic yak with a 50-kb sliding window and 20-kb steps along the autosomes using the VCFtools and R scripts (Danecek et al., 2011; Chen et al., 2018; Chen et al., 2020; Xia et al., 2021). The Tajima'D values were calculated for the candidate genes using VCFtools. To gain a further understanding of the gene functions and signaling pathways of the identified candidate genes, online Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) analyses were conducted using KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/anno_iden.php), and FDR < 0.05 was used as a threshold to detect significantly enriched genes and pathways.

Results

Whole-genome resequencing and single nucleotide polymorphisms

The total raw and clean bases of 42 newly sequenced Qinghai yak were 3,974.07 Gb and 3,942.42 Gb, respectively (Supplementary Table S3). The average sequencing depth of the reads in 42 individual genomes was around ×23.91 (Supplementary Table S1). In total, we obtained 112,960,043 high-quality SNPs in nine yak breeds/populations. The highest number of SNPs (18,061,823) was detected in Datong, followed by Yushu (14,833,892), wild (14,109,444), Huanhu (12,805,435), Tongde (12,059,879), Qilian (11,462,122), Xueduo (10,394,482), and Geermu (10,266,014) yak, while the lowest SNPs were detected in Huzhu white yak (8,966,952) (Supplementary Table S4). Comparison among 113 yak genomes revealed that, among 25,768,146 autosomal SNPs, 35.83% were mapped to intergenic regions, 53.87% to intronic regions, and only 1.17% to exonic regions. The functional annotation of the SNPs assigned to protein-coding regions identified about 61.42% of SNPs to produce silent mutations, 37.99% to cause missense mutations, and 0.59% to result in nonsense mutations (Figure 1B; Supplementary Table S4).

Population genomic diversity, runs of homozygosity, and linkage disequilibrium

Nucleotide diversity (Pi) ranged from 0.0006 to 0.0015 among nine Qinghai yak breeds/populations (Figure 2A; Table 1), with the highest Pi value (0.0015) in Huzhu white yak and the lowest (0.0006) in Geermu yak (Figure 2A; Table 1). The values of He and Ho ranged from 0.2699 to 0.4752 and from 0.1443 to 0.3454 for Datong and Huzhu white yak, from lowest to highest, respectively (Table 1). LD analysis showed that the wild yak presented a very rapid decay rate and the lowest level of LD (Figure 2B). However, in the domestic yak breeds/populations, the Huzhu white yak exhibited a slow decay rate and a high level of LD, followed by Xueduo, Tongde, Yushu, Qilian, Huanhu, and Datong yak, whereas the Geermu yak showed a rapid decay rate and a low level of LD. Among nine Qinghai yak breeds/populations, the wild yak had the highest value of F_{Hom} (0.2841), followed by Qilian, Yushu, Geermu, Huanhu, Datong, Xueduo, Tongde, and Huzhu white yak (Table 1; Figure 2C). We found that the ROH lengths in all yak breeds/populations were mostly between 0.5 and 2 Mb (Figure 2D; Supplementary Table S5). Larger ROHs (>2 Mb) were only identified in Huanhu (2) and Xueduo (1) yak, while medium ROHs (1–2 Mb) were found in almost all the yak breeds/populations, except for the wild yak (i.e., Datong (22), Geermu (1), Huanhu (34), Huzhu white (26), Qilian (9), Tongde (14), Xueduo (21), and Yushu (30)) (Figure 2D; Supplementary Table S5).

Population genetic structure

To determine the population genetic structure and relationships among nine different yak breeds/populations, we conducted a series of analyses, including phylogenetic reconstructions, PCA, and Bayesian clustering, using WGS data. In a phylogenetic analysis, the NJ tree

10.3389/fgene.2022.1034094

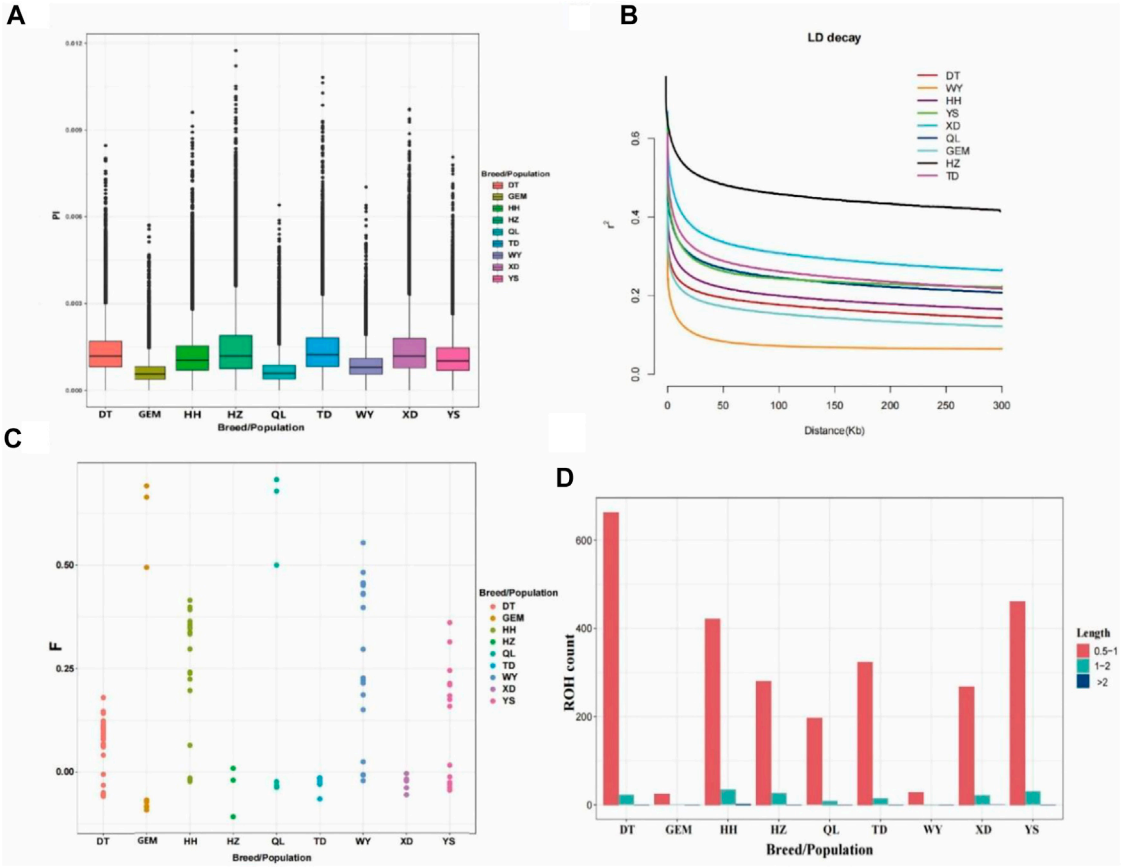


FIGURE 2
Analyses of genomic variation, LD, F_{Hom}, and ROHs for Qinghai yak breeds/populations. **(A)** Box plots of nucleotide diversity (Pi) for each yak breed/population. **(B)** Decay of LD in nine yak breeds/populations, with one line per breed/population. **(C)** F_{Hom} in nine yak breeds/populations. **(D)** Length and average number of ROHs in nine yak breeds/populations.

TABLE 1 Genetic diversity indexes of nine yak breeds/populations in Qinghai, China.

Breed/population	No. of sample/sequence	Pi	Ho	He	Fhom
Datong yak	29	0.0013	0.1443	0.2699	0.0687
Geermu yak	8	0.0006	0.2677	0.2835	0.1817
Huanhu yak	12	0.0012	0.2283	0.3879	0.1747
Huzhu white yak	3	0.0015	0.3454	0.4752	−0.0394
Qilian yak	8	0.0007	0.2429	0.4461	0.2153
Tongde yak	7	0.0014	0.2457	0.2881	−0.0271
Xueduo yak	5	0.0014	0.2932	0.3747	−0.0269
Yushu yak	22	0.0011	0.1892	0.2961	0.2039
Wild yak	19	0.0009	0.2017	0.2798	0.2841

Note: Pi, nucleotide diversity; Ho, observed heterozygosity; He, expected heterozygosity; Fhom, excess of homozygosity inbreeding coefficient.

showed different clades for wild yak and eight domestic yak breeds/populations in Qinghai and inferred two lineages (lineages I and II) (Figure 3A) and three sub-lineages (A, B, and C) in Lineage I.

According to the PCA results, the first principal component (PC1) explained 3.07% of the total variation separating wild yak and Geermu yak from the other seven yak breeds/populations (Figure 3B). The second PC (PC2) explained 2.51% of the total variation, indicating a more complex genetic makeup in wild yak than in other Qinghai domestic yak breeds/populations. In addition, a further PCA analysis excluding the wild and Geermu

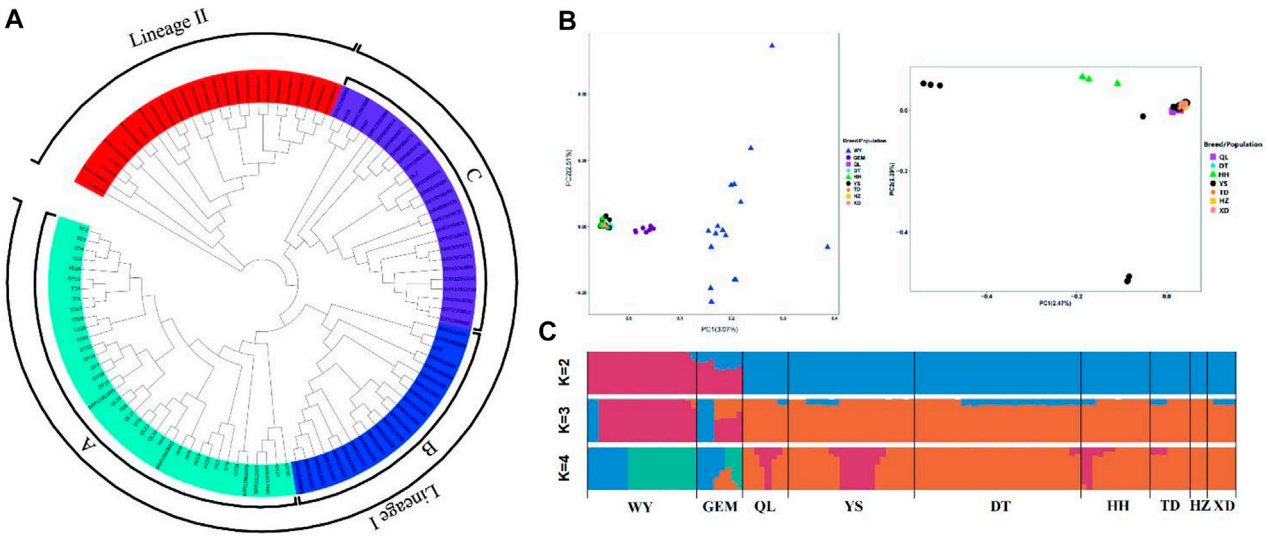


FIGURE 3
Population genetic analyses of Qinghai yak breeds/populations. **(A)** NJ tree of 113 wild and domestic yak. **(B)** PCA with first (PC1) and second (PC2) principal components. The right figure shows a further PCA excluding wild and Geermu yak populations. **(C)** Model-based clustering of Qinghai yak breeds/populations using ADMIXTURE with K = 2–4.

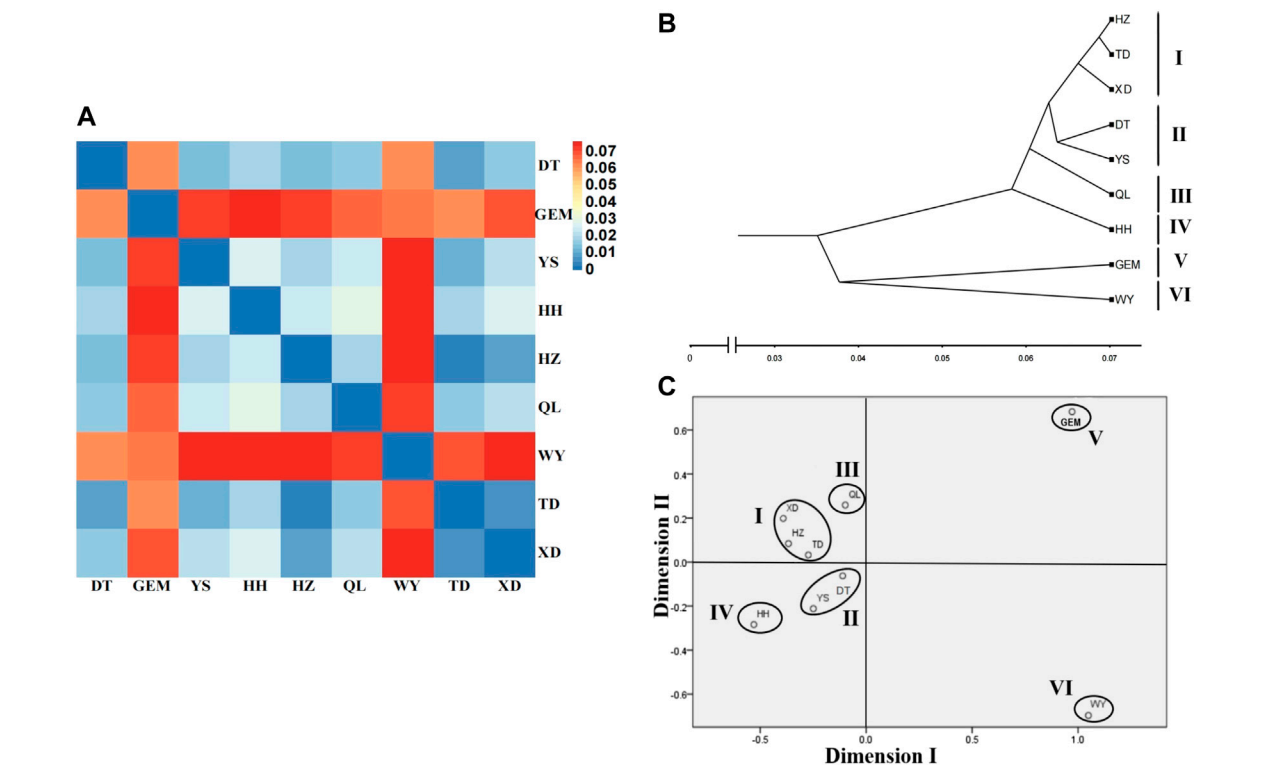


FIGURE 4
Genetic relationships among Qinghai yak breeds/populations. **(A)** Heat map of genetic differentiation among nine yak breeds/populations. **(B)** Cluster relationship among nine yak breeds/populations using the UPGMA tree. **(C)** MDS analysis among nine yak breeds/populations.

yak populations showed that some individuals from Yushu and Huanhu yak were divided (Figure 3B). To further elucidate the population genetic structure of Qinghai yak breeds/populations, we estimated the ancestral components for 113 individuals *via* ADMIXTURE software through clustering models (Figure 3C). In clustering analysis, when K = 2, wild and Qinghai domestic yak

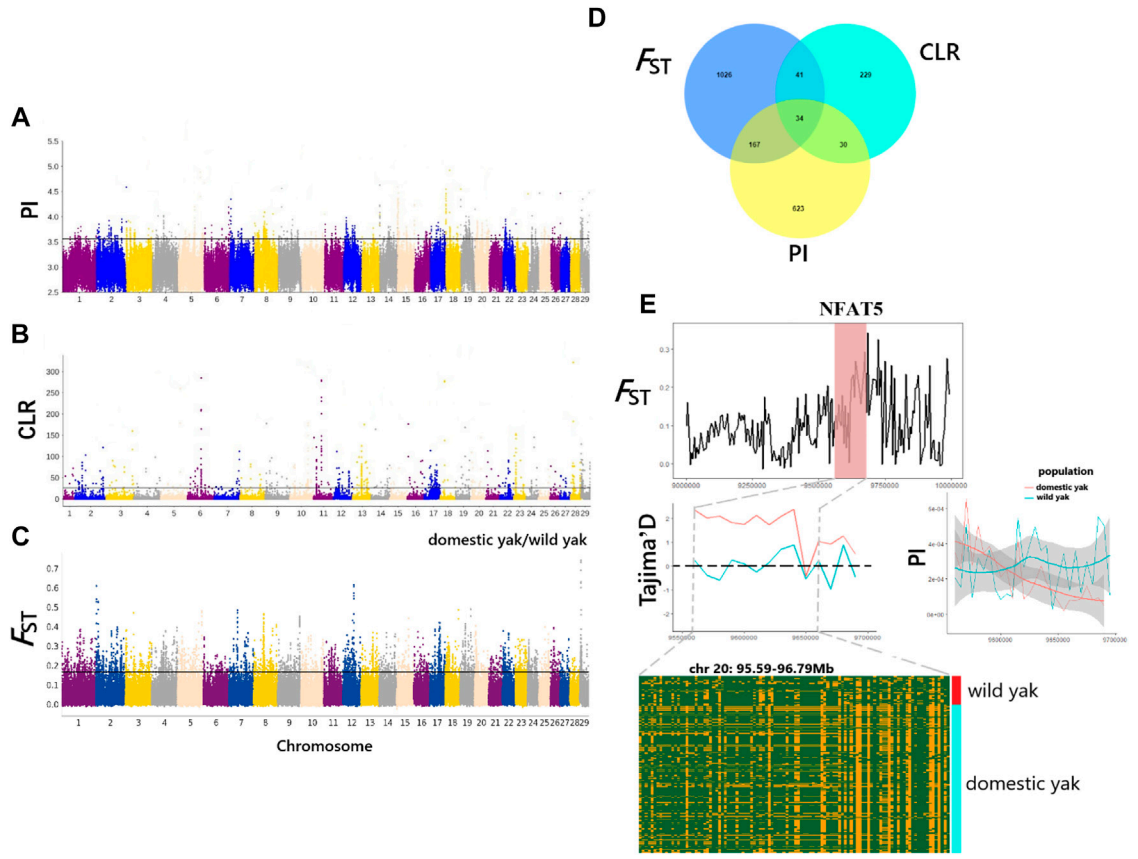


FIGURE 5
Genome-wide selection scan for positively selected genes (PSGs) in Qinghai yak using sliding window analysis (50-kb window size and 20-kb step size). **(A)** Manhattan plot of selected sweeps by the PI method. **(B)** Manhattan plot of selected sweeps by the CLR method. **(C)** Manhattan plot of selected sweeps by the F_{ST} method, based on domestic and wild yak populations. **(D)** Venn diagram showing the candidate genes identified by the CLR, PI, and F_{ST} methods. **(E)** F_{ST}, Tajima'D, and PI of the *NFAT5* gene region between wild and domestic yak.

breeds/populations had two ancestral components, namely, wild yak and domestic yak, with only the Geermu yak presenting two ancestral components. When $K = 3$, a new ancestry was observed in almost all Qinghai yak breeds/populations except Huzhu white yak. Notably, the wild yak and Geermu yak populations accounted for a relatively high proportion of this new ancestry, but the other six Qinghai domestic yak breeds/populations (Qilian, Yushu, Datong, Huanhu, Tongde, and Xueduo) shared a small amount of this new ancestry. When $K \geq 4$, cross-validation errors (≥0.3043) for ancestry models increased gradually, which suggested false results for the ancestry analysis (Supplementary Table S13).

The fixation index (F_{ST}) values between nine yak breeds/populations ranged from 0.003 to 0.076 (Supplementary Table S6), which showed moderate genetic differentiation ($0.05 < F_{ST} < 0.15$) between wild yak, Geermu yak, and the other seven yak breeds/populations. However, there was weak genetic differentiation ($0 < F_{ST} < 0.05$) among the other seven yak breeds/populations (Figure 4A, Supplementary Table S6). Based on the R_{ST} values (Supplementary Table S6), the clustering analysis showed that nine yak breeds/populations were separated into six groups (groups I–VI) (i.e., wild; Geermu; Huanhu; Qilian; Yushu and Datong; and Xueduo, Tongde, and Huzhu) (Figures 4B,C).

Genome-wide selective sweeps

In this study, we used the CLR, PI, and F_{ST} methods to screen for the potentially selected genomic regions in wild yak and eight domestic yak breeds/populations in Qinghai (Figures 5A–C). As a result, a total of 334 (CLR), 854 (PI), and 1268 (F_{ST}) genes were identified (Supplementary Tables S7–S9), with only 34 genes overlapping among the three analyses ($p < 0.05$) (Supplementary Tables S10, Figure 5D). These identified genes were considered candidate positively selected genes (PSGs) (Figures 5A–C). Annotations of the 34 candidate genes revealed functions that may be associated with important traits, including disease resistance (*CDK2AP2*, *PLEC*, and *CYB5B*), heat stress (*NFAT5*, *HSF1* and *SLC25A48*), pigmentation (*MCAM*, *RNF26*, and *BOP1*), vision (*C1QTNF5*, *MFRP*, and *TAX1BP3*), milk quality (*OPLAH* and *GRINA*), neurodevelopment (*SUSD4*, *INSYN1*, and *PPP1CA*), and meat quality (*ZRANB1*) (Supplementary Tables S10). As a suppressor of oxidative phosphorylation-associated gene expression, mitochondrial respiration, and reactive oxygen species (ROS) production in pulmonary artery smooth muscle cells, *NFAT5* gene is vital in limiting ROS-dependent arterial resistance in hypoxic environments (Laban et al., 2021). Here, it is notable that the genomic region harboring *NFAT5* (95.59–96.79 Mb on chromosome 20) exhibited higher F_{ST} and

differential Tajima'D and PI values between wild and domestic yak (Figure 5E). However, the haplotype diagram showed that differences in this gene region between wild and domestic yak were not obvious. To further elucidate the genetic mechanisms related to the candidate genes, we also performed functional enrichment analysis for the 34 candidate genes, using KOBAS (http://kobas.cbi.pku.edu.cn/) to find vital KEGG and GO pathways. In total, 29 KEGG pathways and 383 GO terms were found in our enrichment results (Supplementary Tables S11, S12). As for functional enrichment analysis (Supplementary Table S11), the KEGG pathway had four significant functions, called "regulation of actin cytoskeleton," "glycosylphosphatidylinositol (GPI)-anchor biosynthesis," "Legionellosis", and "glutathione metabolism" (p-value < 0.05), as well as five genes (*ITGAE*, *PPP1CA*, *GPAA1*, *HSF1*, and *OPLAH*) related to disease resistance, growth, and heat stress (Supplementary Table S10). GO terms were particularly enriched in terms of osmotic stress and endoplasmic reticulum stress ("response to osmotic stress, GO: 0006970," "nuclear stress granule, GO:0097165," and "negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway, GO:1902236"). Several genes (*NFAT5*, *MARVELD3*, *HSF1*, and *GRINA*) were found to be associated with immunity, heat stress, and milk quality. We also detected significant GO terms responsible for growth ("cellular response to platelet-derived growth factor stimulus, GO:0036120," "positive regulation of multicellular organism growth, GO: 0040018," and "cellular response to transforming growth factor beta stimulus, GO:0071560") (Supplementary Table S12) involving relevant genes (*CORO1B*, *HSF1*, and *SCX*). In addition, some significant GO terms ("protein binding, GO:0005515," "acyl carnitine transmembrane transporter activity, GO:0015227," and "acyl carnitine transmembrane transport, GO:1902616") related to disease resistance were detected, which may contribute to disease resistance of yak in harsh high-altitude environments.

Discussion

Genetic diversity is an important component of biological diversity. It is the basis of biological evolution and species differentiation, with great significance for population maintenance and adaptation to habitat change. In a previous study, Ji et al. (2021) suggested that wild yak had experienced a genetic bottleneck, yet that their genomic diversity was still higher than that of domestic yak. However, Qiu et al. (2015) concluded that the value of genomic nucleotide diversity in domestic yak (0.0014) was slightly higher than that of wild yak (0.0013). Chai et al. (2020) also showed that the genomic diversity of wild yak (0.0012) was lower than that of some domestic yak breeds/populations (0.0010–0.0016). In this study, the values of nucleotide diversity for nine Qinghai yak breeds/populations ranged from 0.0006 to 0.0015 at the whole-genome level (Table 1; Figure 2A), which indicates a wide range of genomic diversity for wild yak and eight Qinghai domestic yak breeds/populations. Here, the wild yak had nucleotide diversity of 0.0009, Huzhu white yak had the highest nucleotide diversity (0.0015), and Geermu yak had the lowest nucleotide diversity (0.0006). The nucleotide diversities of most of domestic yak breeds/populations, except for Geermu and Qilian yak, were higher than that of wild yak, indicating that most of Qinghai yak breeds/populations had abundant genomic diversity; our result is consistent with previous reports (Table 1; Figure 2A) (Qiu et al., 2015; Chai et al., 2020).

ROH has gradually become an important index for identifying inbreeding degrees and genetic variation patterns in livestock populations (Chen et al., 2018; Xia et al., 2021). Here, the ROH distribution pattern of nine yak breeds/populations showed significant differences (Figure 2D, Supplementary Table S5). The ROH numbers for Geermu and wild yak populations were small, which indicates that the living habitats of these yak populations were remote from human settlements and had less artificial intervention. Among the other domestic yak breeds/populations, Datong yak had the highest ROHs in both length and number (Figure 2D, Supplementery Table S5), followed by Yushu, Huanhu, Tongde, Huzhu white, Xueduo, and Qilian yak. The values indicated that these seven yak breeds/populations have been subjected to significant human intervention. In addition, the pattern of LD decay in each yak breed/population was largely consistent with the results of nucleotide diversity in this study. The results of inbreeding coefficients inferred that Geermu, Huanhu, Qilian, Yushu, and wild yak breeds/populations had high degrees of inbreeding (Figures 2B,C; Table 1). It is noteworthy that high selection pressure may cause inbreeding depression in yak.

In previous studies, Qiu et al. (2015) detected a clear genetic split between wild and domestic yak despite low morphological divergence and continuing gene flow between them. Similarly, Chai et al. (2020) noted that when K = 2, yak populations were divided into domestic and wild yak, while when K = 3–5, yak samples could not be divided into different ancestries, which indicates extensive genetic mixing among domestic yak. In our present study, we explored the population genetic structure of wild yak and Qinghai domestic yak breeds/populations. The ADMIXTURE analysis showed that at K = 2, nine yak breeds/populations had two ancestral components (domestic yak and wild yak), while Geermu yak carried a high genetic component of wild yak. At K = 3, a third new ancestral component appeared in a few wild yak individuals and in all domestic yak breeds/populations except for Huzhu white yak. Geermu yak showed more similar ancestral composition to wild yak, and other domestic yak breeds/populations in Qinghai showed similar ancestral components that differed from wild yak. The size of cross-validation errors could affect the authenticity of ancestry (Chen et al., 2018; Xia et al., 2021). At K = 2, the minimum CV value (0.2878, Supplementary Table S12) indicated the accurate ancestral composition of Qinghai yak, and the obtained result was consistent with previous studies (Qiu et al., 2015; Chai et al., 2020). According to PCA, among 113 individuals of nine breeds/populations in Qinghai, the Qinghai yak were divided into three clusters: wild yak, Geermu yak, and other domestic yak breeds/populations (Datong, Huanhu, Yushu, Qilian, Xueduo, Tongde, and Huzhu white yak) (Figure 3B). In an early PCA, separation among domestic yak samples demonstrated that all domestic yak populations showed a single-origin domestication and close genetic distance (Chai et al., 2020), which is similar to our current findings.

F_{ST} is the index of genetic differentiation among populations and is used to evaluate the degree of differentiation among populations. F_{ST} may show very weak (0–0.05), moderate (0.05–0.15), or significant (0.15–0.25) population differentiation. The degree of differentiation is considered extremely significant when the index exceeds 0.25 (Wright,

1978). In a previous study, based on Y-SNPs and Y-STR markers, the average F_{ST} value between wild yak and 15 domestic yak populations was 0.178, indicating significant paternal genetic differentiation between domestic and wild yak (Ma, 2019). In the present study, the F_{ST} values between nine yak breeds/populations were 0.003–0.076 (Figure 4A, Supplementary Table S6), which suggests moderate genetic differentiation between wild yak, Geermu yak, and the seven other yak breeds/populations, and weak genetic differentiation among the other seven yak breeds/populations. The result confirmed that the Geermu yak population has certain hereditary particularities. Nowadays, Datong, Huanhu, Xueduo, and Yushu yak in Qinghai have been recognized as different breeds by China National Committee of Animal Genetic Resources (National Committee of Animal Genetic Resources, 2021). In our MDS analysis (Figures 4B,C), nine yak breeds/populations were divided into six groups (I–VI) (wild; Geermu; Huanhu; Qilian; Yushu, and Datong yak; and Xueduo, Tongde, and Huzhu white yak). Here, Geermu and Qilian yak populations had significant genetic differences from other yak breeds/populations, which therefore could be considered as potential new genetic resources for further research and utilization. Notably, the clustering results among yak breeds/populations were not closely related to their geographical distributions, consistent with our PCA and ADMIXTURE analysis results (Figures 3B,C). Previously, studies based on Y-chromosome marker variations showed that both wild and domestic yak in Qinghai had relatively high paternal genetic diversity with weak phylogeographic structure and two paternal origins (Ma et al., 2018; Ma, 2019). Maternal genetic diversity in wild and domestic yak breeds/populations (Qinghai-Plateau, Huanhu, Xueduo, and Yushu yak) in Qinghai was recently examined based on nucleotide variants of mitogenomes, and the results suggested that both domestic and wild yak from Qinghai contain a wide range of maternal variability; the genetic differentiation among Qinghai indigenous yak was weak, but each Qinghai indigenous yak breed had unique maternal genetic information (Li GZ et al., 2022). In the phylogeny of yak maternal origins, the wild yak and Qinghai domestic yak were composed of three maternal lineages (lineages Ⅰ, Ⅱ, and Ⅲ) with three possible maternal origins, but only a few wild and domestic yak carried lineage III (Li et al., 2022). In this study, the clustering of nine yak breeds/populations into two large lineages (lineages Ⅰ and Ⅱ) suggests that Qinghai yak might have two origins. Among them, three sub-lineages (A, B, and C) were determined to belong to lineage Ⅰ, and no sub-lineage was found in lineage II. This result is not consistent with previous results of yak maternal origin studies but is similar to results of yak paternal phylogenetic analysis (Ma, 2019; Li et al., 2022). We acknowledge that our population genetic structure results are sensitive to the relatively limited sample sizes, but they do roughly reflect the genetic background of Qinghai yak breeds/populations included in this study.

The identification of genomic signatures of selection helps reveal genetic mechanisms underlying traits of importance in yak. Overall, 34 PSGs identified in nine Qinghai yak breeds/populations using three integrated methods appear to be possibly related to important traits in humans or other animals. For example, *NFAT5*, a member of the NFAT gene family related to immunity in humans (Dalski et al., 2001), was identified as a PSG in Qinghai yak. Therefore, *NFAT5* gene might be related to the adaptation of Qinghai yak to extreme environments and to their evolution of a powerful mitochondrial function to resist hunger, hypoxia, and severe cold. It is noticeable that the region scanned by F_{ST} for *NFAT5* exhibited higher F_{ST}, and differential Tajima'D and PI values, between wild and domestic yak, indicating strong selective sweeps. Furthermore, a diagram of the haplotypes shows that this gene region is not stable in domestic yak. The heat shock factor 1 (*HSF1*) gene is a regulator of the heat stress response, maximizing heat shock protein expression, and thus is certified to be associated with heat tolerance in Jersey, Angus, Simmental, and indicine cattle (de Fátima Bretanha Rocha et al., 2019; Rong et al., 2019; Mohapatra et al., 2021). Here, it was identified as a yak PSG, and we therefore speculated that this gene is related to yak's resistance to high-intensity ultraviolet rays and frigid climates on the QTP. In addition, the CNV of *HSF1* gene was shown to be related to growth and development in Ashidan yak (Ren et al., 2022). However, further research on this vital PSG in yak is needed. The association of economic traits with candidate genes has revealed that *ZRANB1* polymorphisms are related to the muscle pH, conductivity, flesh color, and drip loss in pigs (Huynh et al., 2013). Thus, we speculated that SNPs in this gene might be the causal sites that affect water holding capacity (WHC) in yak meat. Notably, based on the three selected methods, two genes (*GRINA* and *OPLAH*) linked to milk yield and fat contents in Holstein cows (Atashi et al., 2020; Peters et al., 2021) also showed positive selection, which may be related to milk traits in Qinghai yak.

Coat color is an important target of selection in many domestic animals. Black coat color plays an important role in protecting yak from ultraviolet radiation on the QTP. Previously, *MCAM* and *RNF26* were considered key determinants of white/black tail feather color in dwarf chickens (Nie et al., 2021). Additionally, Bao et al. (2020) revealed some modular hub genes highly related to hormones such as *SLF2*, *BOP1,* and *DPP8*, which are involved in hormone regulation related to the hair cycle in yak. In this study, three genes (*MCAM*, *RNF26*, and *BOP1*) were identified as PSGs by all three methods. We believe that these three genes are more likely to participate in hair formation and pigmentation in yak, although additional functional experiments are needed for verification. Additionally, *SUSD4*, a gene coding for a complement-related transmembrane protein involved in neurodevelopment (González-Calvo et al., 2021), may participate in cold adaptation in Qinghai yak. Overall, our study reveals several PSGs in Qinghai yak, contributing to an improved understanding of the genetic mechanisms of population characteristics and providing a molecular basis for yak breeding.

Conclusion

In conclusion, this study provides a comprehensive overview of the genomic diversity, population structure, and selection signatures of Qinghai yak using whole-genome resequencing data. Overall, most Qinghai yak breeds/populations possess abundant genomic diversity. The Qinghai yak have two ancestral components (domestic and wild yak), and the Geermu yak carry a high genetic component originating from wild yak. Moreover, Qinghai yak are clustered into wild, Geermu, and other seven yak breeds/populations, while weak genetic differentiation is displayed among the seven other domestic yak breeds/populations. The candidate genomic regions are involved in disease resistance, heat stress, pigmentation, vision, milk quality, neurodevelopment, and meat quality. This research indicates for the first time that Geermu yak, Qilian yak, and the other domestic yak breeds/populations exhibit genetic differences at the genomic level. These findings provide a theoretical basis for the reasonable

protection and utilization of Qinghai yak genetic resources in the future.

Data availability statement

The datasets presented in this study can be found in online repositories. The data presented in the study are deposited in the NCBI repository, accession BioProject number: PRJNA827919.

Ethics statement

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of the Academy of Animal Science and Veterinary Medicine, Qinghai University.

Author contributions

ZM designed and led the project; GL performed the experiments and analyzed the data; GL made the figures in the manuscript; ZM, GL, JL, FW, DX, SC, and RL contributed to sample collection; GL and ZM drafted the manuscript; JL submitted sequence data; and ZM and ZA edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (31960656) and the CAS "Light of West China" Program (3–1).

Acknowledgments

The authors thank the High-Performance Computing group at Northwest A&F University for providing computing resources. They are grateful to the editor, Zexi Cai, and the three reviewers for their insightful comments. They also thank Chuzhao Lei from Northwest A&F University and Jianlin Han from the International Livestock Research Institute for their helpful suggestions.

Conflict of interest

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

The reviewer RL declared a shared affiliation with the author FW to the handling editor at time of review.

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.2022.1034094/full#supplementary-material

References

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinforma.* 12, 246. doi:10.1186/1471-2105-12-246

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. Anim. Breed. Genet.* 137 (3), 292–304. doi:10.1111/jbg.12442

Bao, P., Luo, J., Liu, Y., Chu, M., Ren, Q., Guo, X., et al. (2020). The seasonal development dynamics of the yak hair cycle transcriptome. *BMC Genomics* 21 (1), 355. doi:10.1186/s12864-020-6725-7

Bao, Q., Ma, X. M., Jia, C. J., Wu, X. Y., Wu, Y., Meng, G. Y., et al. (2022). Resequencing and signatures of selective scans point to candidate genetic variants for hair length traits in long-haired and normal-haired Tianzhu White yak. *Front. Genet.* 13, 798076. doi:10.3389/fgene.2022.798076

Bhati, M., Kadri, N. K., Crysnanto, D., and Pausch, H. (2020). Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics* 21 (1), 27. doi:10.1186/s12864-020-6446-y

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Chai, Z. X., Xin, J. W., Zhang, C. F., Dawayangla, L., Zhang, Q., Pingcuozhandui, et al. (2020). Whole-genome resequencing provides insights into the evolution and divergence of the native domestic yaks of the Qinghai-Tibet Plateau. *BMC Evol. Biol.* 20 (1), 137. doi:10.1186/s12862-020-01702-8

Chen, N. B., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat. Commun.* 9 (1), 2337. doi:10.1038/s41467-018-04737-0

Chen, Q. M., Qu, K. X., Ma, Z. J., Zhan, J. X., Zhang, F. W., Shen, J. F., et al. (2020). Genome-wide association study identifies genomic loci associated with Neurotransmitter concentration in cattle. *Front. Genet.* 11, 139. doi:10.3389/fgene.2020.00139

Dalski, A., Schwinger, and E., and Zühlke, C. (2001). Genomic organization of the human *NFAT5* gene: exon-intron structure of the 14-kb transcript and CpG-island analysis of the promoter region. *Cytogenet Cell Genet.* 93 (3-4), 239–241. doi:10.1159/000056990

Danecek, P., and McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinforma. Oxf. Engl.* 33 (13), 2037–2039. doi:10.1093/bioinformatics/btx100

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

de Fátima Bretanha Rocha, R., Baena, M. M., de Cássia Estopa, A., Gervásio, I. C., Guaratini Ibelli, A. M., Santos Gionbelli, T. R., et al. (2019). Differential expression of *HSF1* and *HSPA6* genes and physiological responses in Angus and Simmental cattle breeds. *J. Therm. Biol.* 84, 92–98. doi:10.1016/j.jtherbio.2019.06.002

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R. (2016). SweepFinder2: Increased sensitivity, robustness and flexibility. *Bioinforma. Oxf. Engl.* 32 (12), 1895–1897. doi:10.1093/bioinformatics/btw051

E., G. X., Basang, W. D., and Zhu, Y. B. (2019a). Whole-genome analysis identifying candidate genes of altitude adaptive ecological thresholds in yak populations. *J. Anim. Breed. Genet.* 136 (5), 371–377. doi:10.1111/jbg.12403

E., G. X., Yang, B. G., Basang, W. D., Zhu, Y. B., An, T. W., and Luo, X. L. (2019b). Screening for signatures of selection of Tianzhu White yak using genome-wide re-sequencing. *Anim. Genet.* 50 (5), 534–538. doi:10.1111/age.12817

Gao, X., Wang, S., Wang, Y. F., Li, S., Wu, S. X., Yan, R. G., et al. (2022). Long read genome assemblies complemented by single cell RNA-sequencing reveal genetic and cellular mechanisms underlying the adaptive evolution of yak. *Nat. Commun.* 13 (1), 4887. doi:10.1038/s41467-022-32164-9

González-Calvo, I., Iyer, K., Carquin, M., Khayachi, A., Giuliani, F. A., Sigoillot, S. M., et al. (2021). Sushi domain-containing protein 4 controls synaptic plasticity and motor learning. *Elife* 10, e65712. doi:10.7554/eLife.65712

Guo, S. C., Peter, S., Su, J. P., Zhang, Q., Qi, D. L., Zhou, J., et al. (2006). Origin of mitochondrial DNA diversity of domestic yaks. *BMC Evol. Biol.* 6, 73–86. doi:10.1186/1471-2148-6-73

Huynh, T. P., Muráni, E., Maak, S., Ponsuksili, S., and Wimmers, K. (2013). UBE3B and ZRANB1 polymorphisms and transcript abundance are associated with water holding capacity of porcine M. longissimus dorsi. *Meat Sci.* 95 (2), 166–172. doi:10.1016/j.meatsci.2013.04.033

Ji, Q. M., Xin, J. W., Chai, Z. X., Zhang, C. F., Dawa, Y., Luo, S., et al. (2021). A chromosome-scale reference genome and genome-wide genetic variations elucidate adaptation in yak. *Mol. Ecol. Resour.* 21 (1), 201–211. doi:10.1111/1755-0998.13236

Jia, C. J., Wang, H., Li, C., Wu, X., Zan, L., Ding, X., et al. (2019). Genome-wide detection of copy number variations in polled yak using the Illumina BovineHD BeadChip. *BMC Genomics* 20 (1), 376. doi:10.1186/s12864-019-5759-1

Jia, G. X., Ding, L. M., Xu, S. R., Fang, Y. G., Fu, H. Y., and Yang, Q. E. (2020). Conservation and utilization of yak genetic resources in Qinghai-Tibet Plateau: Problems and perspectives. *Acta Ecol. Sin.* 40 (18), 6314–6323. doi:10.5846/stxb201912232763

Keller, M. C., Visscher, P. M., and Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189 (1), 237–249. doi:10.1534/genetics.111.130922

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., and Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* 5 (11), e13996. doi:10.1371/journal.pone.0013996

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Laban, H., Siegmund, S., Zappe, M., Trogisch, F. A., Heineke, J., Torre, C., et al. (2021). NFAT5/TonEBP limits pulmonary vascular resistance in the hypoxic lung by controlling mitochondrial reactive oxygen species generation in arterial smooth muscle cells. *Cells* 10 (12), 3293. doi:10.3390/cells10123293

Lan, D., Xiong, X., Mipam, T. D., Fu, C., Li, Q., Ai, Y., et al. (2018). Genetic diversity, molecular phylogeny, and selection evidence of jinchuan yak revealed by whole-genome resequencing. *G3 (Bethesda, Md)* 8 (3), 945–952. doi:10.1534/g3.118.300572

Lan, D., Ji, W., Xiong, X., Liang, Q., Yao, W., Mipam, T. D., et al. (2021). Population genome of the newly discovered Jinchuan yak to understand its adaptive evolution in extreme environments and generation mechanism of the multirib trait. *Integr. Zool.* 16 (5), 685–695. doi:10.1111/1749-4877.12484

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, G. Z., Ma, Z. J., Chen, S. M., Lei, C. Z., Li, R. Z., Xie, Y. L., et al. (2022). Maternal genetic diversity, differentiation and phylogeny of wild yak and local yak breeds in Qinghai inferred from mitogenome sequence variations. *Acta Vet. Zootech. Sin.* 53 (5), 1420–1430. doi:10.11843/j.issn.0366-6964

Liang, C., Wang, L., Wu, X., Wang, K., Ding, X., Wang, M., et al. (2016). Genomewide association study identifies loci for the polled phenotype in yak. *PLoS ONE* 11 (7), e0158642. doi:10.1371/journal.pone.0158642

Ma, Z. J., Xia, X. T., Chen, S. M., Zhao, X. C., Zeng, L. L., Xie, Y. L., et al. (2018). Identification and diversity of Y-chromosome haplotypes in Qinghai yak populations. *Anim. Genet.* 49 (6), 618–622. doi:10.1111/age.12723

Ma, Z. J., Li, G. Z., Chen, S. M., Han, J. L., and Hanif, Q. (2022). Rich maternal and paternal genetic diversity and divergent lineage composition in wild yak (*Bos mutus*). *Anim. Biotechnol.* 33 (6), 1318–1321. doi:10.1080/10495398.2021.1895187

Ma, Z. J. (2019). *Study on the paternal genetic diversity and origin of the yak (*Bos grunniens*)*. [doctoral dissertation]. China: Northwest Agriculture & Forestry University.

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 (9), 1297–1303. doi:10.1101/gr.107524.110

Medugorac, I., Graf, A., Grohs, C., Rothammer, S., Zagdsuren, Y., Gladyr, E., et al. (2017). Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. *Nat. Genet.* 49 (3), 470–475. doi:10.1038/ng.3775

Mohapatra, S., Kundu, A. K., Mishra, S. R., Senapati, S., Jyotiranjan, T., and Panda, G. (2021). HSF1 and GM-CSF expression, its association with cardiac health, and assessment of organ function during heat stress in crossbred Jersey cattle. *Res. Vet. Sci.* 139, 200–210. doi:10.1016/j.rvsc.2021.07.018

National Committee of Animal Genetic Resources (2021). *National list of livestock and poultry genetic resources in China*. Beijing, China. Available at: http://www.moa.gov.cn/govpublic/nybzzj1/202101/t20210114_6359937.htm.

Nie, C., Qu, L., Li, X., Jiang, Z., Wang, K., Li, H., et al. (2021). Genomic regions related to white/black tail feather color in dwarf chickens identified using a genome-wide association study. *Front. Genet.* 12, 566047. doi:10.3389/fgene.2021.566047

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15 (11), 1566–1575. doi:10.1101/gr.4252305

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Peters, S. O., Kızılkaya, K., Ibeagha-Awemu, E. M., Sinecen, M., and Zhao, X. (2021). 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.* 104 (2), 1900–1916. doi:10.3168/jds.2020-18489

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Qiu, Q., Zhang, G. J., Ma, T., Qian, W. B., Wang, J. Y., Ye, Z. Q., et al. (2012). The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44 (8), 946–949. doi:10.1038/ng.2343

Qiu, Q., Wang, L. Z., Wang, K., Yang, Y. Z., Ma, T., Wang, Z. F., et al. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat. Commun.* 6, 10283. doi:10.1038/ncomms10283

Ren, W. W., Huang, C., Ma, X. M., La, Y. F., Chu, M., Guo, X., et al. (2022). Association of *HSF1* gene copy number variation with growth traits in the Ashidan yak. *Gene* 842, 146798. doi:10.1016/j.gene.2022.146798

Rong, Y., Zeng, M. F., Guan, X. W., Qu, K. X., Liu, J. Y., Zhang, J. C., et al. (2019). Association of *HSF1* genetic variation with heat tolerance in Chinese cattle. *Animals* 9 (12), 1027. doi:10.3390/ani9121027

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Wang, K., Li, M. Y., and Hakonarson, H. (2010). Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Wang, K., Hu, Q. J., Ma, H., Wang, L. Z., Yang, Y. Z., Luo, W. C., et al. (2014). Genome-wide variation within and between wild and domestic yak. *Mol. Ecol. Resour.* 14 (4), 794–801. doi:10.1111/1755-0998.12226

Wang, H., Chai, Z. X., Hu, D., Ji, Q. M., Xin, J. W., Zhang, C. F., et al. (2019). A global analysis of CNVs in diverse yak populations using whole-genome resequencing. *BMC Genomics* 20 (1), 61. doi:10.1186/s12864-019-5451-5

Wang, X. D., Pei, J., Bao, P. J., Cao, M. L., Guo, S. K., Song, R. D., et al. (2021). Mitogenomic diversity and phylogeny analysis of yak (*Bos grunniens*). *BMC Genomics* 22 (1), 325. doi:10.1186/s12864-021-07650-x

Wiener, G., Han, J. L., and Long, R. J. (2003). *The Yak*. 2nd edition. Bangkok, Thailand: The Regional Office for Asia and the Pacific of the Food and Agriculture Organization of the United Nations.

Wright, S. (1978). "Evolution and the genetics of populations," in *Variability within and among natural populations* (Chicago: University of Chicago Press).

Xia, X. T., Zhang, S. J., Zhang, H. J., Zhang, Z. J., Chen, N. B., Li, Z. G., et al. (2021). Assessing genomic diversity and signatures of selection in Jiaxian Red cattle using whole-genome sequencing data. *BMC Genomics* 22 (1), 43. doi:10.1186/s12864-020-07340-0

Xie, X., Yang, Y., Ren, Q., Ding, X., Bao, P., Yan, B., et al. (2018). Accumulation of deleterious mutations in the domestic yak genome. *Anim. Genet.* 49 (5), 384–392. doi:10.1111/age.12703

Zhang, X., Wang, K., Wang, L. Z., Yang, Y. Z., Ni, Z. Q., Xie, X. Y., et al. (2016). Genome-wide patterns of copy number variation in the Chinese yak genome. *BMC Genomics* 17, 379. doi:10.1186/s12864-016-2702-6

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinforma. Oxf. Engl.* 35 (10), 1786–1788. doi:10.1093/bioinformatics/bty875

Zhang, K., Lenstra, J. A., Zhang, S., Liu, W., and Liu, J. (2020). Evolution and domestication of the Bovini species. *Anim. Genet.* 51 (5), 637–657. doi:10.1111/age.12974

Zhang, S. Z., Liu, W. Y., Liu, X. F., Du, X., Zhang, K., Zhang, Y., et al. (2021). Structural variants selected during yak domestication inferred from long-read whole-genome sequencing. *Mol. Biol. Evol.* 38 (9), 3676–3680. doi:10.1093/molbev/msab134

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 24 January 2023
DOI 10.3389/fgene.2022.1028662

OPEN ACCESS

EDITED BY
Zexi Cai,
Aarhus University, Denmark

REVIEWED BY
Olga Matveeva,
The University of Utah, United States
Yuchun Pan,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Qianqian Zhang,
zhangqianqian186@hotmail.com

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

RECEIVED 26 August 2022
ACCEPTED 15 November 2022
PUBLISHED 24 January 2023

CITATION
Cui B, Guo Z, Cao H, Calus M and Zhang Q (2023), The computational implementation of a platform of relative identity-by-descent scores algorithm for introgressive mapping.
Front. Genet. 13:1028662.
doi: 10.3389/fgene.2022.1028662

The computational implementation of a platform of relative identity-by-descent scores algorithm for introgressive mapping

Bo Cui[1,2,3], Zhongxu Guo[1,2,3], Hongbo Cao[1,2,3], Mario Calus[4] Qianqian Zhang[1,3]*

[1]School of Chemistry and Biological Engineering, University of Science and Technology, Beijing, China, [2]College of Water Resources and Civil Engineering, China Agricultural University, Beijing, China, [3]Institute of Biotechnology, Beijing Academy of Agricultural and Forestry Sciences, Beijing, China, [4]Department of Animal Science, Animal Breeding and Genetics Group, Wageningen University, Wageningen, Netherlands

With the development of genotyping and sequencing technology, researchers working in the area of conservation genetics are able to obtain the genotypes or even the sequences of a representative sample of individuals from the population. It is of great importance to examine the genomic variants and genes that are highly preferred or pruned during the process of adaptive introgression or long-term hybridization. To the best of our knowledge, we are the first to develop a platform with computational integration of a relative identity-by-descent (rIBD) scores algorithm for introgressive mapping. The rIBD algorithm is designed for mapping the fine-scaled genomic regions under adaptive introgression between the source breeds and the admixed breed. Our rIBD calculation platform provides compact functions including reading input information and uploading of files, rIBD calculation, and presentation of the rIBD scores. We analyzed the simulated data using the rIBD calculation platform and calculated the average IBD score of 0.061 with a standard deviation of 0.124. The rIBD scores generally follow a normal distribution, and a cut-off of 0.432 and −0.310 for both positive and negative rIBD scores is derived to enable the identification of genomic regions showing significant introgression signals from the source breed to the admixed breed. A list of genomic regions with detailed calculated rIBD scores is reported, and all the rIBD scores for each of the considered windows are presented in plots on the rIBD calculation platform. Our rIBD calculation platform provides a user-friendly tool for the calculation of fine-scaled rIBD scores for each of the genomic regions to map possible functional genomic variants due to adaptive introgression or long-term hybridization.

KEYWORDS

adaptive introgression, hybridization, identity by descent, gene flow, algorithm, computational platform

1 Introduction

Adaptive introgression in situations with gene flow between different breeds or even species holds high interest in evolutionary genetics, as the genetic basis of this phenomenon is largely unknown when there is gene flow between different breeds, varieties, species, etc. (Voight et al., 2006; Bosse et al., 2014a; Ai et al., 2015). Identifying the genes or genomic regions that are due to introgression will help understand the genetic basis of long-term hybridization and admixture. It is of great importance to identify which genomic regions or genes influence the phenotypic character or traits of hybrids and how these have interacted with each other to form the genetic basis of hybrid breeds (Bosse et al., 2014b; Jagoda et al., 2017; Zhang et al., 2018). The genomic regions or genes identified with an important role in adaptive introgression hold a higher possibility as a causal variant affecting the phenotypes or traits. For example, Tibetans with altitude adaptation are caused by a genetic background traced long back to Denisovan-like DNA (Huerta-Sanchez et al., 2014; Ai et al., 2015). By identifying these important facts, we are able to disentangle the genetic basis of hybridization and the genetic effects of introgression during adaptation for important phenotypes, traits, or even diseases, which has long been an important research topic receiving much attention.

With the development of genotyping and sequencing technology, it is possible to sequence individuals deeply for all nucleotides with no ascertainment bias (Daetwyler et al., 2014; Zhang et al., 2015). This enables detailed examination of single genomes to understand the phenomenon of introgression and hybridization at the population level, which provides a valid basis for the information of mapping the functional genomic variants in individuals (Ai et al., 2015; Chen et al., 2016; Qanbari et al., 2011; vonHoldt et al., 2016). After the hybridization, some of the genomic variants can be favored or pruned out with a high or low frequency in the population across several generations of directional selection (Hedrick, 2013; Bosse et al., 2014a; Ai et al., 2015; Galov et al., 2015; Hartwig et al., 2015). These introgressed genomic variants are highly likely to be functional and play a key role during the process of introgression and hybridization (Bosse et al., 2014b; Hasenkamp et al., 2015; Deschamps et al., 2016; Figueiro et al., 2017). Mapping these genomic variants that have been subjected to adaptive introgression using a valid method provides information about these variants that can be used for further functional validations.

Many studies have identified these functional genomic variants that have been subjected to adaptive introgression (Bosse et al., 2014b; Deschamps et al., 2016; Figueiro et al., 2017; Wu et al., 2018; Zhang et al., 2018). The aim of these studies is to examine the genome-wide signatures of adaptive divergence and introgression in depth and further disentangle the genetic basis of the complex traits formed during this process by the utilization of genomic analysis from phylogeny. Among these studies, Zhang et al. (2018) utilized the relative identity-by-descent (rIBD) algorithm to identify the genomic regions in an admixed Red cattle breed that arose from adaptive introgression from Holstein and Brown Swiss cattle breeds. Figueiro et al. (2017) found complex genomic signatures of introgression using phylogeny, comparative analysis, and demographic reconstructions, and genes involved in craniofacial and limb development were identified.

So far, there are many tools which can calculate the proportion of admixed genomes obtained from ancestral breeds at the level of an individual, such as ADMIXTURE (Alexander et al., 2009). It is also possible to construct the phylogeny and demographic history using tools such as RAxML (Stamatakis, 2014) and PSMC (Li and Durbin, 2011). These tools and methods could examine the demographic history and population structure on the basis of an individual genome. There is, however, no tool which enables to map these genomic variants at a fine scale, that is, up to base-pair resolution. The algorithm of rIBD was previously applied in our studies for adaptive introgression mapping (Zhang et al., 2018). IBD inference is to detect the haplotypes which are inherited from the common ancestor, and the IBD states could reflect the general pattern of demographic history on the population scale (Sticca et al., 2021). The advantage of IBD detection is that it is even possible to trace back to the recent common ancestor of the shared pattern of rare variation, which has been long mysterious for the research scientists in the field. Therefore, IBD detection is of great importance for addressing the unanswered questions in genomics and genetics.

Here, we develop a user-friendly platform with the implementation of the algorithm of rIBD (Bosse et al., 2014b; Zhang et al., 2018) for identifying genomic variants and genomic regions that have been subjected to introgression in a fine-scaled sliding window across the whole genome from an evolutionary perspective. These genomic variants that have been subjected to adaptive introgression are basically identity by descent with the ancestral breed from which it was derived when comparing pair-wise between individual genomes. The objective of this study is to 1) implement the rIBD algorithm with a fine-scaled sliding window, 2) integrate rIBD algorithm in a user-friendly platform so that the users can utilize it as an online tool, and 3) demonstrate the different aspects of the results calculated using rIBD algorithm and different compact modules of the rIBD calculation platform based on a small simulated dataset.

2 Materials and methods

We implement the method of rIBD algorithm into the platform in three steps: 1) input and the design of the front interface: files, options, and illustrations; 2) the rIBD calculations: calculation options; and 3) output: files and illustrations.

We first design the front interface with the options of inputting the files. Here, it is designed to have the input information from source breed 1, source breed 2, and the breed with admixture. It is necessary to calculate the basic identity-by-descent (IBD) information first between source breed 1 and the admixed breed and between source breed 2 and the admixed breed based on the genomic marker input information. Our rIBD scores are calculated based on the posterior inferences of the basic IBD states of shared IBD tracts between two individuals. It is also possible to use the sample data as input for a quick rIBD calculation and demonstration. At the moment, our platform only supports rIBD calculations for one admixed breed at a time that is derived from no more than two ancestral breeds. When there are more ancestral breeds to be considered, the user should perform one separate analysis for each possible pair of ancestral breeds.

The next step is to utilize the input information described previously to calculate the rIBD scores using a sliding window across the whole genome. Here, we design the rIBD algorithm for a genome with a sliding window of 10 kbp, that is, the minimum resolution for the possibility to map a gene on a single genome, is selected as the size of the sliding window. In this way, we could reach the highest resolution to scan the genomes with all possible adaptive introgression signals. Using a sliding window of 10 kbp, we make pair-wise comparisons between each admixed individual and all individuals in source breed 1, and all individuals in source breed 2. The proportion of genomic regions which are calculated as significant IBD is outputted between source breed 1 and the breed with admixture and between source breed 2 and the breed with admixture. A relative IBD (rIBD) score is finally calculated when considering source breed 1, source breed 2, and the breed with admixture in the following format:

$$IBD_{S1} - IBD_{S2}, \quad (1)$$

where IBD_{S1} refers to the proportion of genomes in IBD between source breed 1 and the breed with admixture and IBD_{S2} is the proportion of genomes in IBD between source breed 2 and the breed with admixture. The proportion of genomes in IBD is calculated as the proportion of admixed individuals' genomes which are IBD with source breed 1 or 2. There are many algorithms to infer the state of IBD, and it is suggested to use the hidden Markov model to infer the tract of the IBD haplotype for each pair of individuals (BrianBrowning, 2011). The obtained rIBD scores are then used for mapping the introgressed genomic regions in the admixed breed. Our designed rIBD algorithm intends to identify these introgressed genomic regions in a pair-wise comparison between individuals so that an average rIBD score is calculated at a population level between the different breeds or populations.

The scale of the rIBD scores ranges, by definition, from −1 to 1. In an admixed breed, genomic regions with rIBD scores between −1 and 0 apparently have been introgressed more strongly from source breed 2, while genomic regions with rIBD scores between 0 and 1 apparently have been introgressed more strongly from source breed 1. Noting that with two breeds involved in an admixed population, the sum of IBD_{S1} and IBD_{S2} is 1, and an rIBD value of, for instance, 0.5 implies that this region has IBD_{S1} = 0.75 and IBD_{S2} = 0.25. Generally, the rIBD scores will follow a normal distribution, while genomic regions with extreme rIBD scores in the tails of the distribution indicate that these in the admixed breed are more similar to source breed 1 or source breed 2. It is also possible to calculate average rIBD scores from the sliding windows across the whole genome, which will reflect the genome-wide level of introgression between source breed 1 or 2 and the admixed breed. However, we focused more on the genomic regions showing clear signals with significant introgression from a certain source breed, that is, extreme positive rIBD scores or missing the rIBD scores from one side of the source breed, that is, extreme negative rIBD scores. Normally more efforts should be made to explore more in these genomic regions with a significant positive rIBD score or an extended genomic region with negative rIBD scores.

Finally, output options include outputting the rIBD scores in the sliding windows and plotting the rIBD scores across the chromosome. In this way, the user could save the rIBD scores calculated using the platform and also directly obtain an illustration of the rIBD scores across the chromosome. The users can use the sample data to demonstrate the output rIBD scores and the rIBD plot, and we used this as an example to show the results here.

3 Results and discussion

3.1 The relative identity-by-descent calculation platform

rIBD algorithm is based on the relative difference of IBD on genomes between the source breed and the admixed breed (Figure 1A), and this algorithm is designed to identify genomic regions that have been subjected to adaptive introgression when the admixed breed has gene flow from two source breeds. We implement rIBD algorithm described in the Materials and methods section using the C programming language and Perl programming language and then re-write the programming part of the algorithm on the rIBD calculation platform. The rIBD calculation platform is available at https://cuibobetter.shinyapps.io/example-1/.

The rIBD platform is a user-friendly calculation platform with different options on the front interface. In the main section, the first step is to register as a user on our rIBD calculation platform. Then, the user should input the information from source breed 1, source breed 2, and an admixed breed and the

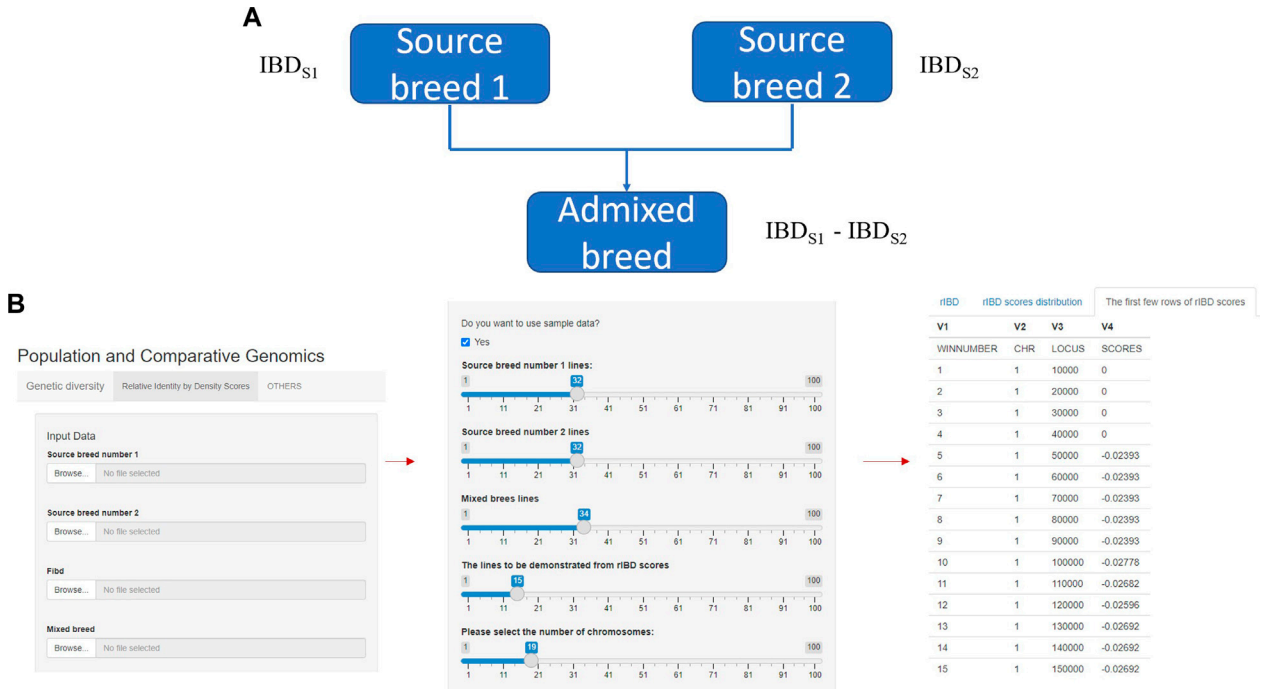


FIGURE 1
The core and main section of rIBD calculation platform and the process of calculating rIBD scores on the platform. **(A)** The core flow of the rIBD calculation algorithm; **(B)** The main section of rIBD calculation platform and the process of calculating rIBD scores on the platform.

basic IBD information in the main section (Figure 1A). The basic IBD information could be calculated using software packages such as Beagle (BrianBrowning, 2011). Usually, the breeds are clearly defined and the history of the breed is recorded, so it is possible to clearly define the source breeds and the admixed breed. The user also needs to specify the number of individuals from source breed 1, source breed 2, and the admixed breed and the number of chromosomes for this species (Figure 1B). After inputting these details in the main section, the rIBD calculation platform will output the rIBD scores for a sliding window of 10 kbp (Figure 1B). We only demonstrate the first few lines according to the user's specification on the calculation platform so that the user has a direct illustration of the rIBD scores (Figure 1B). This output includes the chromosome number, the start and end positions of the corresponding sliding window in the size of 10 kbp, and the calculated rIBD scores in this window (Figure 1B). Meanwhile, it is also possible to utilize an executable of the rIBD calculation package for analyzing large amounts of data due to the limited space on the online platform as it is not possible to upload large-scale data on the platform, and the users should refer to the corresponding author of this work, for e.g., the executable of the rIBD algorithm.

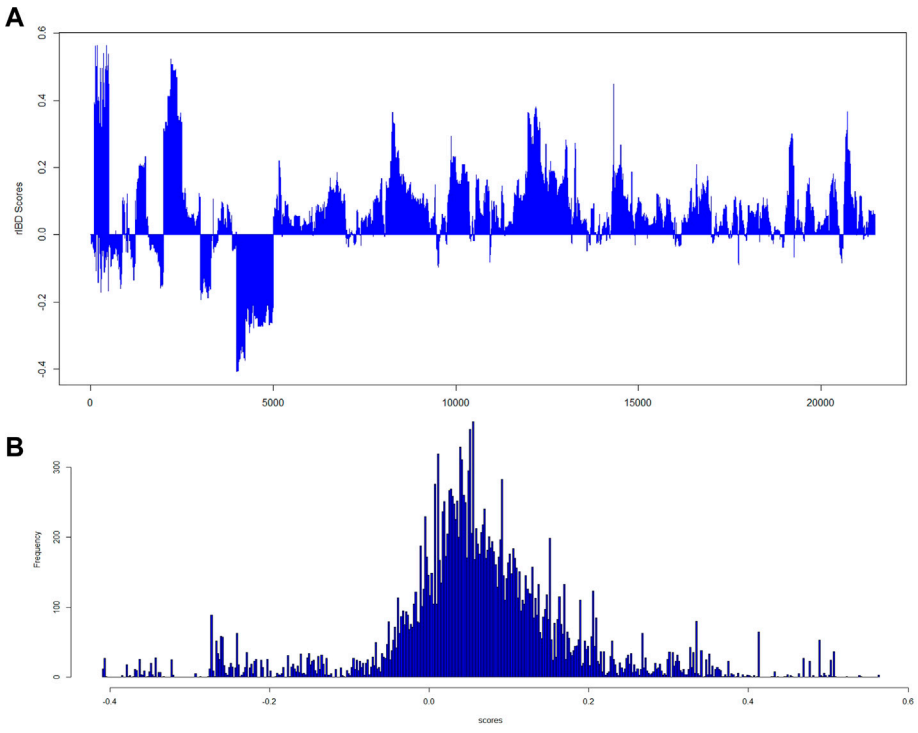
Our rIBD calculation platform will be published as an online calculation platform at the end for the users around the world to calculate the rIBD scores and explore the genomic regions and the genetic basis of adaptive introgression. Our platform provides the users to automatically calculate either the rIBD scores from a specific genomic region or across the whole genome according to the capacity of our platform following the user's requirement.

3.2 Analysis of data and illustrations of the calculated relative identity-by-descent scores on the platform

Here, we analyze a set of simulated genomic data from source breed 1, source breed 2, and one admixed breed. The aim of the analysis is to map the genomic regions that have been subjected to adaptive introgression in the admixed breed. Our rIBD platform is based on powerful Beagle, that is, the basic IBD states are first inferred from Beagle using HMM algorithm (BrianBrowning, 2011).

3.2.1 Interpretation of output results

Then, our rIBD platform is utilized to calculate the rIBD scores for this set of genomic data, and the rIBD scores are plotted across the whole genomic region (Figure 2A). The average rIBD score across this genomic region is 0.061, and



C

WINNUMBER	CHR	LOCUS	SCORES
174	1	1740000	0.5632
435	1	4350000	0.5632
123	1	1230000	0.5621
350	1	3500000	0.5411
177	1	1770000	0.5386
475	1	4750000	0.5386
2187	7	21870000	0.5236
340	1	3400000	0.5086
2225	7	22250000	0.5068
2226	7	22260000	0.5068
2227	7	22270000	0.5068
2228	7	22280000	0.5068
2229	7	22290000	0.5068
2230	7	22300000	0.5068
2231	7	22310000	0.5068

D

WINNUMBER	CHR	LOCUS	SCORES
4000	9	40000000	-0.4082
4001	9	40010000	-0.4082
4002	9	40020000	-0.4082
4003	9	40030000	-0.4082
4004	9	40040000	-0.4082
4005	9	40050000	-0.4082
4006	9	40060000	-0.4082
4007	9	40070000	-0.4082
4008	9	40080000	-0.4082
4009	9	40090000	-0.4082
4010	9	40100000	-0.4082
4011	9	40110000	-0.4082
4012	9	40120000	-0.4063
4013	9	40130000	-0.4063
4014	9	40140000	-0.4063

FIGURE 2
The distribution of rIBD scores across the chromosome and the significant positive or negative rIBD scores lists. **(A)** The calculated rIBD scores across the chromosome; **(B)** Tthe distribution of the rIBD scores; **(C)** The list of the significant positive rIBD scores list; **(D)** The list of the significant negative rIBD scores list.

the standard deviation of the rIBD scores is 0.124 (Figure 2B). This reflects that there is an overall introgression level of 0.061, which suggests that the introgression level from source breed 1 to the admixed breed is higher than that from source breed 2 to the admixed breed. The genomic regions with positive rIBD scores are more related with the introgression from source breed 1 to the admixed breed, which are the important genomic regions related with, for example, directional selection during the adaptive introgression process. However, the genomic regions with negative rIBD scores reflect the genes in the region with higher frequency due to introgression from source breed 2 to the admixed breed and directional selection later in the admixed breed during the adaptive introgression process.

3.2.2 Identification of signals due to adaptive introgression with a high level of significance

Generally, the rIBD scores follow a normal distribution, and there are two extreme tails on the distribution (Figure 2B), showing that there are genomic regions with rIBD scores exceeding the significant level. We further take a very stringent cut-off of 0.432 and −0.310 calculated as three times the standard deviation from the mean in the normal distribution corresponding to 0.3% of the distribution for both positive and negative rIBD scores in order to identify the genomic regions with significant signals due to adaptive introgression. These significant genomic regions are mapped and listed as shown in Figures 2C,D. In this way, the genes in these genomic regions due to adaptive introgression can be mapped, and the gene list can be used to perform further functional studies.

The output of rIBD calculation can also be briefly demonstrated on the rIBD calculation platform. For example, the users could plot the rIBD scores across the genomic region to demonstrate the significant genomic regions with extreme high or low rIBD scores showing adaptive introgression signals from different source breeds (Figure 3A). Meanwhile, the users could also plot the distribution of the calculated rIBD scores to show the general pattern of the calculated rIBD scores and the extreme values of the calculated rIBD scores (Figure 3B). Generally, our rIBD calculation platform has compact functions and modules including inputting information, calculating rIBD scores, outputting the calculated rIBD scores, and illustrating the plots using the calculated rIBD scores for the users.

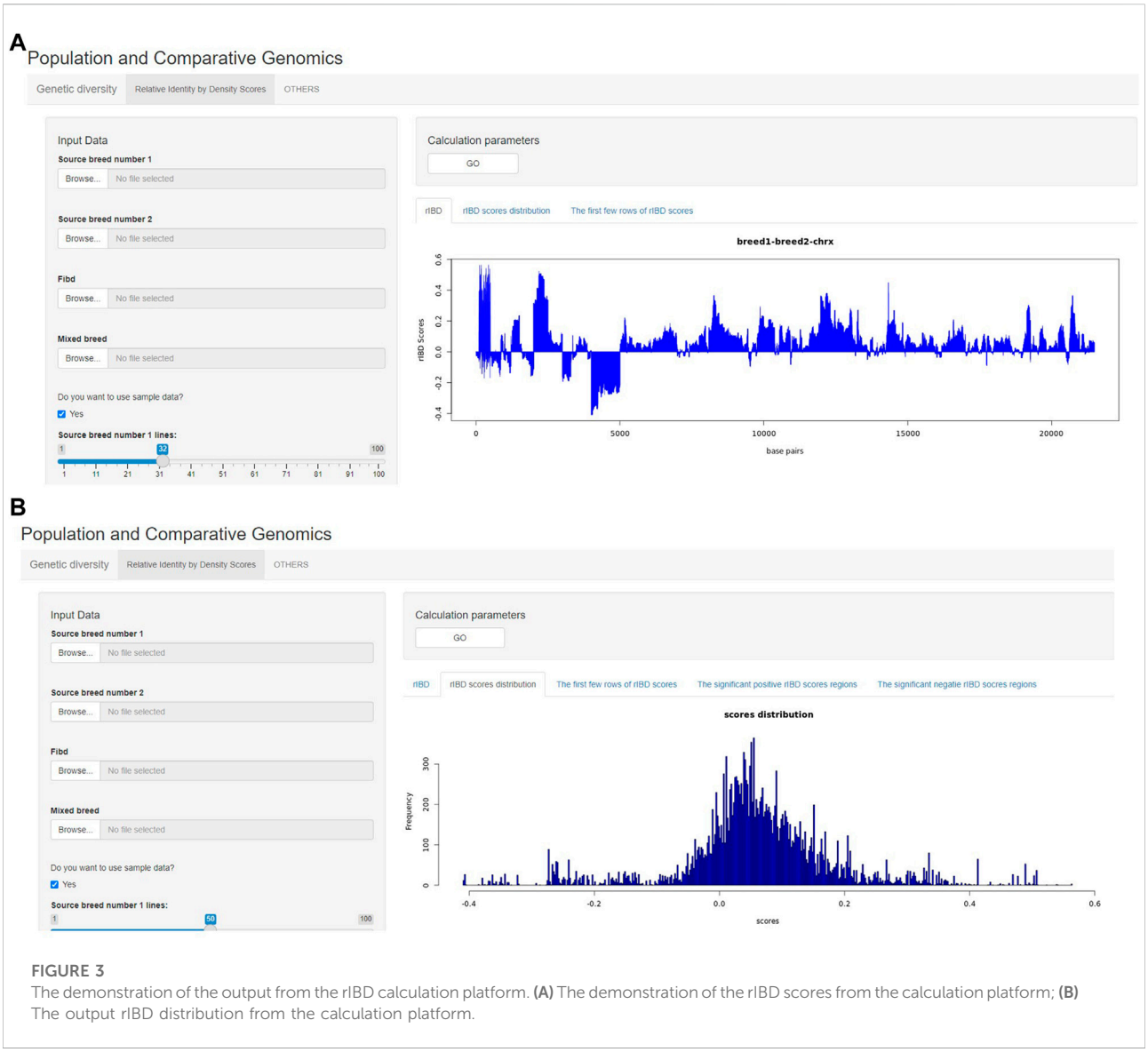


FIGURE 3
The demonstration of the output from the rIBD calculation platform. **(A)** The demonstration of the rIBD scores from the calculation platform; **(B)** The output rIBD distribution from the calculation platform.

3.2.3 Advantages of the relative identity-by-descent calculation platform

The advantage of our rIBD calculation platform is that the users are able to input the related information and calculate the rIBD scores directly without requiring any basic knowledge of bioinformatics. This would aid many agricultural breeders, especially from developing countries to obtain the introgression or long-term hybridization information between different breeds. We develop and implement a tool for the first time for rIBD score calculation integrated into a platform that can conveniently be used by agricultural breeders and researchers working in the conservation genetics area. It is necessary to mention that the amount of data which could be analyzed is limited due to the limited support to pay for the running of the standard online server at the moment. The amount of data to be analyzed and the ability of the calculations will also be increased and enlarged with the enlarging impact and usage of our rIBD calculation platform.

4 Conclusion

Exploring the genetic basis of long-term hybridization and adaptive introgression is extremely important for explaining evolutionary phenomena in biology, and a method with a corresponding tool to identify these genomic regions in detail during adaptive introgression is extremely useful for the researchers in the area. In this study, we developed the rIBD method and integrated into a calculation platform

that is particularly useful for agricultural breeders and researchers working in the area of conservation breeding. Our rIBD method is based on pair-wise IBD comparisons between individual genomes from different source breeds and the admixed breed to map the specific genomic regions due to adaptive introgression. We then integrate the calculation platform with illustrations for the convenient utilization of users. We demonstrate the structure of our rIBD calculation platform and the usage of different modules of our rIBD calculation platform. Our rIBD calculation platform is a useful tool for the researchers working in conservation genetics and conservation breeding area and the agricultural breeders to study the genetic basis and map the genes in detail during adaptive introgression and long-term hybridization.

Data availability statement

Publicly available calculation platform is available and can be found at: https://39.106.33.52:3838/sample-apps/rIBD/. The executable version of the implemented rIBD algorithm can be required from the corresponding author when necessary.

Author contributions

BC implemented the computational programming part of this study and participated in the analysis of data in this study. ZG and HC assisted BC in his work in this study. MC participated in the design of the study and drafting of the manuscript. QZ developed and planned the design of the study, coordinated the study, instructed in implementing the computational programming, analysis of the data and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by Beijing Nova Program, Beijing Committee of Science and Technology, China (grant No. Z201100006820091).

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

Ai, H. S., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* 47 (3), 217–225. doi:10.1038/ng.3199

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Bosse, M., Megens, H. J., Frantz, L. A. F., Madsen, O., Larson, G., Paudel, Y., et al. (2014). Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat. Commun.* 5, 4392. doi:10.1038/ncomms5392

Bosse, M., Megens, H. J., Madsen, O., Frantz, L. A. F., Paudel, Y., Crooijmans, R. P. M. A., et al. (2014). Untangling the hybrid nature of modern pig genomes: A mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. *Mol. Ecol.* 23 (16), 4089–4102. doi:10.1111/mec.12807

BrianBrowning, L. S. R. B. (2011). A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 103 (3), 338–348.

Chen, M. H., Pan, D., Ren, H., Fu, J., Li, J., Su, G., et al. (2016). Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. *Genet. Sel. Evol.* 48, 76. doi:10.1186/s12711-016-0254-5

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brondum, R. F., 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

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J. L., et al. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* 98 (1), 5–21. doi:10.1016/j.ajhg.2015.11.014

Figueiro, H. V., Li, G., Trindade, F. J., Assis, J., Pais, F., Fernandes, G., et al. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci. Adv.* 3 (7), e1700299. doi:10.1126/sciadv.1700299

Galov, A., Fabbri, E., Caniglia, R., Arbanasic, H., Lapalombella, S., Florijancic, T., et al. (2015). First evidence of hybridization between golden jackal (*Canis aureus*) and domestic dog (*Canis familiaris*) as revealed by genetic markers. *R. Soc. Open Sci.* 2 (12), 150450. doi:10.1098/rsos.150450

Hartwig, S., Wellmann, R., EmmeRling, R., Hamann, H., and Bennewitz, J. (2015). Short communication: Importance of introgression for milk traits in the German Vorderwald and Hinterwald cattle. *J. Dairy Sci.* 98 (3), 2033–2038. doi:10.3168/jds.2014-8571

Hasenkamp, N., Solomon, T., and Tautz, D. (2015). Selective sweeps versus introgression - population genetic dynamics of the murine leukemia virus receptor Xpr1 in wild populations of the house mouse (*Mus musculus*). *BMC Evol. Biol.* 15, 248. doi:10.1186/s12862-015-0528-5

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22 (18), 4606–4618. doi:10.1111/mec.12415

Huerta-Sanchez, E., Jin, X., Bianba, Z., Peter, B. M., Vinckenbosch, N., Yu, L., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512 (7513), 194–197. doi:10.1038/nature13408

Jagoda, E., Lawson, D. J., Wall, J. D., Lambert, D., Muller, C., Westaway, M., et al. (2017). Disentangling immediate adaptive introgression from selection on standing introgressed variation in humans. *Mol. Biol. Evol.* 35, 623–630. doi:10.1093/molbev/msx314

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475 (7357), 493–496. doi:10.1038/nature10231

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., et al. (2011). Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *Bmc Genomics* 12, 318. doi:10.1186/1471-2164-12-318

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30 (9), 1312–1313. doi:10.1093/bioinformatics/btu033

Sticca, E. L., Belbin, G. M., and Gignoux, C. R. (2021). Current developments in detection of identity-by-descent methods and applications. *Front. Genet.* 12, 722602. doi:10.3389/fgene.2021.722602

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4 (3), e72–e458. doi:10.1371/journal.pbio.0040072

vonHoldt, B. M., Kays, R., Pollinger, J. P., and Wayne, R. K. (2016). Admixture mapping identifies introgressed genomic regions in North American canids. *Mol. Ecol.* 25 (11), 2443–2453. doi:10.1111/mec.13667

Wu, D. D., Ding, X. D., Wang, S., Wojcik, J. M., Zhang, Y., Tokarska, M., et al. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat. Ecol. Evol.* 2 (7), 1139–1145. doi:10.1038/s41559-018-0562-y

Zhang, Q. Q., Calus, M. P. L., Bosse, M., Sahana, G., Lund, M. S., and Guldbrandtsen, B. (2018). Human-Mediated introgression of haplotypes in a modern dairy cattle breed. *Genetics* 209 (4), 1305–1317. doi:10.1534/genetics.118.301143

Zhang, Q. Q., Guldbrandtsen, B., Bosse, M., Lund, M. S., and Sahana, G. (2015). Runs of homozygosity and distribution of functional variants in the cattle genome. *Bmc Genomics* 16, 542. doi:10.1186/s12864-015-1715-x

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fgene.2023.1083106

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
George R. Wiggans,
Council on Dairy Cattle Breeding, United States
Setegn Worku Alemu,
Massey University, New Zealand

*CORRESPONDENCE
Maryam Esrafili Taze Kand Mohammaddiyeh,
m.m.esrafili@gmail.com
Seyed Abbas Rafat,
abbasrafat@hotmail.com

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

RECEIVED 28 October 2022
ACCEPTED 28 February 2023
PUBLISHED 17 March 2023

CITATION
Esrafili Taze Kand Mohammaddiyeh M, Rafat SA, Shodja J, Javanmard A and Esfandyari H (2023), Selective genotyping to implement genomic selection in beef cattle breeding.
Front. Genet. 14:1083106.
doi: 10.3389/fgene.2023.1083106

Selective genotyping to implement genomic selection in beef cattle breeding

Maryam Esrafili Taze Kand Mohammaddiyeh[1]*, Seyed Abbas Rafat[1]*, Jalil Shodja[1], Arash Javanmard[1] and Hadi Esfandyari[2]

[1]Department of Animal Sciences, University of Tabriz, Tabriz, Iran, [2]Norwegian Beef Cattle Organizations, TYR, Hamar, Norway

Genomic selection (GS) plays an essential role in livestock genetic improvement programs. In dairy cattle, the method is already a recognized tool to estimate the breeding values of young animals and reduce generation intervals. Due to the different breeding structures of beef cattle, the implementation of GS is still a challenge and has been adopted to a much lesser extent than dairy cattle. This study aimed to evaluate genotyping strategies in terms of prediction accuracy as the first step in the implementation of GS in beef while some restrictions were assumed for the availability of phenotypic and genomic information. For this purpose, a multi-breed population of beef cattle was simulated by imitating the practical system of beef cattle genetic evaluation. Four genotyping scenarios were compared to traditional pedigree-based evaluation. Results showed an improvement in prediction accuracy, albeit a limited number of animals being genotyped (i.e., 3% of total animals in genetic evaluation). The comparison of genotyping scenarios revealed that selective genotyping should be on animals from both ancestral and younger generations. In addition, as genetic evaluation in practice covers traits that are expressed in either sex, it is recommended that genotyping covers animals from both sexes.

KEYWORDS

beef, genomic estimated breeding value, pedigree, SSGblup, meta-founder

1 Introduction

In livestock breeding programs, genomic selection (GS) is a method that uses genomic information to estimate breeding values and rank selection candidates. GS has shaped modern breeding programs and contributed substantially to the increase of genetic progress for various economically important traits, especially in dairy cattle (VanRaden et al., 2009; Meuwissen et al., 2016). The advantages of GS over traditional selection include; shorter generation intervals, increased selection intensity, greater selection accuracies, not limited to sex, and can be generalized to any trait that is recorded in the reference population (Schaeffer, 2006; Aguilar et al., 2010).

Genomic selection has a high potential for improving the genetic gain in beef cattle because reproduction, health, growth rate, meat quality, and feed efficiency are vital traits that contribute to the profitability of this industry, which are difficult and expensive to measure routinely (Van Eenennaam et al., 2011; Montaldo et al., 2012; Hayes et al., 2013). However, the accuracies of genomic breeding values for economic traits in beef cattle are low to moderate (Saatchi et al., 2011; Van Eenennaam et al., 2014). This is for two possible

reasons: i) the reference populations that have been assembled for beef cattle are generally smaller than those for dairy cattle, and there are fewer sires with highly accurate progeny tests in comparison with dairy cattle; and ii) unlike dairy cattle, where populations around the world are dominated by just a couple of breeds, there are numerous breeds of importance and even two subspecies (Bos *taurus* and B. *indicus*) in the beef industry (Hayes et al., 2013).

Genomic selection in beef cattle was first performed based on pseudo-data with multiple-step methods, such as estimated breeding value (EBV) or daughter yield deviation (VanRaden et al., 2009). This method needs many animals (hundreds of thousands) to be genotyped and have phenotypic measurements for the trait of interest to serve as the reference population. The reference population also needs to be updated, i.e., new animals with both phenotype and genotype need to be added. Although the multiple-step method is practical, it rests on several assumptions that are not met in all situations; for instance, it is impossible to genotype all animals. Also, the predicted accuracy using the multistep procedure is lower when compared to single-step BLUP. Also, the large number of breeds and crossbreds, poor extent of phenotyping, limited use of artificial insemination, less advanced structures and breeding programs, low number of offspring per female, incomplete relationships between identical traits in different countries, and limited data recording on economically important traits have resulted in limited adoption of GS in beef cattle (Goddard, 2009; Johnston et al., 2012; Van Eenennaam et al., 2014). Despite these difficulties, results of applying GS have been reported in some studies (Hayes et al., 2019; Wang et al., 2019; Zhu et al., 2019). All studies reported the benefits of applying GS in beef cattle and showed that GS could be a practical alternative to traditional selection approaches. Due to the mentioned limitations, the single-step genomic best linear unbiased prediction (ssGBLUP) method that combines all types of information (phenotype records, pedigree, genotypes) seems to work best in practical genetic evaluation in beef cattle. The main benefit of this method is that all animals in evaluation can get genomic-enhanced breeding values, even if not all have been genotyped (Misztal et al., 2009; Legarra et al., 2014).

Berry et al., 2016 reviewed prediction accuracy with the use of genomic data for some traits in beef cattle. However, guidelines to implement the method in practice are lacking. The main challenge in the implementation of GS for a breeding organization would be which and how many animals to be genotyped as the initial step. Even though, the cost of genotyping is not an obstacle nowadays, but contrary to dairy, the beef industry is much smaller, and genetic evaluation is performed on a much smaller scale in most countries. This is because beef production is highly influenced by the dairy sector with calves and cattle not required for dairy products being fattened to produce meat (Deblitz, 2008). Smaller industries can easily translate to the fact that breed associations and companies have much fewer resources to spend on genotyping. In the ideal situation, there may be possibilities to spend some funding on genotyping of the semen sires that are in service. However, a large proportion of genotyping costs for younger animals would be paid by the farms which traditionally are slower adopters of technology than dairy farmers. This could be due to a multitude of reasons, including the lower business margin.

TABLE 1 Parameters of the simulation process.

Population structure	
Step 1: Historical generations (HG)	
Number of generations phase 1	1,000
Size	1,000
Number of generations phase 2	200
Size	2020
Step 2: Expanded generations (EG)	
Number of founder males from HG	100
Number of founder females from HG	100
Number of generations	8
Number of offspring per dam	5
Selection and mating	Random
Step 3: Breed formation (BF)	
Number of males/females from BF for all 5 breeds	100/100
Number of generations	30
Number of offspring per dam	2
Selection and mating	Random
Step 4: Breeds A, B, C, D and E	
Number of males/females from A	220/1800
Sire replacement and growth rate	0.5065 0.072
Dam replacement and growth rate	0.30 0.098
Number of males/females from B	160/1,100
Sire replacement and growth rate	0.5851 0.1038
Dam replacement and growth rate	0.30 0.1629
Number of males/females from C	140/1,200
Sire replacement and growth rate	0.5252 0.073
Dam replacement and growth rate	0.30 0.103
Number of males/females from D	120/600
Sire replacement and growth rate	0.6256 0.118
Dam replacement and growth rate	0.30 0.182
Number of males/females from E	100/500
Sire replacement and growth rate	0.5392 0.06
Dam replacement and growth rate	0.30 0.117
Selection	High EBV
Mating system	Random
Number of generations	15
Number of offspring per dam	1
Genome	
Number of chromosomes	29
Number of SNPs	50,000
SNP distribution	Evenly spaced
Number of QTL	800
QTL distribution	Random
MAF of SNPs	0.1
MAF of QTL	0.1
Additive allelic effects for QTL	Gamma
Rate of recurrent mutation	2.5×10^{-5}

While genotyping of the ancestral sires in a sense that they have contributed much more than their contemporaries to the current generations seems logical, however, sustainable genetic gain in a breeding scheme needs accurate selection in younger generations as well. In addition, similar to dairy cattle, genetic progress in a breeding scheme in beef is not only driven by bulls but also

TABLE 2 Number of male and female animals with genotyping record in each scenario according to breeds.

[a]Scenarios	Sex	Breeds					
		A	B	C	D	E	Total
Sc. 1	Male progeny	1,349	1,317	1,087	759	488	5,000
Sc. 2	Ancestral sires	621	463	382	305	229	2000
	Male progenies	780	811	647	468	294	3,000
Sc. 3	Male selection candidates	648	670	545	394	239	2,496
	Female selection candidates	654	665	547	402	236	2,504
Sc. 4	Ancestral sires	464	305	303	252	176	1,500
	Ancestral dams	450	386	310	228	126	1,500
	Male selection candidates	249	258	198	167	116	988
	Female selection candidates	259	247	234	175	97	1,012

[a]Sc. 1: 5,000 randomly selected male progenies from 15th generation were genotyped, Sc. 2: 2000 ancestral sires with more than 10 progenies and 3,000 randomly selected male progenies from 15th generation were genotyped, Sc. 3: 5,000 selection candidates (both males and females) from 15th generation were genotyped, SC 4: randomly selected 1,500 ancestral sires, 1,500 ancestral dams and 2000 selection candidates (both males and females) from 15th generation were genotyped.

depends on the superiority of the dams of the candidates. Based on this, we hypothesized that prediction accuracy in selection candidates might differ when genotyping is only on ancestors, younger generation, and is restricted for males or females. Thus, the main aim of this study was to evaluate genotyping strategies in terms of prediction accuracy as the first step in the implementation of GS in beef. In particular, the goal was to compare and to contrast the importance of genotyping the ancestors and distributing the genotyped individuals for the two sexes on the predicting accuracy. The study was conducted under the assumption that a number of animals being genotyped is the main constraint.

2 Materials and methods

2.1 Definition of the population structure

Using QMSim software, a historical population of beef cattle was simulated based on the forward-in-time process (Sargolzaei and Schenkel, 2009). In total, 2020 generations were considered for the historical population. For the first 1,000 generations, the population size (n = 1,000) was constant and gradually decreased to 200 individuals to generate linkage disequilibrium (LD) during generations 1,001 to 2020.

In the second step, to enlarge the base population, 100 Founder males and 100 Founder females were selected randomly from the last generation of the historical population (Expanded generations) and were mated randomly for another eight generations. In the third step (Breed formation), five random samples, as the base for five breeds (A-E), were randomly chosen from the last generation of the previous step. In this step, mating and selection were also random within each breed, producing two offspring per dam for 30 generations. In the last step (step 4), population structure was simulated to mimic the production and genetic evaluation system in practice such that parameters were chosen to be realistic five breeds with different sizes were simulated for 15 generations. Selection in all breeds was based on EBVs using pedigree-based BLUP, and the culling of animals was based on age. It was assumed that pedigree was available for all breeds without error, and base animals in this step were considered as meta founders (i.e., one meta-founder per breed). Sire and dam replacement ratio was different across breeds (Table 1).

2.2 Scenarios

Five scenarios were compared in terms of prediction accuracy of selection candidates in the last generation (15th generation). Scenarios included a reference scenario (Ref. Sc) without genotypic data and scenarios with both genotypic and phenotypic information (Sc. 1 to Sc. 4). In all scenarios, it was assumed that phenotypic and pedigree data were available for the last 15 generations, and animals could be genotyped from the 7th generation onward. In all genomic scenarios, 5,000 animals could be genotyped. Genotyping scenarios differed in the method applied for the selection of 5,000 animals to be genotyped. In Sc. 1, the genotyping strategy was focused on young animals and only male progenies from the pool of selection candidates of generation 15 were selected randomly. In Sc. 2, genotyping was only on males but both young animals and ancestral sires could be genotyped. The criteria for selection of ancestral sires were that a sire should have at least 10 progenies in the population to be selected for genotyping. In Sc. 3, both male and female progenies from generation 15 could be genotyped. In Sc. 4, randomly selected ancestral sires, ancestral dams, and selection candidates (both males and females) were selected to be genotyped. The number of male and female animals with genotypic record in each scenario across breeds is presented in Table 2.

In all scenarios, irrespective of the availability of genotypic data, two other factors were considered to evaluate their impact on the prediction accuracy of the selection candidates. The first factor was the number of phenotypic records which could be collected and used

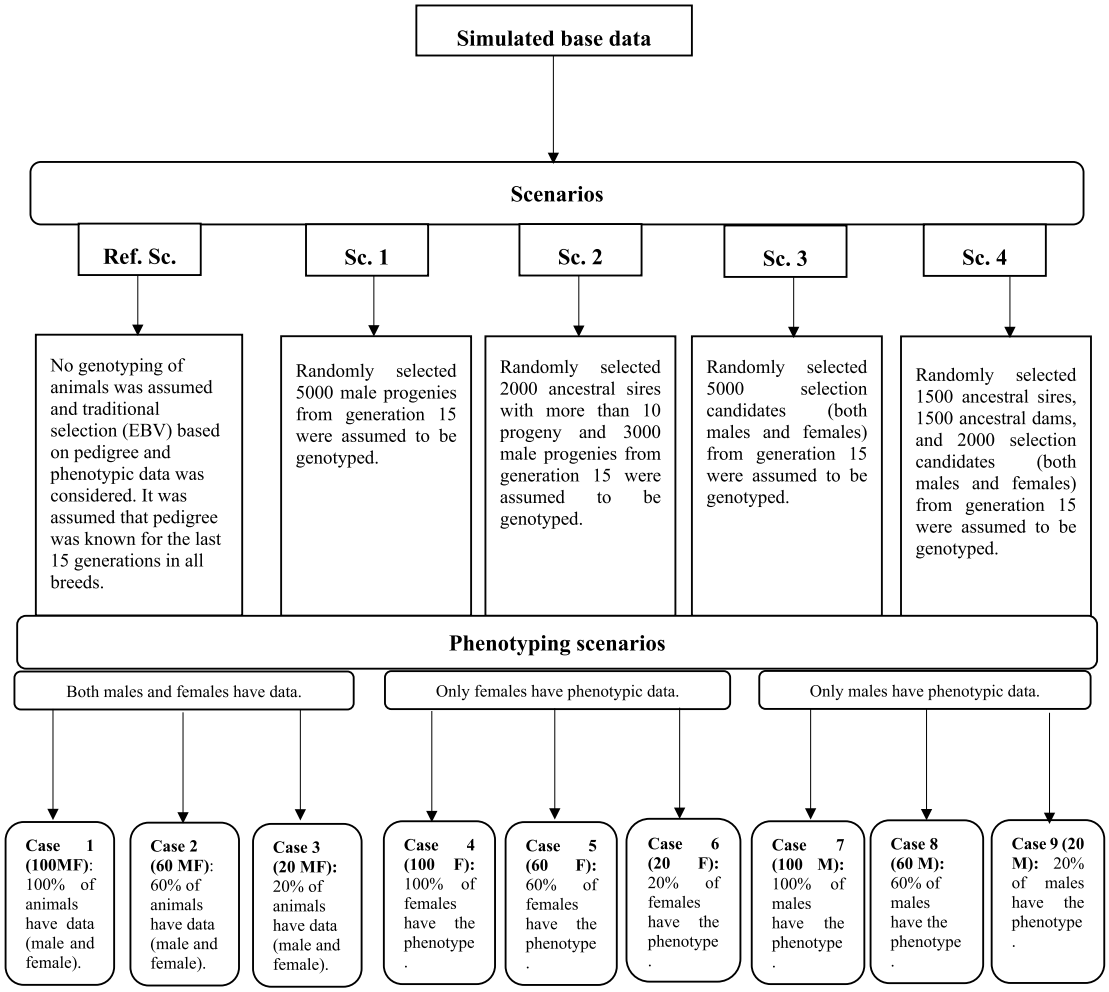


FIGURE 1
Schematic representation of the genotyping and phenotyping scenarios.

for the genetic evaluation. It was assumed that 20, 60, and 100 percent of the animals could have phenotypic observations (cases 1–3: 100 MF, 60 MF, and 20 MF). This phenotyping scenario was considered to cover a range of traits such as birth weight, where data are collected routinely in practice, and a scarcely recorded trait such as meat quality, where normally 20% of animals in the evaluation would have records available in practice. The second factor was the simulation of a sex-limited trait where either males (e.g., scrotal circumference) (cases 3–6: 100 M, 60 M, and 20 M) or females (cases 6–9: 100 F, 60 F, and 20 F) could have phenotypic records. A schematic representation of the genotyping and phenotyping scenarios is in Figure 1.

2.3 Genome architecture

A genome consisting of 29 pairs of chromosomes with a total length of 2,319 cM was simulated. For each animal, single nucleotide polymorphisms (SNPs) markers with the density of 50 K and n = 800 QTL were considered. Both SNPs and QTL were selected from the segregating loci of the last generation of the historical population with a Minor Allele Frequency (MAF) of greater than 0.1 and were randomly spaced across the genome. Recurrent mutation rate of 2.5×10^{-5} for both marker and QTL was considered. The additive allelic effect for each QTL was sampled from gamma distribution with shape parameters equal to 0.4 (Table 1).

2.4 Simulation of phenotypes and GEBV

A single trait with a heritability of 0.3 and phenotypic variance of 1.0 was simulated. The True Breeding Values (TBV) for each animal were calculated as follows:

$$TBV_k = \sum_{j=1}^{nQTL} \beta_j . Q_{k_j} \quad (1)$$

Where β_j is the additive effect of QTL j, Q_{k_j} is the QTL genotype at locus j, coded as 0, 1, or 2, as the number of copies of a specified QTL allele is carried by an individual (k). The phenotypes (y_i) were simulated by adding residual term sampled as $\varepsilon_i \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance.

2.5 Genetic evaluations

2.5.1 BLUP with unknown parent groups (UPGs)

For the reference scenario, a single-trait BLUP with UPG was used to estimate the breeding values. In this scenario, all five breeds were analyzed together in a multi-breed model. EBVs were estimated based on the model (2):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{ZQs} + \boldsymbol{e} \quad (2)$$

Where $\boldsymbol{y}$ is the vector of simulated phenotypes, μ is the constant average, $\boldsymbol{X}$ is the design matrix connecting records to fixed effects (sex, breed), $\boldsymbol{Z}$ is an incidence matrix relating animals to observations that connects animals to observations; $\boldsymbol{a}$ is the vector of random additive genetic effects; $\boldsymbol{Q}$ is a matrix that contains UPG compounds for all individuals; $\boldsymbol{s}$ is the vector of UPG effects, and $\boldsymbol{e}$ is the vector of random residuals. The trait of interest was considered to be the same across all breeds (i.e., rg = 1). Random effects were assumed to be independent and normally distributed:

$$\boldsymbol{a} \sim N\left(0, \boldsymbol{A}\sigma_a^2\right) \text{ and } \boldsymbol{e} \sim N\left(0, \boldsymbol{I}\sigma_e^2\right)$$

Where $\boldsymbol{A}$ is the numerator relationship matrix, $\boldsymbol{I}$ is the identity matrix, σ_a^2 is the direct additive genetic variance, and σ_e^2 is the residual variance. In this model, the EBV was

$$\boldsymbol{u} = \boldsymbol{Qs} + \boldsymbol{a} \quad (3)$$

where $\boldsymbol{u}$ was the total EBV, including UPG effects.

2.5.2 ssGBLUP with meta-founder

For genomic scenarios (Sc. 1–Sc. 4), the ssGBLUP with meta-founder was used. In the meta-founder approach, a modified $(\boldsymbol{H}^{\Gamma})^{-1}$ is substituted for the traditional $\boldsymbol{H}^{-1}$ matrix (Christensen et al., 2015; Legarra et al., 2015)

$$\left(\boldsymbol{H}^{\Gamma}\right)^{-1} = \left(\boldsymbol{A}^{\Gamma}\right)^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \left(\boldsymbol{G}^{\Gamma}\right)^{-1} - \left(\boldsymbol{A}_{22}^{\Gamma}\right)^{-1} \end{pmatrix} \quad (4)$$

Where $(\boldsymbol{H}^{\Gamma})^{-1}$ is the inverse of the realized relationship matrix with meta-founder, $\boldsymbol{A}^{\Gamma}$ is pedigree relationship matrix formed with a $\boldsymbol{\Gamma}$ matrix, $\boldsymbol{A}_{22}^{\Gamma}$ is a submatrix of $\boldsymbol{A}^{\Gamma}$ for the genotyped animals, and $\boldsymbol{G}^{\Gamma}$ is genomic relationship matrix with meta-founder constructed as:

$$\boldsymbol{G}^{\Gamma} = \frac{\boldsymbol{ww}'}{s} \quad (5)$$

Where $\boldsymbol{w}$ is the incidence matrix with elements of 1, 0 and −1 for AA, Aa, and aa, respectively; s is the half of the number of markers.

Matrix $\boldsymbol{\Gamma}$ represents within and across population relationship matrix. The structure of variance-covariance of meta-founder was estimated as $\boldsymbol{\Gamma} = 8Cov(\boldsymbol{P})$, according to the method presented by Christensen et al. (2015), where $\boldsymbol{P}$ is a matrix with $\boldsymbol{m}$ columns (m = total number of meta-founder) and n rows (n = total number of markers), containing the frequency of the second allele per breed.

The genetic evaluation analysis was performed under the restricted maximum likelihood (REML) approach using an animal model in the BLUPF90 family software (Misztal et al., 2015). The prediction accuracy in each scenario was computed as the correlation between TBV and (G)EBV in the 15th generation.

TABLE 3 Prediction accuracy across scenarios under different phenotyping scenarios.

Cases	Scenarios				
	Ref. Sc	Sc. 1	Sc. 2	Sc. 3	Sc. 4
100 MF	0.34	0.45	0.49	0.39	0.50
60 MF	0.25	0.36	0.43	0.28	0.44
20 MF	0.14	0.21	0.28	0.19	0.34
100 F	0.24	0.35	0.40	0.27	0.41
60 F	0.21	0.30	0.38	0.27	0.35
20 F	0.19	0.24	0.32	0.20	0.38
100 M	0.27	0.36	0.42	0.36	0.45
60 M	0.25	0.30	0.39	0.31	0.42
20 M	0.17	0.26	0.34	0.26	0.39

Details about cases, and scenarios are in Figure 1.

3 Results

In all scenarios, the use of genomic information irrespective of phenotyping strategy, increased the prediction accuracy compared to the Ref. Sc (Table 3). The range of prediction accuracy in Ref. Sc scenario was between 0.14 and 0.34 and for GS scenarios between 0.19 and 0.50. The average prediction accuracy across scenarios were 0.23, 0.31, 0.38, 0.28 and 0.41 for Ref. Sc and Sc. 1 to 4, respectively. Among the GS scenarios, Sc. 4 had the highest accuracy across phenotyping strategies, followed by Sc. 2. In both Sc. 4 and Sc. 2, ancestral animals with contributions to the population (i.e., had some progenies) were genotyped in addition to selection candidates. For scenarios where genotyping was limited to the young selection candidates (Sc. 1 and Sc. 3), prediction accuracy was lower than the scenarios where genotyping was on animals from young and older generations.

Figure 2 shows the prediction accuracy in males and females. Nearly in all scenarios and cases, males were predicted more accurately than females. Even when only 20% of males had a phenotypic record, the accuracy is higher than when 20% of male and female animals or only 20% of female animals had a phenotypic record, which shows the more significant effect of male phenotypic records on the prediction accuracy.

As expected within each scenario, higher percentage of phenotype availability, resulted in higher prediction accuracy. The trend was similar for all scenarios and irrespective of the availability of records on both or either of sex. Results also show that when the trait of interest could be measured on both sex (Cases 100, 60, 20 MF), on average prediction accuracy was higher than when trait was sex-limited. For sex-limited traits, prediction accuracy was similar whether it was measured on males or females, however, mean accuracy were slightly higher when trait of interest was measured on male animals (0.34 in females and 0.36 in males).

Table 4 shows the prediction accuracy for animals without phenotypic records in different cases for selection candidates. The aim would be to realize how genotyping strategy would

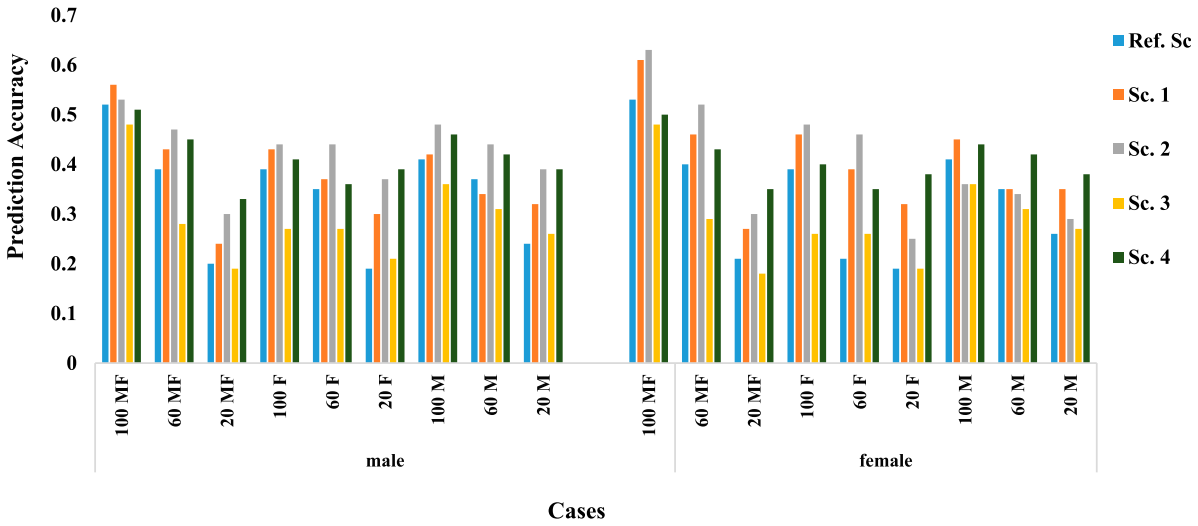


FIGURE 2
Prediction accuracy in males and females.

TABLE 4 Prediction accuracy for animals without phenotypic records in males and females according to cases.

Sex	Cases	Number of animals	Accuracy in scenarios				
			Ref. Sc	Sc. 1	Sc. 2	Sc. 3	Sc. 4
Male	100 MF	—	—	—	—	—	—
	60 MF	5,330	0.39	0.43	0.46	0.28	0.44
	20 MF	5,445	0.20	0.24	0.29	0.19	0.33
	100 F	6,797	0.38	0.43	0.44	0.27	0.41
	60 F	2,719	0.53	0.37	0.43	0.27	0.35
	20 F	5,437	0.18	0.29	0.37	0.20	0.38
	100 M	—	—	—	—	—	—
	60 M	2,719	0.36	0.34	0.43	0.31	0.42
	20 M	5,437	0.24	0.31	0.37	0.26	0.38
Female	100 MF	—	—	—	—	—	—
	60 MF	2,686	0.40	0.45	0.51	0.28	0.43
	20 MF	5,452	0.21	0.26	0.29	0.18	0.35
	100 F	—	—	—	—	—	—
	60 F	2,719	0.20	0.38	0.46	0.25	0.35
	20 F	5,437	0.18	0.31	0.25	0.19	0.37
	100 M	6,797	0.40	0.45	0.36	0.35	0.44
	60 M	2,719	0.35	0.35	0.33	0.31	0.41
	20 M	5,437	0.25	0.34	0.28	0.26	0.38

Note that in 100 MF, 100% of animals have phenotypic records and as a result accuracy was not calculated for this case. Details about cases, and scenarios are in Figure 1.

affect prediction accuracy in animals without records in the last generation. Prediction accuracy in males without records was highest based on Sc. 2 and lowest based on Sc. 3 with an average of 0.41 and 0.24, respectively. Prediction accuracy for females without records was highest in Sc. 4 and Sc. 3 with an average of 0.41 and 0.28, respectively. Note that prediction accuracy for some

phenotyping cases is not presented as either all animals had phenotypic records (Case 100 MF) or trait was sex limited (Case 100 M and Case 100 F). (The solution for UPG was similar among breeds and Supplementary Table S1 shows mean solution for UPG each breed across cases.)

4 Discussion

We investigated the potential of applying GS in beef cattle when the aim was improving prediction accuracy in selection candidates. Four selective genotyping scenarios were compared to traditional pedigree-based evaluation. Results showed fair improvement in prediction accuracy, albeit a limited number of animals being genotyped.

One challenge with applying GS in practice is that there are many selection candidates, and genotyping all of them is often impractical. Genotype scenarios should only genotype small proportions of selected candidates welcomed by breeders because results have shown that significant investments in the genotype of selected candidates are not necessary to take full advantage of the benefits of genomic selection (Pryce and Daetwyler, 2012; Howard et al., 2018). As the results showed that the determination of the genotype of only the selected candidates, regardless of gender, has the lowest prediction accuracy (Sc. 3). Our study investigated the importance and effect of genotyping and phenotyping scenarios on prediction accuracy. The results showed that genotyping of both the selection candidates), and male and female ancestors, could be used to maximize the advantage of genomic selection (Sc. 4). In fact, Sc. 4 confirms the more significant effect of the female genotype than the male genotype and the more significant influence of the ancestral genotype than the selected candidates on the prediction accuracy (because they have more offspring, more information). The lower effect of male genotypes on the accuracy of predictions in this study can be allocated to the effect of the pedigree relationship between individuals. This makes it difficult to make accurate sire selection decisions (Nwogwugwu et al., 2020). In addition, in beef cattle, the offspring are smaller per male and more significant in each female than in dairy cattle. Determining the female genotype in beef cattle can significantly contribute to genetic accuracy and development. As a result, selective genotyping of only part of the selection candidates, males and females of the ancestors, can increase the prediction accuracy. In addition, we can make the most of the benefits of genome selection while saving on genotype costs.

Natural mating of multiple sires is the most common mating system in beef cattle production, despite the management advantages of this mating system, it does not allow for identification the paternity of the progeny (Tonussi et al., 2017). So, because the information of the cows is known, the genotype of females is easier Given that the genotypic data of females are usually more available than males (Mrode, 2019), accuracy can be increased by increasing the genotypic data of females (Tsuruta et al., 2013). Various studies have examined the effect of female genotype on prediction accuracy, including a study using a multi-step method that reported a decrease in accuracy using female genotype (Wiggans et al., 2011). However, in our study using the ssGBLUP approach, accuracy was increased by including the female genotype, and also consistent with the results reported by Tsuruta et al., 2013; Lourenco et al., 2014. In addition, in all genomic scenarios, by examining prediction accuracy in males and females separately, it can be concluded that increasing the genotypic information of females, the prediction accuracy increases (Figure 2).

In practical situation, all animals being evaluated rarely have phenotypic information, and the records are not widely available for traits such as disease and meat quality. To address this issue, we considered different phenotypic scenarios in our simulation. In all studied scenarios, with decreasing phenotypic records, prediction accuracy also decreases, which indicates a direct relationship between phenotypic records and prediction accuracy (Goddard, 2009; Takeda et al., 2020).

Genomic prediction in beef cattle provides accuracy higher than the average of parents based on the pedigree of selected candidates. It can be equivalent to progeny tests based on a maximum of 10 offspring (Garrick, 2011). In addition, ssGBLUP is superior to traditional evaluation methods because ssGBLUP uses phenotypes instead of pseudo-phenotypes and considers the entire population structure for GEBV estimation (Lourenco et al., 2014). This can be used for beef cattle selection in that only a tiny proportion of animals have pedigree and genotype. Estimated breeding value assessment with BLUP depends on the phenotype, parents, and progeny. But the ssGBLUP method is less sensitive to scenarios where animals selectively genotyped and, or genomic preselection exists compared to the multiple-step methods. Hence, ssGBLUP in conventional evaluations is attractive (Masuda et al., 2018). The ssGBLUP method is conceptually and practically simpler than the multiple-step GBLUP method, and in addition, it does not have shortcomings such as bias and loss of information of with few progenies, as well as operational complexity (Christensen and Lund, 2010; Legarra et al., 2014). Therefore, the ssGBLUP method is simpler and applicable to complex models and is generally as accurate as multiple-step methods (VanRaden, 2012; Mehrban et al., 2019). Also, Due to the lower sensitivity of the ssGBLUP method to genotyping scenarios, this method can be used to determine the best genotyping scenario to reduce genotype costs while increasing accuracy (Howard et al., 2018).

In our study, genetic evaluation was based on multiple breeds information. When several breeds are combined in one assessment, there is generally no pedigree information among breeds. As a result, UPG (Quaas, 1988) has been developed to model missing pedigrees and to explain breed differences in multi-breed evaluations(Legarra et al., 2007; VanRaden et al., 2007). However, UPG solutions, when evaluated with the ssGBLUP model, may be biased due to genomic incompatibility ($\boldsymbol{G}$) and pedigree-based relationship ($\boldsymbol{A}$) matrices due to the lack of genotypes of all animals in the pedigree (Misztal and Legarra, 2017; Kudinov et al., 2020). Legarra et al. (2015) developed a meta-founders theory to solve this problem and consider the relationships within and between the founding population. Several studies have reported improved genetic evaluation performance using meta-founder (Bradford et al., 2019; Junqueira et al., 2020). Accordingly, we used the method to account for the genetic level of base animals of each breed in step 4.

The pedigrees used in genetic evaluations may go back to a few base populations thought to be unrelated due to a lack of access to information (Legarra et al., 2015). In addition, information is not available at the beginning of the pedigree, and animals of several generations may have missing pedigree information (Tsuruta et al., 2019). Also, populations are selected, and animals with missing parents are unlikely to be chosen as parents of the next-generation

because their breeding value is reduced to zero (Legarra et al., 2015; Kluska et al., 2021). Unknown parent groups and Meta-Founder can be used to calculate missing pedigree and breed structure in multi-breed populations such as beef cattle (Kluska et al., 2021).

We investigated the potential of applying GS in beef cattle when the aim was improving prediction accuracy in selection candidates. Four selective genotyping scenarios were compared to traditional pedigree-based evaluation. Results showed fair improvement in prediction accuracy, albeit a limited number of animals being genotyped. Comparison of GS scenarios revealed that selective genotyping should be on animals from both ancestral and younger generations. In addition, as genetic evaluation in practice cover traits that are expressed on either sex, it is recommended that selective genotyping covers animals from both sexes as well.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding authors.

Author contributions

SR and HE conceived and designed this simulation. ME performed the simulation and data analysis. SR and HE participated in the simulation and data analysis. ME drafted the article. All authors participated in editing and approved the final version.

Conflict of interest

HE was employed by the company Norwegian Beef Cattle Organizations, TYR.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1083106/full#supplementary-material

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J., et al. (2010). *Hot topic*: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *J. Dairy Sci.* 93 (2), 743–752. doi:10.3168/jds.2009-2730

Berry, D. P., Garcia, J. F., and Garrick, D. J. (2016). Development and implementation of genomic predictions in beef cattle. *Anim. Front.* 6 (1), 32–38. doi:10.2527/af.2016-0005

Bradford, H. L., Masuda, Y., VanRaden, P. M., LegArrA, A., and Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. *J. Dairy Sci.* 102 (3), 2336–2346. doi:10.3168/jds.2018-15434

Christensen, O. F., Legarra, A., Lund, M. S., and Su, G. (2015). Genetic evaluation for three-way crossbreeding. *Genet. Sel. Evol.* 47 (1), 98–13. doi:10.1186/s12711-015-0177-6

Christensen, O., and Lund, M. (2010). Genomic relationship matrix when some animals are not genotyped. *Genet. Sel. Evol.* 42 (2), 1–8.

Deblitz, C. (2008). Brömmer and D. Brüggemann/Landbauforschung-vTI agriculture and forestry research. *Landbauforschung* 1 (2), 29–44.

Garrick, D. J. (2011). The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genet. Sel. Evol.* 43, 17. doi:10.1186/1297-9686-43-17

Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136, 245. doi:10.1007/s10709-008-9308-0

Hayes, B. J., Corbet, N. J., Allen, J. M., Laing, A. R., Fordyce, G., Lyons, R., et al. (2019). Towards multi-breed genomic evaluations for female fertility of tropical beef cattle. *J. Animal Sci.* 97 (1), 55–62. doi:10.1093/jas/sky417

Hayes, B. J., Lewin, H. A., and Goddard, M. E. (2013). The future of livestock breeding: Genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends Genet.* 29 (4), 206–214. doi:10.1016/j.tig.2012.11.009

Howard, J. T., Rathje, T. A., Bruns, C. E., Wilson-Wells, D. F., Kachman, S. D., and Spangler, M. L. (2018). The impact of selective genotyping on the response to selection using single-step genomic best linear unbiased prediction. *J. Animal Sci.* 96 (11), 4532–4542. doi:10.1093/jas/sky330

Johnston, D. J., Tier, B., and Graser, H. U. (2012). Beef cattle breeding in Australia with genomics: Opportunities and needs. *Animal Prod. Sci.* 52 (2–3), 100–106. doi:10.1071/AN11116

Junqueira, V. S., Lopes, P. S., Lourenco, D., Silva, F. F. E., and Cardoso, F. F. (2020). Applying the metafounders approach for genomic evaluation in a multibreed beef cattle population. *Front. Genet.* 11 (12), 556399. doi:10.3389/fgene.2020.556399

Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., et al. (2021). Metafounders may reduce bias in composite cattle genomic predictions. *Front. Genet.* 12 (8), 678587. doi:10.3389/fgene.2021.678587

Kudinov, A. A., Mantysaari, E. A., Aamand, G. P., Uimari, P., and Stranden, I. (2020). Metafounder approach for single-step genomic evaluations of Red Dairy cattle. *J. Dairy Sci.* 103 (7), 6299–6310. doi:10.3168/jds.2019-17483

Legarra, A., Christensen, O. F., Aguilar, I., and Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livest. Sci.* 166 (1), 54–65. doi:10.1016/j.livsci.2014.04.029

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., and Misztal, I. (2015). Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. *Genetics* 200 (2), 455–468. doi:10.1534/genetics.115.177014

Legarra, A., Strabel, T., Sapp, R. L., Sanchez, J. P., and Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population. *J. Animal Breed. Genet.* 124 (5), 286–295. doi:10.1111/j.1439-0388.2007.00671.x

Lourenco, D. A. L., Misztal, I., TSuruta, S., AguIlar, I., Ezra, E., Ron, M., et al. (2014). Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *J. Dairy Sci.* 97 (3), 1742–1752. doi:10.3168/jds.2013-6916

Masuda, Y., VanRaden, P. M., Misztal, I., and Lawlor, T. J. (2018). Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *J. Dairy Sci.* 101 (6), 5194–5206. doi:10.3168/jds.2017-13310

Mehrban, H., Lee, D. H., Naserkheil, M., Moradi, M. H., and Ibáñez-Escriche, N. (2019). 'Comparison of conventional BLUP and singlestep genomic BLUP evaluations for yearling weight and carcass traits in Hanwoo beef cattle using single trait and multi-trait models. *PLoS ONE* 14, e0223352. doi:10.1371/journal.pone.0223352

Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* 6 (1), 6–14. doi:10.2527/af.2016-0002

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92 (9), 4648–4655. doi:10.3168/jds.2009-2064

Misztal, I., and Legarra, A. (2017). Invited review: Efficient computation strategies in genomic selection. *Animal* 11 (5), 731–736. doi:10.1017/S1751731116002366

Misztal, I., Tsuruta, S., Aguilar, I., Legarra, A., and Vitezica, Z. (2015). *BLUPF90 family of programs*. Athens, USA: University of Georgia, 1–125. Available at: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf.

Montaldo, H. H., Casas, E., Ferraz, J. B. S., Vega-Murillo, V. E., and Roman-Ponce, S. I. (2012). Opportunities and challenges from the use of genomic selection for beef cattle breeding in Latin America. *Anim. Front.* 2 (1), 23–29. doi:10.2527/af.2011-0029

Mrode, R. (2019). Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: Current status and future prospects. *Front. Genet.* 9, 694. doi:10.3389/fgene.2018.00694

Nwogwugwu, C. P., Kim, Y., Choi, H., Lee, J. H., and Lee, S. H. (2020). Assessment of genomic prediction accuracy using different selection and evaluation approaches in a simulated Korean beef cattle population. *Asian-Australasian J. Animal Sci.* 33 (12), 1912–1921. doi:10.5713/ajas.20.0217

Pryce, J. E., and Daetwyler, H. D. (2012). Designing dairy cattle breeding schemes under genomic selection: A review of international research. *Animal Prod. Sci.* 52 (2–3), 107–114. doi:10.1071/AN11098

Quaas, R. L. (1988). Additive genetic model with groups and relationships. *J. Dairy Sci.* 71 (5), 1338–1345. doi:10.3168/jds.S0022-0302(88)79691-5

Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., et al. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol.* 43 (1), 40–16. doi:10.1186/1297-9686-43-40

Sargolzaei, M., and Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics* 25 (5), 680–681. doi:10.1093/bioinformatics/btp045

Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *J. Animal Breed. Genet.* 123 (4), 218–223. doi:10.1111/j.1439-0388.2006.00595.x

Takeda, M., Uemoto, Y., and Satoh, M. (2020). Effect of genotyped bulls with different numbers of phenotyped progenies on quantitative trait loci detection and genomic evaluation in a simulated cattle population. *Animal Sci. J.* 91 (1), e13432–e13439. doi:10.1111/asj.13432

Tonussi, R. L., de Oliveira Silva, R. M., Magalhaes, A. F. B., Espigolan, R., Peripolli, E., Olivieri, B. F., et al. (2017). Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. *PLoS ONE* 12 (9), 01817522. doi:10.1371/journal.pone.0181752

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Misztal, I., and Lawlor, T. J. (2019). Controlling bias in genomic breeding values for young genotyped bulls. *J. Dairy Sci.* 102 (11), 9956–9970. doi:10.3168/jds.2019-16789

Tsuruta, S., Misztal, I., and Lawlor, T. J. (2013). Short communication: Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *J. Dairy Sci.* 96 (5), 3332–3335. doi:10.3168/jds.2012-6272

Van Eenennaam, A. L., van der Werf, J. H. J., and Goddard, M. E. (2011). The value of using DNA markers for beef bull selection in the seedstock sector. *J. Animal Sci.* 89 (2), 307–320. doi:10.2527/jas.2010-3223

Van Eenennaam, A. L., Weigel, K. A., Young, A. E., Cleveland, M. A., and Dekkers, J. C. M. (2014). Applied animal genomics: Results from the field. *Annu. Rev. Animal Biosci.* 2, 105–139. doi:10.1146/annurev-animal-022513-114119

VanRaden, P. M. (2012). Avoiding bias from genomic pre-selection in converting daughter information across countries. *Interbull Bull.* 45 (45), 1–5. Available at: https://journal.interbull.org/index.php/ib/article/view/1243%5Cnhttps://journal.interbull.org/index.php/ib/article/download/1243/1241.

VanRaden, P. M., Tooker, M. E., Cole, J. B., Wiggans, G. R., and Megonigal, J. H. (2007). Genetic evaluations for mixed-breed populations. *J. Dairy Sci.* 90 (5), 2434–2441. doi:10.3168/jds.2006-704

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92 (1), 16–24. doi:10.3168/jds.2008-1514

Wang, X., Miao, J., Chang, T., Xia, J., Li, Y., Xu, L., et al. (2019). Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese simmental beef cattle. *PLoS One* 14 (2), 02104422. doi:10.1371/journal.pone.0210442

Wiggans, G. R., Cooper, T. A., Vanraden, P. M., and Cole, J. B. (2011). Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. *J. Dairy Sci.* 94 (12), 6188–6193. doi:10.3168/jds.2011-4481

Zhu, B., Guo, P., Wang, Z., Zhang, W., Chen, Y., Zhang, L., et al. (2019). Accuracies of genomic prediction for twenty economically important traits in Chinese Simmental beef cattle. *Anim. Genet.* 50 (6), 634–643. doi:10.1111/age.12853

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 31 March 2023
DOI 10.3389/fvets.2023.1148070

OPEN ACCESS

EDITED BY
Zexi Cai,
Aarhus University, Denmark

REVIEWED BY
Zhixin Chai,
Southwest Minzu University, China
Wenlin Bai,
Shenyang Agricultural University, China

*CORRESPONDENCE
Zhijie Ma
✉ zhijiema@126.com
Chuzhao Lei
✉ leichuzhao1118@126.com

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

RECEIVED 19 January 2023
ACCEPTED 01 March 2023
PUBLISHED 31 March 2023

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

Analysis of genomic copy number variations through whole-genome scan in Chinese Qaidam cattle

Yangkai Liu[1,2,3], Yanan Mu[3], Wenxiang Wang[3], Zulfiqar Ahmed[4], Xudong Wei[1,2], Chuzhao Lei[3]* and Zhijie Ma[1,2]*

[1]Plateau Livestock Genetic Resources Protection and Innovative Utilization Key Laboratory of Qinghai Province, Academy of Animal Science and Veterinary Medicine, Qinghai University, Xining, China, [2]Key Laboratory of Animal Genetics and Breeding on Tibet Plateau, Ministry of Agriculture and Rural Affairs, Xining, China, [3]Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Xianyang, China, [4]Faculty of Veterinary and Animal Sciences, University of Poonch Rawalakot, Rawalakot, Pakistan

Qaidam cattle (CDM) are indigenous breed inhabiting Northwest China. In the present study, we newly sequenced 20 Qaidam cattle to investigate the copy number variants (CNVs) based on the ARS-UMD1.2 reference genome. We generated the CNV region (CNVR) datasets to explore the genomic CNV diversity and population stratification. The other four cattle breeds (Xizang cattle, XZ; Kazakh cattle, HSK; Mongolian cattle, MG; and Yanbian cattle, YB) from the regions of North China embracing 43 genomic sequences were collected and are distinguished from each of the other diverse populations by deletions and duplications. We also observed that the number of duplications was significantly more than deletions in the genome, which may be less harmful to gene formation and function. At the same time, only 1.15% of CNVRs overlapped with the exon region. Population differential CNVRs and functional annotations between the Qaidam cattle population and other cattle breeds revealed the functional genes related to immunity (*MUC6*), growth (*ADAMTSL3*), and adaptability (*EBF2*). Our analysis has provided numerous genomic characteristics of some Chinese cattle breeds, which are valuable as customized biological molecular markers in cattle breeding and production.

KEYWORDS

Qaidam cattle, whole genome resequencing, copy number variation (CNV), genome selection, population structure

1. Introduction

Domestic cattle are one of the important animals that have been used as a source of materials for production and development by human civilization. Approximately 850,000 years ago, domestic cattle diverged into two groups, namely, humpless taurine (*Bos Taurus*) and humped indicine (*Bos Indicus*) (1, 2). Moreover, environmental factors, geographical isolation, and human activities also contributed to the development of present-day cattle. Through a long period of domestication, megabases (Mb) of DNA gradually enriched the genomic diversity among cattle breeds (3).

As of 2021 (4), there are already 55 Chinese indigenous breeds. The Qaidam cattle (CDM) is one of the 55 breeds reared in Northwest China (36°21'−39°23' N, 90°30'−99°30' E, Qinghai Province, China), where the drought (annual precipitation < 200 mm) and high altitude (2,600–3,000 m) environment is predominant, and these conditions made the

Qaidam cattle breed have more stress resistance, rough feeding tolerance, and environmental adaptability. During the Yuan dynasty (AD 1271–1638) period, the Mongolian army introduced the Mongolian cattle (*Bostaurus*) into the present-day Qinghai and Gansu Provinces of China during a southward invasion, which might have influenced the breeding herds of the present Qaidam cattle. Paternal and maternal diversity studies indicated that Qaidam cattle included two lineages (1, 5). The autosomal genetic evidence suggests that the Qaidam cattle was closer to Mongolian cattle, which is a hybrid of *Bos Taurus* × *Bos Indicus* (1). The purebred Qaidam cattle have not been effectively protected for their low economic returns. There was a 47.80% decrease in the Qaidam cattle population by 2006 compared to the 1981 Qaidam cattle population (6).

The copy number variations (CNVs) are defined as the deletion or duplication of a genome copy number, ranging from 50 bp to several Mb in length (7). As compared to SNP mutations, the CNV fragments are large in length and cover a wider range of genomes that have broader prospects in studying animal genetics and breeding application. Recently, next-generation genome sequencing technologies have been continuously used to detect the genome-wide CNVs of livestock (8, 9). However, numerous genomic studies exploring CNVs in commercial cattle breeds have underestimated the role of native breeds in the adaptation process (10, 11).

In the present study, we performed a genome-wide CNV analysis using genomic resequencing data in six Chinese cattle breeds. The purpose was to generate a comprehensive CNV landscape in Qaidam cattle to investigate and compare the diversity and population–genetic properties of the CNV regions (CNVRs) among them and to explore the diverse selection patterns involved with the CNV genes for local adaptation in Chinese native cattle.

2. Materials and methods

2.1. Genome resequencing and samples collection

Qaidam Basin is the highest basin in China with an altitude of 2,600–3,000 m and is located in the northwest region of the Qinghai Province and the northeast region of the Qinghai–Tibet Plateau. The climate of the basin is characterized as extremely dry and cold, with an annual average precipitation of <200 mm and an annual average temperature of ∼3.0–6.5°C. To reflect the sample representativeness of the Qaidam cattle, 20 samples were collected from five different counties/cities (Dulan, Golmud, Mangya, Wulan, and Dachaidan) in the Qaidam Basin (Supplementary Table 1, Figure 2A).

The ear tissues of selected samples were used for DNA extraction by the standard phenol–chloroform protocol. Genomic DNA was constructed into 350-bp libraries and sequenced using Illumina NovaSeq at Novogene Bioinformatics Institute (Beijing, China). Moreover, 42 publicly available data of four Chinese cattle breeds were downloaded in this study (10 Mongolian cattle, MG; 9 Xizang cattle, XZ; 15 Yanbian cattle, YB; and 8 Kazakh cattle, HSK) (Supplementary Table 2). It is worth noticing that the resequencing data of one Xizang cattle (Sample ID: Xizang9) was offered by the Key Laboratory of Animal Genetics, Breeding and Reproduction of Northwest A&F University (Supplementary Table 2).

2.2. Genome data generation and CNV calling

Read pairs were aligned to the *B. taurus* reference assembly (ARS-UCD1.2) using the Burrows–Wheeler Aligner (BWA) program with default parameters (12). Then, CNVcaller (13) was applied to call the CNV in each individual. First, to create a *B. taurus* reference database, the ARS-UCD1.2 was split and the overlapping windows were recommended as 800 bp (13). Second, the reads number in each window was calculated, and high similarity (≧97%) reads were merged into segments of the autosomes. Third, the GC bias was used to standardize the copy number in each window, and it was used to classify the different genotypes of each sample. Finally, various steps of CNVcaller filtering parameters were carried out: -f 0.1 -h 3 -r 0.1; a Silhouette score of > 0.6; the length of CNVR of ≤50 kb (deletion and both), with the length of CNVR of < 500 kb (duplication) (15).

2.3. Breed/population differentiation

Principal component analysis (PCA) was used to stratify and cluster the close breeds/populations, which plays a positive role in understanding the genetic differences among cattle subpopulations. According to the smartPCA module of EIGENSOFT (Program 2006), the PCA calculation was performed based on the four different CNVR datasets.

2.4. Differential CNVR identification

We calculated V_{ST} (15) between Qaidam cattle and the other four cattle breeds (XZ, YB, HSK, and MG) to identify the differential CNVRs. The V_{ST} is a method to calculate selection between populations similar to the F_{ST} method. The formula is $V_{ST} = (V_T - V_S)/V_T$, where V_T represents the variance among all the unrelated individuals and V_S is the average variance within each population, weighted for population size (16). Finally, the top 1% gene cluster of the V_{ST} method was kept out by the outlier method.

The ANNOVAR was applied to annotate the CNVRs in our results (14). Further, the Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene ontology (GO) analysis were performed on the candidate CNV genes by KOBAS 3.0. Since the enriched terms were retained with a p-value of < 0.05, we preferred showing some of the top pathways; for more information, please see Supplementary Tables.

3. Results

3.1. CNV discovery and CNVR set statistics

We collected 63 Chinese cattle whole genomes representing five breeds from the north and northwest regions of China,

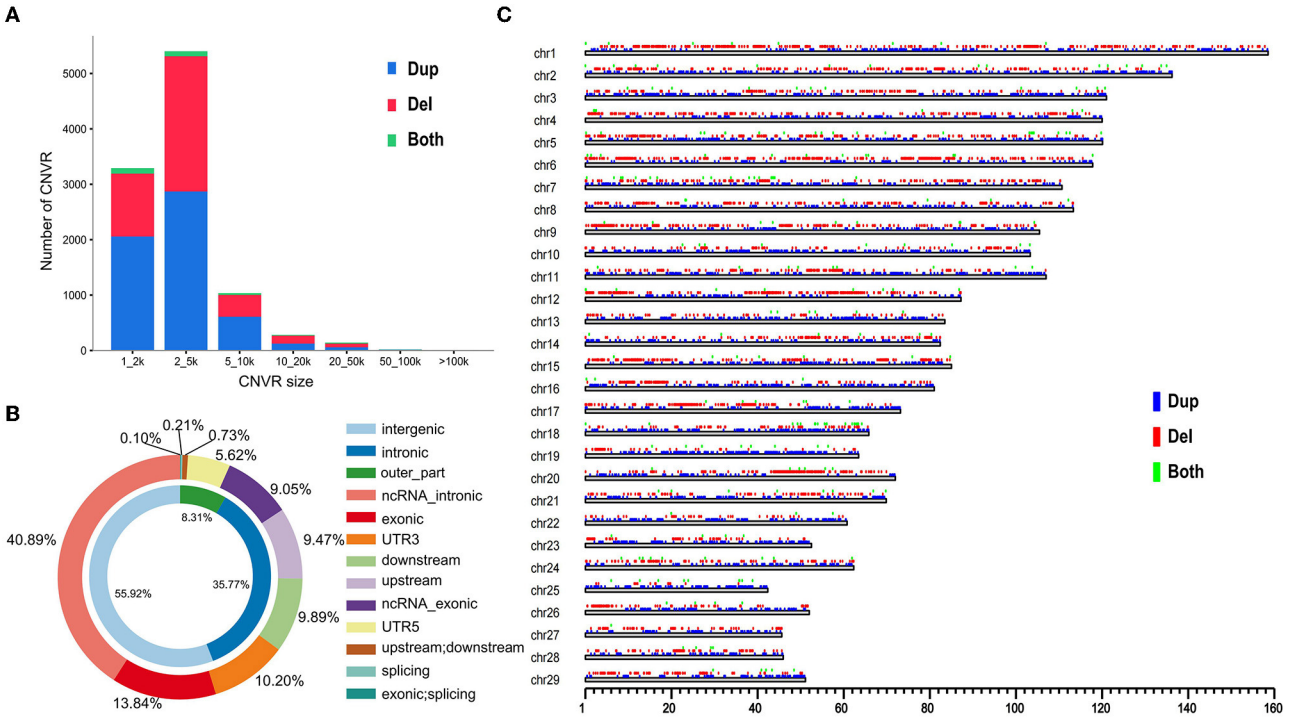


FIGURE 1
Genomic diversity and distribution of CNVRs. **(A)** The number of the detected CNVR. **(B)** Annotation of CNVRs with various genomic features. The inner circle indicates the intronic region, intergenic region, and the remaining set of functional regions (outer_part). The outer circle includes ncRNA intronic; exonic; UTR3; downstream; upstream; ncRNA exonic; UTR6; splicing; upstream and downstream; exonic and splicing. **(C)** The autosomal distribution of CNVRs. The location with different colors represent duplication (blue), deletion (red), and both duplication and deletion (green).

including 20 Qaidam cattle, 10 Mongolian cattle, 15 Yanbian cattle, nine Kazakh cattle, and nine Xizang cattle (Figure 2A, Supplementary Table 2). The mean sequencing depth was performed to 12-fold coverage of the *Bos taurus* genome (Supplementary Table 2). Among the 63 genomes, we newly sequenced 20 Qaidam samples and one Xizang sample at ~9-fold coverage each (Supplementary Table 2), and the other 42 genomic sequences were available online.

We applied a read–depth-based bio-software (CNVcaller) to discover autosomal CNVs among individuals relative to the ARS-UCD1.2 reference genome. We generated the CNVR datasets from each cattle breed. The CNVR set contained 10,178 CNVRs, which were detected from 63 cattle genome datasets. There were 5,743 duplication CNVRs; 4,187 deletion CNVRs; and 248 both duplication and deletion CNVRs (Supplementary Table 3). Here, 10,178 CNVRs (duplication, deletion, and both duplication and deletion) were divided into different length groups (Figure 1A). The CNVRs annotation showed that the number of CNVRs was 5,398 (53.04%), which were detected in 2–5 kb size. It was observed that 55.92% CNVRs were located in the intergenic region followed by the intron region (35.77%). However, only 1.16 % CNVRs were detected in the coding exonic region (Figure 1B). And the CNVRs distribute randomly in the chromosome both in number and length (Figure 1C).

3.2. Population structure

With the effect of balancing selection, abundant polymorphisms of the genomic copy number variation are found in Chinese *Bos taurus*. A principal component analysis (PCA) was carried out with an obvious distinction from deletions (Figure 2B), duplications (Figure 2C), and total CNVRs datasets (Figure 2D). Qaidam cattle population is broadly distinguished from Mongolian, Kazakh, and Xizang breeds and closely clustered with the Yanbian breed. The PC1 explained ~30.13–50.86% of the genetic variation. For deletions, PC1 (30.13% of the variance) could separate Qaidam and Yanbian breeds from the other breeds, and PC2 (3.52% of the variance) could distinguish Xizang cattle (Figure 2B) from the other breeds (Kazakh and Mongolian cattle). Compared to deletions, duplications separated Qaidam cattle from all other breeds, in general, as shown in the PCA, but its clustering had less accuracy (Figure 2C). The effect of the PCA using both CNVR types was not optimistic in the clustering populations (Supplementary Figure 1). Interestingly, Kazakh and Mongolian cattle populations showed greater separation within these breeds by duplication. Unlike Qaidam, Xizang, and Yanbian cattle, Kazakh and Mongolian cattle may have less pressure of selection, which caused numerous meaningless duplications. These data suggest that artifical selection has shaped the CNVR diversity of each cattle breed during animal domestication.

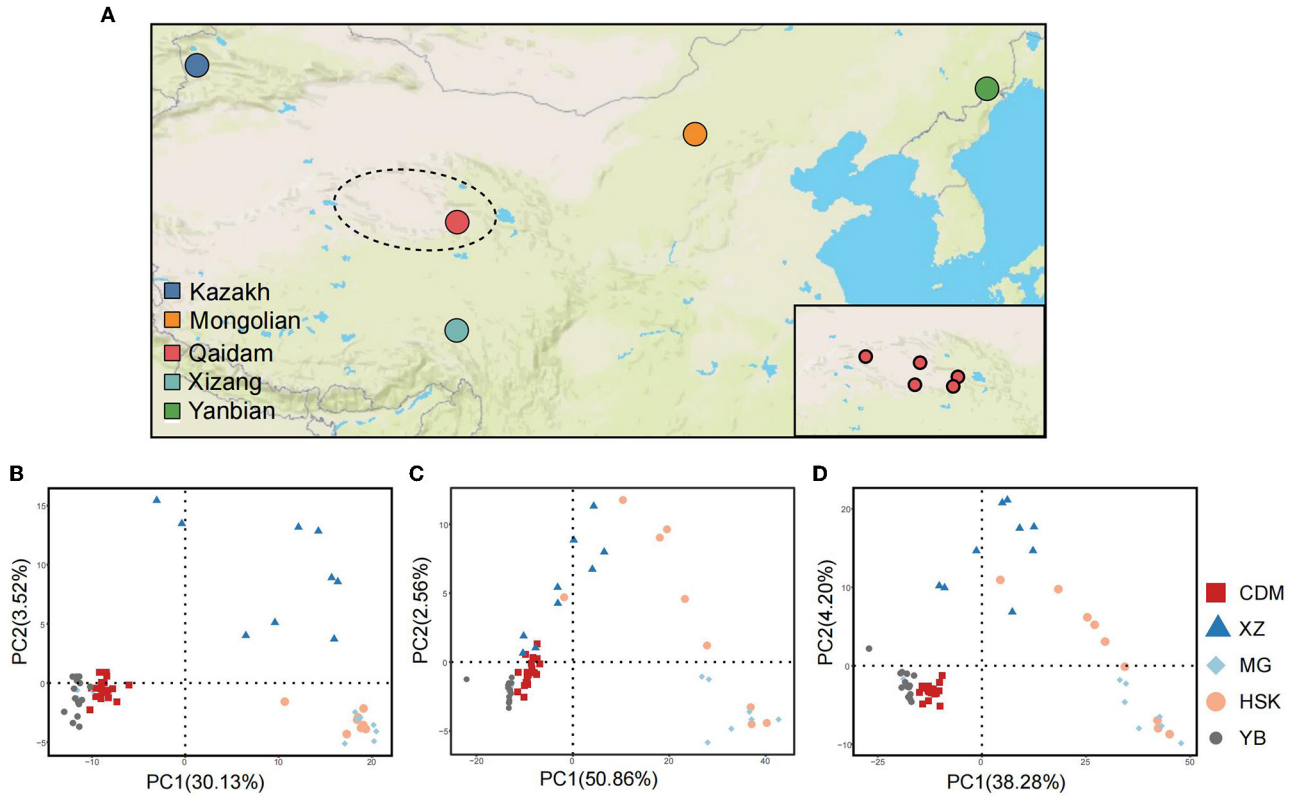


FIGURE 2
Geographic distribution and population stratification in five Chinese cattle. **(A)** The major geographical locations of the five Chinese cattle breeds. The dotted circle represents the Qaidam Basin, which is the main geographical distribution area of Qaidam cattle. In this study, Qaidam cattle were selected from five geographical locations along the Qaidam Basin, as shown in the lower right corner, which are Dulan, Golmud, Mangya, Wulan, and Dachaidan. **(B)** The principal component analysis (PCA) was derived from CNVRs (deletions). **(C)** PCA derived from CNVRs (duplications). **(D)** PCA derived from all CNVRs in five population genomes.

3.3. Differentiated CNVRs between Qaidam cattle and other cattle breeds

We calculated the V_{ST} between Qaidam cattle (CDM) and other cattle breeds from the regions of North China (XZ, HSK, YB, and MG) (Supplementary Figure 2, Supplementary Tables 4–7). First, the top 1% signal value regions were kept out; then, it was annotated by the cattle reference genome (ARS-UCD1.2).

The selection signals between CDM and YB were enriched to "MAPK signaling pathway" ($p = 4.49 \times 10^{-7}$), "Pi3k-akt signaling pathway" ($p = 1.82 \times 10^{-8}$), "mTOR signaling pathway" ($p = 3.22 \times 10^{-6}$), "HIF-1 signaling pathway" ($p = 4.95 \times 10^{-5}$), and "aldosterone-regulated sodium reabsorption" ($p = 5.91 \times 10^{-4}$) (Figure 3A, Supplementary Table 8). There were five candidate genes (*MUC6*, *WDR25*, *CNNM4*, *MGAM*, and *GFRA2*) in the study (Figure 4A, Supplementary Table 4). The selection signals between CDM and MG were enriched to "ErbB signaling pathway" ($p = 2.73 \times 10^{-4}$), "calcium signaling pathway" ($p = 2.69 \times 10^{-3}$), "GnRH signaling pathway" ($p = 0.01122$), and "insulin signaling pathway" ($p = 0.04588$) (Figure 3B, Supplementary Table 9). Among these annotated genes, four genes (*PTPRT*, *BOLL*, *PLIN4*, and *ADGRL3*) deserved more attention in copy number between CDM and MG (Figure 4B, Supplementary Table 5). The selection signals between CDM and XZ were enriched to "axon guidance" ($p = 5.48 \times 10^{-6}$) and "ErbB signaling pathway" ($p = 7.49 \times 10^{-5}$) (Figure 3C, Supplementary Table 10). Among these annotated genes, five genes (*PLIN4*, *CDH13*, *SYCP1*, *PTPRC*, and *ADAMTSL3*) were noteworthy in copy number between CDM and XZ (Figure 4C, Supplementary Table 6). There was a difference in copy numbers between CDM and HSK. The selection signal enrichment pathways between CDM and HSK include "bacterial invasion of epithelial cells" ($p = 1.10 \times 10^{-3}$), "Salmonella infection" ($p = 0.02195$), and "human immunodeficiency virus 1 infection" ($p = 0.02446$) (Figure 3D, Supplementary Table 11). Among these annotated genes, three genes (*KHDRBS2*, *THRDE*, and *EBF2*) were notable in copy number between CDM and HSK (Figure 4D, Supplementary Table 7).

4. Discussion

During domestication and diversification, the frequency of copy number variation in the species' genome responds to selective pressure. Considerable effort has been applied to identify the causal mutations and genes. However, screening the selected genomic copy number genetic markers is complex. Over the past decades, high-throughput sequencing techniques and bioinformatics tools have been increasingly used to construct genome-wide CNV maps (1, 15, 17, 18). The diversity of CNVs

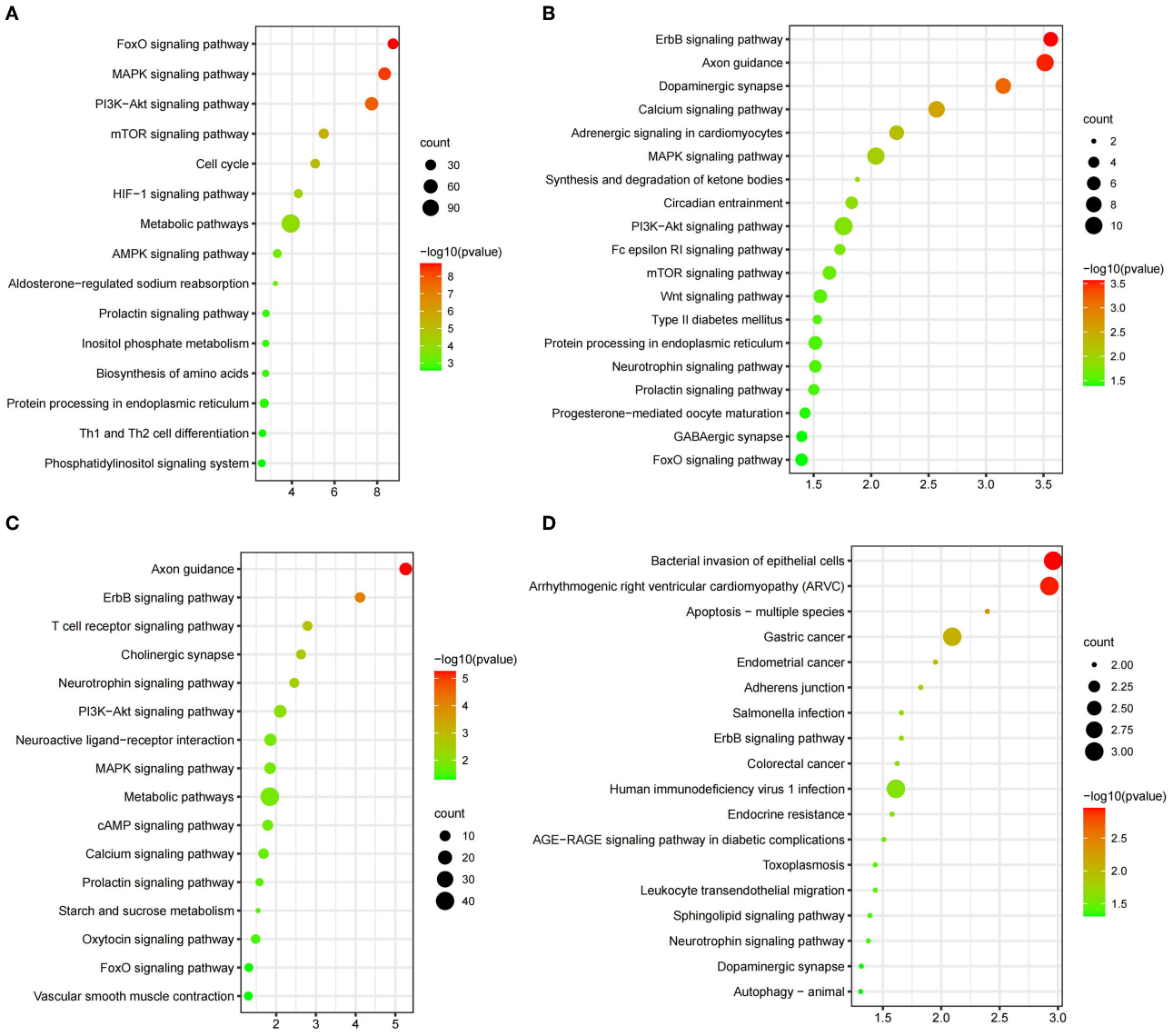


FIGURE 3
KEGG pathways from the enrichment analysis. **(A)** CDM vs. YB; **(B)** CDM vs. MG; **(C)** CDM vs. XZ; **(D)** CDM vs. HSK) ($p < 0.05$). We preferred showing part of the pathways; for more information, please see Supplementary Tables 8–11.

has been extensively explored in *Bos Taurus*, *Bos Indicus*, and their crossing populations.

In our study, we investigated the CNV of 20 newly resequenced Qaidam cattle genomes based on the ARS_UCD 1.2 cattle reference genome. It improved the reliability of screening CNVs more than through UMD 3.1 assembly (19, 20). A total of 10,178 CNVRs were detected in five Chinese indigenous cattle breeds, and more than 99.9% CNVRs in length ranged from 1 to 100 kb. It was suggested that CNVs were widespread in Chinese cattle and may have been caused by the rapid adaptation during population expansion. For better statistics, variants were divided into three categories: duplication, deletion, and both duplication and deletion. The duplication was higher than deletions in number (Supplementary Table 3). And most of the CNVRs ranging from 2 to 5 kb in length (Figure 1A). In addition, the location of CNVRs is not uniformly distributed in the cattle genome (Figure 1C), and they are also not randomly distributed on chromosomes. The annotation uncovered that CNVRs are mostly annotated in the intergenic or intronic regions in the cattle genome. A previous study has also supported that many CNVRs are located on highly variable genes (15).

Compared to the analysis of the genome CNV in Qaidam cattle (18) for the first time, the role that CNVs have in the evolution of Qaidam cattle is becoming clear through our present study. The Qaidam cattle have strong adaptability to the arid environment, exhibiting dry, hypoxia, low air pressure, and large diurnal temperature difference (relative humidity 29–42%, precipitation 140–210.4 mm). Interestingly, YB cattle have almost opposite living conditions (relative humidity 68.6%; precipitation 500–700 mm) than Qaidam cattle. By consulting scientific articles,

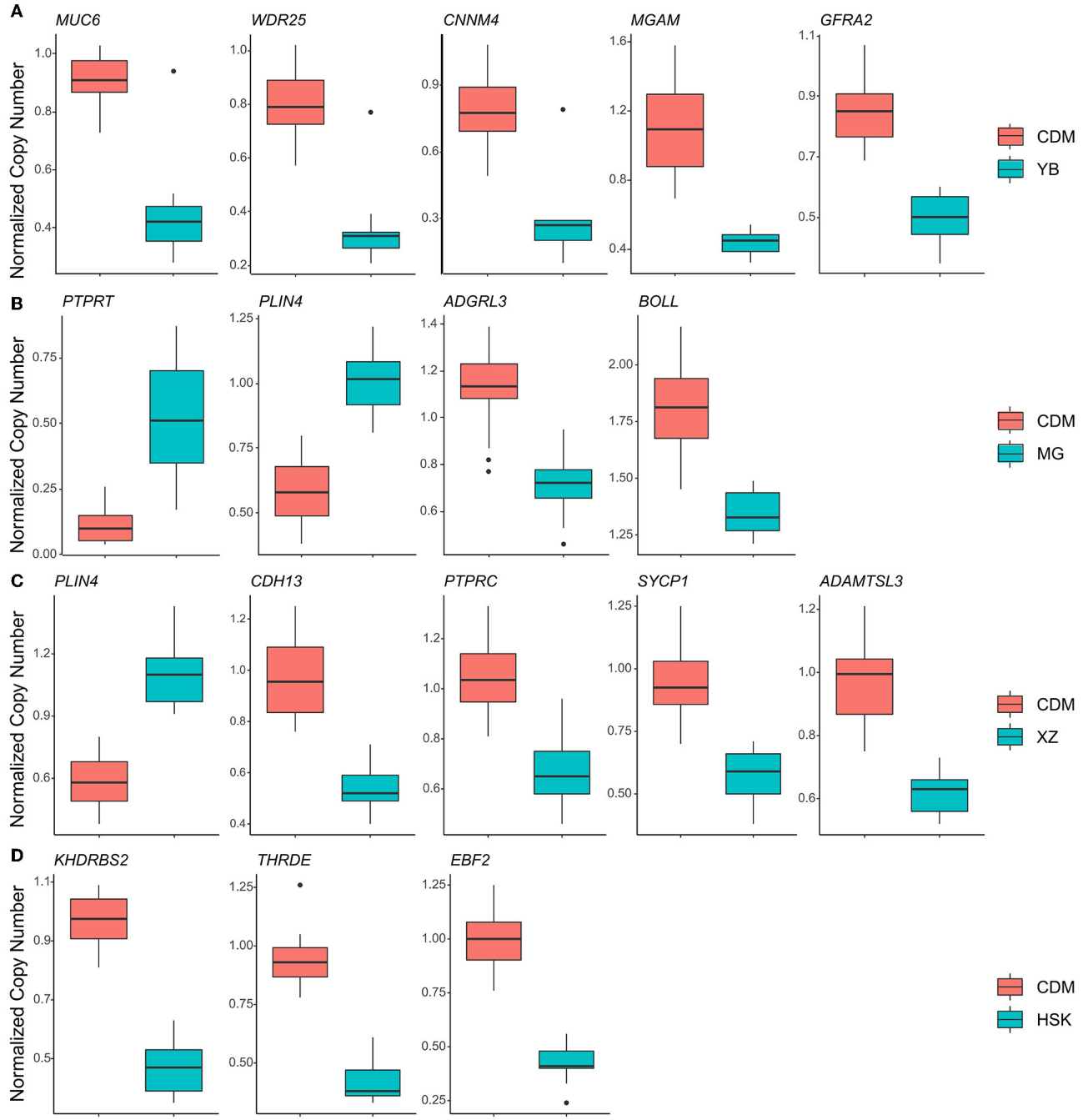


FIGURE 4
Normalized copy number comparison of the top 1% V_{ST} genes between CDM and other populations. **(A)** CDM compared with the YB population. **(B)** CDM compared with the MG population. **(C)** CDM compared with the XZ population. **(D)** CDM compared with the HSK population.

we found that *EBF1* and *ZNF521* related to fat development (21, 22) and *VEGFA*, *EGLN2*, and *ENO3* were associated with high altitude hypoxic adaptation (23–25). In the enrichment analysis, the *MGAM* gene was significantly enriched in the "metabolic pathways (bta01100, *P*-value = 0.000113)" (Figure 3A), and was also clustered in the "carbohydrate metabolic process (GO:0005975, *P*-value = 0.014482)" (Supplementary Table 8). A previous study reported on the CNVR overlapping with the *MAGM* gene, and that it was related to starch digestion (26). Specifically, we found that *MUC6* in Qaidam cattle was a normal-type CNVR, but it is a deletion CNVR in the YB cattle genome (Figure 4A). A previous study found CNV polymorphism in the *MUC6* gene of domestic sheep, and this CNVR presents normal or duplication under arid environments, and deletion in warm and humid environments (27). Structurally, large numbers of tandem repeats rich in Pro, Thr, and Ser residues in *MUC6* can

affect the covalent attachment of O-glycans (28). In ruminants, such as sheep and cattle, the *MUC6* gene has been associated with gastrointestinal parasite resistance (29, 30). Therefore, we hypothesized that the copy number difference of the *MUC6* gene may influence the ability of antiparasitic immunity in Qaidam cattle and YB cattle.

High-quality beef is the breeding target of Qaidam cattle. In the comparison between Qaidam cattle and MG cattle (Supplementary Figure 2B), we observed that the *PRKCA*, *CAMK2D*, *PHKB*, and *GRID2* genes (Supplementary Table 5) (V_{ST} value $>$ 0.43) were related to muscle growth and development by searching previous research studies (31–34). Moreover, we identified *PTPRT*, *BOLL*, *PLIN4,* and *ADGRL3* gene regions in the CNVRs of the top V_{ST} values which have obvious copy number differences between Qaidam cattle and MG cattle (Figure 4B). The *ADGRL3* gene is associated with the nervous system of the Fuzhong buffalo (34). In addition, *PTPRT* (Chr13: s17021.1) was associated with body weight for pre-weaning growth in Esme sheep (35). In addition, the functional enrichment analysis of candidate genes with top 1% signal V_{ST} values revealed that the "GnRH signaling pathway" and "calcium signaling pathway" were significantly overrepresented. These results imply that the selected genes might contribute to the characteristics of growth rate and meat quality in Qaidam cattle.

Body size is one of the important traits in the evaluation of beef selection. In this study, we identified *ADAMTSL3*, *PLIN4*, *CDH13*, *SYCP1*, and *PTPRC* genes of the top 1% signal regions between Qaidam and XZ cattle (Supplementary Table 6). According to previous research, *ADAMTSL3* plays an important role in chondrogenesis, morphogenesis, and skeletal growth in humans (36). A previous study reported that the bovine *ADAMTSL3* gene has specific polymorphisms in individuals and the SNPs (T1532C and C1899T) were significantly associated with body size traits (37). Our results further suggested that copy number in the *ADAMTSL3* gene may be one of the reasons for the difference in body size between XZ cattle and Qaidam cattle.

By comparing the copy number differences between the HSK and Qaidam breeds on the genome (Supplementary Figure 2C), we identified CNVRs with significant differences including *KHDRBS2*, *THRDE*, and *EBF2* (Figure 4D, Supplementary Table 7). One of the eye-catching genes was *EBF2*, which has copy number polymorphism and showed a normal type in Qaidam cattle but a deletion type in HSK cattle (Figure 4D). Previous studies showed that *EBF2* promotes brown adipocyte differentiation (38) and that its loss in mouse adipocytes abrogates brown adipose tissue (BAT) characteristics and function, leading to cold intolerance (39, 40). The cold tolerance of Qaidam cattle is an essential characteristic and it was speculated to be related to the copy number variation of *EBF2*.

5. Conclusion

Based on the high-quality *Bos taurus* reference genome, we constructed a CNV map of Northern Chinese Qaidam cattle using whole-genome resequencing data. Moreover, there are many copy number differences between Qaidam cattle and other cattle breeds from the regions of North China. It may play a crucial role in understanding the Qaidam cattle's adaptability, growth, and developmental characteristics In conclusion, these results provide a wealth of CNVR information to explore the valuable molecular markers in the Qaidam cattle genome.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of Qinghai Academy of Animal Science and Veterinary Medicine, Qinghai University. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

YL drafted the manuscript and took part in the analysis of genome data. ZM, CL, and XW contributed to the sample collection. WW and YM performed the primary analysis of genome data. ZA and ZM revised the writing. ZM and CL designed the experiment and provided the funding for this research. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of Qinghai Province of China (2021-ZJ-914), Kunlun Talent. High-end Innovation and Entrepreneurship Talents Program of Qinghai Province and the earmarked fund for China Agriculture Research System of MOF and MARA (CARS-37).

Acknowledgments

The authors would like to thank the High-Performance Computing (HPC) of Northwest A&F University (NWAFU), China for providing computing resources.

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/fvets.2023.1148070/full#supplementary-material

References

1. Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat Commun.* (2018) 9:2337. doi: 10.1038/s41467-018-04737-0

2. Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. Evidence for two independent domestications of cattle. *Proc Nat Acad Sci.* (1994) 91:2757–61. doi: 10.1073/pnas.91.7.2757

3. Lye ZN, Purugganan MD. Copy number variation in domestication. *Trends Plant Sci.* (2019) 24:352–65. doi: 10.1016/j.tplants.2019.01.003

4. National Committee of Animal Genetic Resources (2021). *National List of Livestock and Poultry Genetic Resources in China*. Beijing, China. Available online at: http://www.moa.gov.cn/govpublic/nybzzj1/202101/t20210114_6359937.htm.

5. Ma Z, Li R, Xia X, Luo J, Xie Y, Sun Y, et al. Y-SNPs genetic diversity, population genetic structure and paternal origin of Qaidamcattle. *Genomics Appl Biol.* (2018) 37:1920–5. doi: 10.13417/j.gab.037.001920

6. Zhao XZ, Zhao LX. Conservation status and development strategy of livestock and poultry genetic material in Qinghai Province. *Qinghai J Animal Husbandry Vet Med.* (2022) 52:65–8.

7. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature.* (2011) 470:59–65. doi: 10.1038/nature09708

8. He Y, Hong Q, Zhou D, Wang S, Yang B, Yuan Y, et al. Genome-wide selective detection of Mile red-bone goat using next-generation sequencing technology. *Ecol Evol.* (2021) 11:14805–12. doi: 10.1002/ece3.8165

9. Yuan X, Li J, Bai J, Xi J. A local outlier factor-based detection of copy number variations from NGS data. *IEEE/ACM Trans Comput Biol Bioinform.* (2021) 18:1811–20. doi: 10.1109/TCBB.2019.2961886

10. Zhang Y, Hu Y, Wang X, Jiang Q, Zhao H, Wang J, et al. Population structure, and selection signatures underlying high-altitude adaptation inferred from genome-wide copy number variations in Chinese indigenous cattle. *Front Genet.* (2019) 10:1404. doi: 10.3389/fgene.2019.01404

11. Yang L, Niu Q, Zhang T, Zhao G, Zhu B, Chen Y, et al. Genomic sequencing analysis reveals copy number variations and their associations with economically important traits in beef cattle. *Genomics.* (2021) 113: 812–20. doi: 10.1016/j.ygeno.2020.10.012

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

13. Wang X, Zheng Z, Cai Y, Chen T, Li C, Fu W, et al. CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. *Gigascience.* (2017) 6:1–12. doi: 10.1093/gigascience/gix115

14. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010) 38:e164. doi: 10.1093/nar/gkq603

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

16. Yang L, Xu L, Zhu B, Niu H, Zhang W, Miao J, et al. Genome-wide analysis reveals differential selection involved with copy number variation in diverse Chinese Cattle. *Sci Rep.* (2017) 7:14299. doi: 10.1038/s41598-017-14768-0

17. Zhou B, Ho SS, Zhang X, Pattni R, Haraksingh RR, Urban AE, et al. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet.* (2018) 55:735. doi: 10.1136/jmedgenet-2018-105272

18. Guo S, Wu X, Pei J, Wang X, Bao P, Xiong L, et al. Genome-wide CNV analysis reveals variants associated with high-altitude adaptation and meat traits in Qaidam cattle. *Electron J Biotechnol.* (2021) 54:8–16. doi: 10.1016/j.ejbt.2021.07.006

19. Zhou J, Liu L, Reynolds E, Huang X, Garrick D, Shi Y, et al. Discovering copy number variation in dual-purpose Xinjiang brown cattle. *Front Genet.* (2022) 12:747431. doi: 10.3389/fgene.2021.747431

20. Lloret-Villas A, Bhati M, Kadri NK, Fries R, Pausch H. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genom.* (2021) 22:363. doi: 10.1186/s12864-021-07554-w

21. Chiarella E, Aloisio A, Codispoti B, Nappo G, Scicchitano S, Lucchino V, et al. ZNF521 Has an inhibitory effect on the adipogenic differentiation of human adipose-derived mesenchymal stem cells. *Stem Cell Rev Rep.* (2018) 14:901–14. doi: 10.1007/s12015-018-9830-0

22. Dang TN, Taylor JL, Kilroy G, Yu Y, Burk DH, Floyd ZE, et al. SIAH2 is Expressed in adipocyte precursor cells and interacts with EBF1 and ZFP521 to promote adipogenesis. *Obesity.* (2021) 29:98–107. doi: 10.1002/oby.23013

23. Wu DD, Ding XD, Wang S, Wójcik JM, Zhang Y, Tokarska M, et al. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat Ecol Evol.* (2018) 2:1139–45. doi: 10.1038/s41559-018-0562-y

24. Droma Y, Hanaoka M, Kinjo T, Kobayashi N, Yasuo M, Kitaguchi Y, et al. The blunted vascular endothelial growth factor-A (VEGF-A) response to high-altitude hypoxia and genetic variants in the promoter region of the VEGFA gene in Sherpa highlanders. *Peer J.* (2022) 10:e13893. doi: 10.7717/peerj.13893

25. Zhang B, Chamba Y, Shang P, Wang Z, Ma J, Wang L, et al. Comparative transcriptomic and proteomic analyses provide insights into the key genes involved in high-altitude adaptation in the Tibetan pig. *Sci Rep.* (2017) 7:3654. doi: 10.1038/s41598-017-03976-3

26. Lee Y, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC, et al. Functional and population genetic features of copy number variations in two dairy cattle populations. *BMC Genom.* (2020) 21:89. doi: 10.1186/s12864-020-6496-1

27. Zheng Z, Wang X, Li M, Li Y, Yang Z, Wang X, et al. The origin of domestication genes in goats. *Sci Adv.* (2020) 6:z5216. doi: 10.1126/sciadv.aaz5216

28. Moniaux N, Escande F, Porchet N, Aubert JP, Batra SK. Structural organization and classification of the human mucin genes. *Front Biosci.* (2001) 6:D1192–206. doi: 10.2741/A579

29. Simpson HV, Umair S, Hoang VC, Savoian MS. Histochemical study of the effects on abomasal mucins of Haemonchus contortus or Teladorsagia circumcincta infection in lambs. *Vet Parasitol.* (2016) 226:210–21. doi: 10.1016/j.vetpar.2016.06.026

30. Rinaldi M, Dreesen L, Hoorens PR, Li RW, Claerebout E, Goddeeris B, et al. Infection with the gastrointestinal nematode Ostertagia ostertagi in cattle affects mucus biosynthesis in the abomasum. *Vet Res.* (2011) 42:61. doi: 10.1186/1297-9716-42-61

31. Luo, R., Dai, X., Zhang, L., Li, G., and Zheng, Z. (2022). Genome-wide DNA methylation patterns of muscle and tail-fat in dairymeade sheep and mongolian sheep. *Animals.* 12:1399. doi: 10.3390/ani12111399

32. Dou D, Shen L, Zhou J, Cao Z, Luan P, Li Y, et al. Genome-wide association studies for growth traits in broilers. *BMC Genom Data.* (2022) 23:1. doi: 10.1186/s12863-021-01017-7

33. Smith JL, Wilson ML, Nilson SM, Rowan TN, Schnabel RD, Decker JE, et al. Genome-wide association and genotype by environment interactions for growth traits in U.S. *Red Angus cattle. BMC Genom.* (2022) 23:517. doi: 10.1186/s12864-022-08667-6

34. Sun T, Huang GY, Wang ZH, Teng SH, Cao YH, Sun JL, et al. Selection signatures of Fuzhong Buffalo based on whole-genome sequences. *BMC Genom.* (2020) 21:674. doi: 10.1186/s12864-020-07095-8

35. Yilmaz O, Kizilaslan M, Arzik Y, Behrem S, Ata N, Karaca O, et al. Genome-wide association studies of preweaning growth and in vivo carcass composition traits in Esme sheep. *J Animal Breed Genet.* (2022) 139:26–39. doi: 10.1111/jbg.12640

36. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet.* (2008) 40:575–83. doi: 10.1038/ng.121

37. Liu Y, Zan L, Zhao S, Xin Y, Jiao Y, Li K, et al. Molecular characterization, expression pattern, polymorphism and association analysis of bovine ADAMTSL3 gene. *Mol Biol Rep*. (2012) 39:1551–60. doi: 10.1007/s11033-011-0894-z

38. Rajakumari S, Wu J, Ishibashi J, Lim HW, Giang AH, Won KJ, et al. EBF2 determines and maintains brown adipocyte identity. *Cell Metab*. (2013) 17:562–74. doi: 10.1016/j.cmet.2013.01.015

39. Angueira AR, Shapira SN, Ishibashi J, Sampat S, Sostre-Colón J, Emmett MJ, et al. Early B cell factor activity controls developmental and adaptive thermogenic gene programming in adipocytes. *Cell Rep*. (2020) 30:2869–78. doi: 10.1016/j.celrep.2020.02.023

40. Stine RR, Shapira SN, Lim H, Ishibashi J, Harms M, Won K, et al. EBF2 promotes the recruitment of beige adipocytes in white adipose tissue. *Mol Metabol*. (2016) 5:57–65. doi: 10.1016/j.molmet.2015.11.001

frontiers | Frontiers in Veterinary Science

TYPE Original Research
PUBLISHED 29 August 2023
DOI 10.3389/fvets.2023.1132833

OPEN ACCESS

EDITED BY
Anupama Mukherjee,
Indian Council of Agricultural Research (ICAR), India

REVIEWED BY
Zhuanjian Li,
Henan Agricultural University, China
Tatiana Deniskova,
L.K. Ernst Federal Science Center for Animal Husbandry (RAS), Russia

*CORRESPONDENCE
Chuanying Pan
✉ chuanyingpan@126.com
Mei Liu
✉ Mei.Liu@hunau.edu.cn

†These authors have contributed equally to this work

RECEIVED 28 December 2022
ACCEPTED 17 July 2023
PUBLISHED 29 August 2023

CITATION
Wang Q, Song X, Bi Y, Zhu H, Wu X, Guo Z, Liu M and Pan C (2023) Detection distribution of CNVs of *SNX29* in three goat breeds and their associations with growth traits. *Front. Vet. Sci.* 10:1132833. doi: 10.3389/fvets.2023.1132833

Detection distribution of CNVs of *SNX29* in three goat breeds and their associations with growth traits

Qian Wang[1†], Xiaoyue Song[2,3†], Yi Bi[1], Haijing Zhu[2,3], Xianfeng Wu[4], Zhengang Guo[5], Mei Liu[6*] and Chuanying Pan[1*]

[1]College of Animal Science and Technology, Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, Northwest A&F University, Yangling, Shaanxi, China, [2]Shaanxi Provincial Engineering and Technology Research Center of Cashmere Goats, Yulin University, Yulin, Shaanxi, China, [3]Life Science Research Center, Yulin University, Yulin, Shaanxi, China, [4]Institute of Animal Husbandry and Veterinary, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian, China, [5]Animal Husbandry and Veterinary Science Institute of Bijie City, Bijie, Guizhou, China, [6]College of Animal Science and Technology, Hunan Agricultural University, Changsha, Hunan, China

As a member of the SNX family, the *goat sorting nexin 29* (*SNX29*) is initially identified as a myogenesis gene. Therefore, this study aimed to examine the polymorphism in the *SNX29* gene and its association with growth traits. In this study, we used an online platform to predict the structures of the *SNX29* protein and used quantitative real-time PCR to detect potential copy number variation (CNV) in Shaanbei white cashmere (SBWC) goats ($n = 541$), Guizhou black (GB) goats ($n = 48$), and Nubian (NB) goats ($n = 39$). The results showed that goat *SNX29* protein belonged to non-secretory protein. Then, five CNVs were detected, and their association with growth traits was analyzed. In SBWC goats, CNV1, CNV3, CNV4, and CNV5 were associated with chest width and body length ($P < 0.05$). Among them, the CNV1 individuals with gain and loss genotypes were superior to those individuals with a median genotype, but CNV4 and CNV5 of individuals with the median genotype were superior to those with the loss and gain genotypes. In addition, individuals with the gain genotype had superior growth traits in CNV3. In brief, this study suggests that the CNV of *SNX29* can be used as a molecular marker in goat breeding.

KEYWORDS

***sorting nexin 29* (*SNX29*) gene, copy number variation (CNV), growth traits, goats, marker-assisted selection (MAS)**

Introduction

Members of the SNX family are located in membrane-binding cytoplasm and can bind to phosphatidylinositol via the PX domain and interact with membrane-associated protein complexes, which play an important role in regulating endocytosis and protein transport through cell membrane compartments (1, 2). To date, 32 members have been identified, and they are divided into five subgroups based on protein domain. Among them, the *SNX29* gene belongs to the SNX-PX subfamily (3), which has been reported to be involved in disease, nervous system development, and animal growth. Studies have linked the *SNX29* gene to schizophrenia (SCZ), autism, and other psychiatric disorders (4, 5). The deletion of *SNX29* intron 14 may lead to primary testicular lymphoma (6). Zhu et al. found that downregulation of the *SNX29* gene was associated with epithelial ovarian carcinoma cells (7). Furthermore, Sparks et al. showed a strong association between IgA levels and the region between 6.89

and 14.95 Mb on sheep chromosome 24, which corresponds to the *SNX29* gene (8). A circRNA of the *SNX29* gene regulated the proliferation and differentiation of muscle cells (9). Studies have shown that the *SNX29* gene plays a key role in subcutaneous fat deposition in Xiangdong black (XDB) goats, and the *SNX29* CNV is significantly associated with the chest and abdominal girth of XDB goats ($P < 0.01$) (10). Based on the above, the *SNX29* gene was selected to be studied in this study.

Copy number variation (CNV) exists widely in the genomes of organisms, and it is considered to be an important source of genetic differences between individuals (11, 12). In recent years, some studies reported that CNV was significantly correlated with the economic traits of livestock, such as litter size (13), meat quality (14), milk production (15), weight gain rate (16), and feed conversion rate (17). The advantages of CNV-promoting population diversity, simplicity, and efficiency were discovered by more people (18). As a applicable molecular marker, CNV can make marker-assisted selection (MAS) better play the advantages of convenience, simplicity, and so on. In short, it provides new ideas and methods for breeding work.

Shaanbei white cashmere (SBWC) goats were bred from Liaoning white cashmere goat and Ziwuling black goat (19), which has high cashmere value and meat value (20). Guizhou black (GB) goats are an excellent local breed with good meat quality and coarse feeding tolerance (21). Nubian (NB) goats have good value for meat and milk and have higher meat content than other dual-purpose goats (22). However, their growth performance fails to achieve the expected results, so it is helpful to increase the economic value of goats by improving their growth traits through MAS.

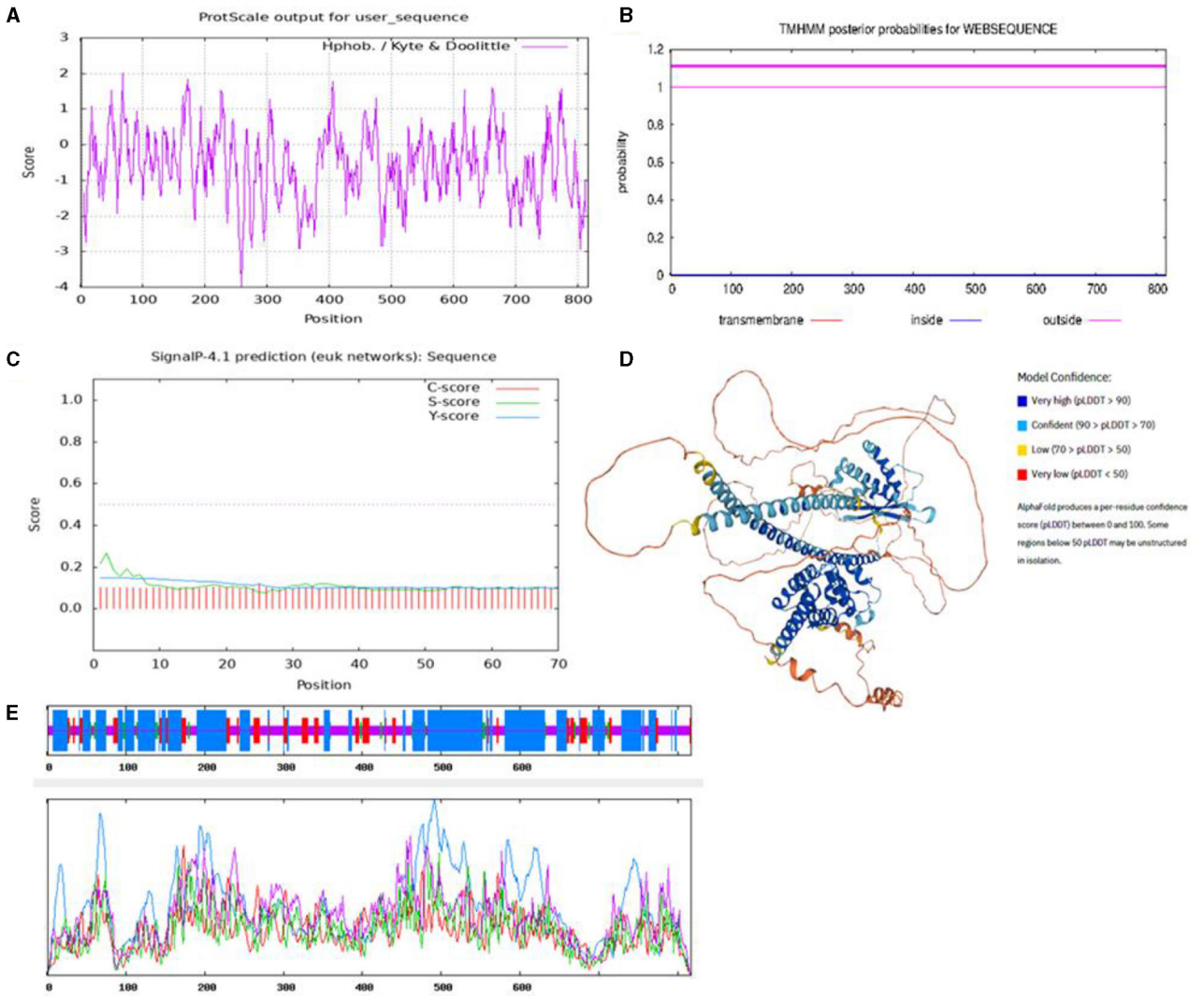


FIGURE 1
Prediction of physicochemical properties and structure of the *SNX29* protein. **(A)** Hydrophobicity of goat *SNX29* protein. **(B)** Goat *SNX29* protein transmembrane signal peptide. **(C)** Goat *SNX29* protein transmembrane signal peptide prediction. The abscissa axis represents the sequence number of amino acid residues corresponding to the submitted protein sequence; the value of the ordinate axis is the probability value of each amino acid located on the inside, outside, and TMhelix on the abscissa axis. **(D)** Three-dimensional model of goat *SNX29* protein tertiary structure. **(E)** Secondary structural parameters of goat *SNX29* protein. Blue means a-helix, red means extended backbone, green means β-folding, and yellow means random crimping.

Currently, the CNVs of the *SNX29* gene and its association with growth traits in SBWC goats have not been reported. Therefore, this study is characterized based on the aspects of protein structure, physicochemical properties, and DNA variation. Next, we explored five potential CNVs, which were detected in SBWC goats, GB goats, and NB goats by quantitative real-time PCR (qRT-PCR). An association analysis was carried out between the *SNX29* gene and the growth traits of goats. These results will have a deeper understanding of gene variation and livestock growth traits, in order to lay a theoretical foundation for MAS breeding of goats.

Materials and methods

Animal welfare explanation

The samples used in this experiment comply with the Regulations on the Administration of Experimental Animals at Northwest A&F University (NWAFU-314020038).

Prediction of *SNX29* protein physicochemical properties and structure

Using NCBI-searched *SNX29* protein sequences, the goats' *SNX29* protein amino acid number, molecular weight, and isoelectric point were calculated using the Expasy online platform, and the ProtScale application and ProtParam were used to predict the protein hydrophobicity. The *SNX29* protein of transmembrane signal peptide was predicted using the TMHMM database and SignalP 4.1. The AlphaFold and SOPMA online platforms were used to predict the advanced structure of the *SNX29* protein (23) (Supplementary Table S1).

Sample collection and genomic DNA extraction

Under the same feeding conditions, ear tissues of 541 SBWC goats, 48 GB goats, and 39 NB goats were selected from the Yulin goat farm in Shaanxi province, the Bijie goat farm in Guizhou province, and the Zhangzhou Nubian goat breeding cooperative in Fujian province. All the individuals were female goats (2–3 years) and were not related to each other. Genomic DNA was extracted from goat ear tissue using the high salt extraction method (24) and stored at 70% alcohol at −80°C (25). A NanoDrop™2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) was used to measure the $OD_{260/280}$ ratio, and a ratio between 1.8 and 2.0 means that the nucleic acid concentration is qualified (26). Then, the extracted DNA was placed at −40°C.

Primer designing

We searched the Animal Omics database (27) (Supplementary Table S1) and found five CNV loci of the *SNX29* gene in goats. Five pairs of amplified primers were referenced in a previous article (28).

CNV genotyping detection of the *SNX29* gene

To ensure that the primers can amplify the target fragment, the primers are detected through the mixed pool (CNV1 = 137 bp, CNV2 = 138 bp, CNV3 = 104 bp, CNV4 = 151 bp, and CNV5 = 109 bp). Next, 541 SBWC goat samples, 48 GB goat samples, and 39 NB goat samples were used to detect the CNV loci. qRT-PCR amplification systems and procedures refer to previous laboratory articles (29, 30). The result was processed using method $2^{*}2(-\delta Ct)$ (31).

Statistical analyses

The association between the variants and growth traits was explored using the analysis of variance (ANOVA) and independent sample *t*-test in SPSS 26.0 (IBM, USA), and the chi-square (χ^2) test was used to analyze the significance between the three breeds (32). And the line model was used as a reference by Liu et al. (33). Where $Y_{ijk} = \alpha_i + \beta_j + e_{ijk} + u$ acts as an analysis model, Y_{ijk} is the evaluation of growth traits at the i level of fixed factor age (α_i) and j level of fixed factor genotype (β_j), u is the overall mean, and e_{ijk} is the random error.

Results

Prediction of *SNX29* protein physicochemical properties and structure

To characterize the functions of the *SNX29* gene, the protein structure and physicochemical properties were predicted. The results showed that the protein contained 817 amino acids, the molecular weight was 9,143.14, and the isoelectric point was 5.90 by the Expasy online platform. ProtScale online software predicted the hydrophobicity of the protein, and the results showed that there were more hydrophilic residues in the goat *SNX29* protein, which indicated that this protein was hydrophilic (Figure 1A). The results were consistent with ProtParam online software predictions. TMHMM prediction results showed that the protein encoded by the *SNX29* gene did not have transmembrane helix (Figure 1B). SignalP 4.1 prediction results showed that the D critical value of signal peptide and non-signal peptide of this protein was 0.450, and the D critical value of the *SNX29* protein was 0.155 (Figure 1C). According to the signal peptide hypothesis, the *SNX29* protein had no signal peptide and belonged to non-secretory protein. The SOPMA online platform predicted the detailed information on the secondary structure of *SNX29* protein, and the results showed that alpha helix accounted for 47.98%, extended strand accounted for 12.24%, β-turn accounted for 4.04%, and random coil accounted for 35.74% (Figure 1E). AlphaFold online software predicted the three-dimensional structure of the *SNX29* protein (Figure 1D).

Frequency of CNV genotypes in goats

After mixed pool detection, it was found that the five CNVs were consistent with the target band (Figure 2). Then, by expanding the sample size for testing, the following results were obtained: In CNV1, the proportion of gain genotype was greater than that of median and loss genotypes in goats. There were 85.61% individuals with gain genotypes in the SBWC goats; however, and all individuals in the GB goats and NB goats were gain genotypes; in CNV2, all three goat breeds were gain genotype; in CNV3, there were 72.18% individuals of gain genotype, 3.31% individuals of median genotype, and 24.52% individuals of loss genotype in SBWC goats, and all GB goats and NB goats were gain genotype; in CNV4, there were 51.25% individuals of gain genotype, 31.67% individuals of median genotype, and 17.08% individuals of loss genotype in SBWC goats, there were 80.43% individuals of gain genotype, 19.57% individuals of median genotype in GB goats, and NB goats were all gain genotype; and in CNV5, there were 56.72% individuals of gain genotype, 31.45% individuals of median genotype, and 11.83% individuals of loss genotype in SBWC goats, there were 48.94% individuals of gain genotype, 51.06% individuals of median genotype in GB goats, there were 84.21% individuals of gain genotype, and 15.79% individuals of median genotype in NB goats (Figure 3).

Association analysis between CNVs and the goat *SNX29* gene

The association analysis results showed that four CNVs were related to growth traits in SBWC goats. CNV1 was significantly associated with chest width ($P = 0.002$), body length ($P = 1.230E-4$), body height ($P = 0.008$), cannon circumference ($P = 1.300E-5$), and heart girth ($P = 0.033$). CNV3 was significantly associated with chest width ($P = 0.004$) and cannon circumference ($P = 0.009$). CNV4 was significantly associated with chest width ($P = 8.166E-7$), heart girth ($P = 2.620E-4$), and cannon circumference ($P = 0.001$). CNV5 was significantly associated with chest depth ($P = 0.008$) and body length ($P = 0.025$). Additionally, in the association analysis between growth traits of SBWC goats and CNVs, we found that in CNV1 individuals, gain and loss genotypes were superior to those with median genotype on the aspect of growth traits, but in CNV4 and CNV5 individuals, median genotypes were superior to loss and gain. In addition, in the CNV3, the gain genotype performed better growth traits (Table 1). The χ^2 test results showed that except

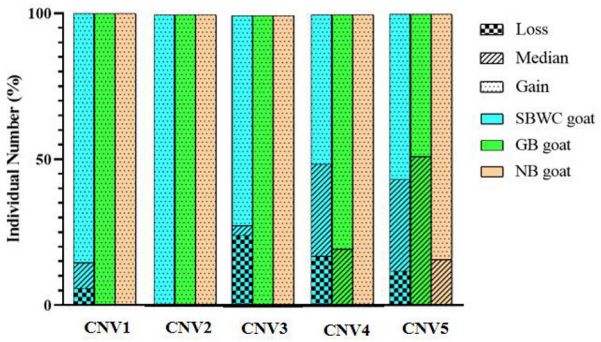


FIGURE 3
Genotyping proportion of CNVs in SBWC goat, GB goat, and NB goat. In CNV1: the total individual number of SBWC goats was 278, GB goats was 48, and NB goats was 38; in CNV2: the total individual number of SBWC goats was 290, GB goats was 48, and NB goats was 39; in CNV3: the total individual number of SBWC goats was 363, GB goats was 48, and NB goats was were 281, GB goats was 46, and NB goats was 38; in CNV5: the total individual number of SBWC goats was 372, GB goats was 48, and NB goats was 39.

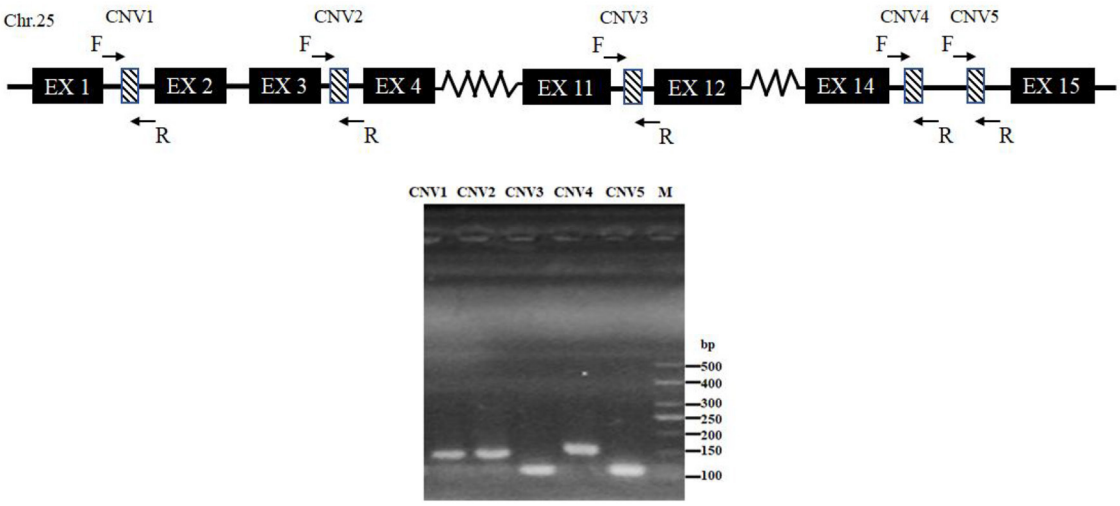


FIGURE 2
Schematic diagram of the PCR assay for CNVs of the goat *SNX29* gene. Chr, chromosome; EX, exon; F, forward primer; R, reverse primer; M, means DNA marker.

TABLE 1 Association analysis between growth traits and the CNVs in SBWC goats.

CNV loci	Trait types	Typical frequencies (AVG ± SE)			*P*-values
		Loss	Median	Gain	
CNV1	Height at hip cross (cm)	60.84 ± 0.99 ($n = 16$)	58.60 ± 0.80 ($n = 24$)	60.75 ± 0.29 ($n = 235$)	0.066
	Chest width (cm)	**21.28 ± 0.95^{A} ($n = 16$)**	**18.04 ± 0.40^{B} ($n = 24$)**	**20.05 ± 0.20^{A} ($n = 235$)**	**0.002**
	Chest depth (cm)	29.56 ± 0.63^{Ab} ($n = 16$)	28.40 ± 0.34^{b} ($n = 24$)	29.69 ± 0.20^{A} ($n = 235$)	0.135
	Body length (cm)	**65.34 ± 0.88^{AB} ($n = 16$)**	**63.65 ± 0.53^{B} ($n = 24$)**	**66.49 ± 0.34^{A} ($n = 233$)**	**1.230E-4**
	Cannon circumference (cm)	**7.98 ± 0.13^{B} (n =16)**	**7.75 ± 0.10^{B} ($n = 24$)**	**8.39 ± 0.05^{A} ($n = 236$)**	**1.300E-5**
	Heart girth (cm)	**88.38 ± 1.60^{a} ($n = 16$)**	**83.15 ± 1.12^{b} ($n = 24$)**	**86.69 ± 0.46^{a} ($n = 236$)**	**0.033**
	Body height (kg)	**60.16 ± 0.96^{A} ($n = 16$)**	**56.08 ± 0.79^{B} ($n = 24$)**	**58.22 ± 0.27^{A} ($n = 235$)**	**0.008**
CNV3	Height at hip cross (cm)	59.98 ± 0.74 ($n = 23$)	60.08 ± 1.24 ($n = 12$)	60.75 ± 0.26 ($n = 262$)	0.625
	Chest width (cm)	**18.78 ± 0.44^{AB} ($n = 23$)**	**17.46 ± 0.61^{B} ($n = 12$)**	**20.03 ± 0.19^{A} ($n = 263$)**	**0.004**
	Chest depth (cm)	30.48 ± 0.88 ($n = 23$)	29.71 ± 1.00 ($n = 12$)	29.22 ± 0.16 ($n = 263$)	0.109
	Body length (cm)	66.44 ± 0.90 ($n = 23$)	65.29 ± 1.42 ($n = 12$)	66.54 ± 0.28 ($n = 261$)	0.649
	Cannon circumference (cm)	**7.87 ± 0.12^{B} (n =22)**	**7.92 ± 0.16^{AB} ($n = 12$)**	**8.31 ± 0.05^{A} ($n = 262$)**	**0.009**
	Heart girth (cm)	89.65 ± 1.51 ($n = 22$)	87.71 ± 2.09 ($n = 12$)	88.81 ± 0.51 ($n = 262$)	0.795
	Body height (kg)	60.84 ± 0.48 ($n = 89$)	56.96 ± 0.93 ($n = 12$)	58.21 ± 0.24 ($n = 262$)	0.552
CNV4	Height at hip cross (cm)	60.63 ± 0.55 ($n = 48$)	60.84 ± 0.48 ($n = 89$)	60.39 ± 0.37 ($n = 143$)	0.742
	Chest width (cm)	**18.08 ± 0.30^{c} ($n = 48$)**	**20.50 ± 0.34^{A} ($n = 89$)**	**19.39 ± 0.25^{b} ($n = 144$)**	**8.166E-7**
	Chest depth (cm)	29.68 ± 0.51 ($n = 48$)	29.44 ± 0.31 ($n = 89$)	29.14 ± 0.23 ($n = 144$)	0.503
	Body length (cm)	65.27 ± 0.69 ($n = 48$)	66.64 ± 0.51 ($n = 89$)	66.51 ± 0.39 ($n = 144$)	0.223
	Cannon circumference (cm)	**7.90 ± 0.10^{B} (n =47)**	**8.35 ± 0.08^{A} ($n = 89$)**	**8.23 ± 0.06^{A} ($n = 143$)**	**0.001**
	Heart girth (cm)	**88.42 ± 1.25^{b} ($n = 47$)**	**91.61 ± 0.87^{A} ($n = 89$)**	**87.12 ± 0.66^{b} ($n = 143$)**	**2.620E-4**
	Body height (kg)	58.05 ± 0.55 ($n = 48$)	58.05 ± 0.42 ($n = 89$)	58.00 ± 0.35 ($n = 144$)	0.996
CNV5	Height at hip cross (cm)	60.33 ± 0.60 ($n = 44$)	60.21 ± 0.42 ($n = 116$)	60.50 ± 0.29 ($n = 212$)	0.837
	Chest width (cm)	19.40 ± 0.50 ($n = 44$)	20.16 ± 0.26 ($n = 116$)	19.82 ± 0.23 ($n = 212$)	0.360
	Chest depth (cm)	**28.75 ± 0.44^{B} ($n = 44$)**	**29.84 ± 0.20^{A} ($n = 116$)**	**28.95 ± 0.19^{B} ($n = 212$)**	**0.008**
	Body length (cm)	**64.59 ± 0.78^{B} ($n = 43$)**	**66.85 ± 0.38^{a} ($n = 116$)**	**65.93 ± 0.35^{aB} (n = 211)**	**0.025**
	Cannon circumference (cm)	8.08 ± 0.09 ($n = 43$)	8.32 ± 0.06 ($n = 117$)	8.18 ± 0.05 ($n = 211$)	0.097
	Heart girth (cm)	90.01 ± 1.18 ($n = 43$)	87.74 ± 0.69 ($n = 117$)	87.83 ± 0.56 ($n = 211$)	0.220
	Body height (kg)	57.92 ± 0.66 ($n = 44$)	57.86 ± 0.35 ($n = 117$)	57.83 ± 0.29 ($n = 211$)	0.990

Values with different letters (A, B, C/a, b, c) within the same row differ significantly at $P < 0.01$/$P < 0.05$. AVG, means average; SE, means standard error. The bold values indicate the value of $P < 0.05$.

for CNV2, the remaining CNV loci were significantly associated among the SBWC goats, GB goats, and NB goats ($P < 0.01$) (Table 2).

Discussion

Relevant studies have shown that *SNX7* (34, 35), *SNX8* (36), *SNX9* (37), *SNX10* (38), and *SNX19* genes (39) were associated with animal growth traits. As a member of the same family, we speculated that the *SNX29* CNVs may have a remarkable influence on growth traits. To preliminarily explore the function of the *SNX29* gene, the goat *SNX29* protein structure was predicted using an online platform. The results showed that the *SNX29* protein was hydrophilic and had no transmembrane helix and signal peptide, and it is a non-secretory protein and performed a relevant function in the cytoplasm, which was consistent with the previous description (40).

To further explore the relationship between this gene and growth traits, we conducted population validation. Five CNVs were retrieved from the database. After population distribution detection, it was found that the genotypes of goats of the three breeds were different at different loci. This is because genetic variations vary from breed to breed (41). In the three goat breeds, more individuals performed gain genotype. This may be because the gain genotype showed better economic efficiency and was retained in artificial selection. Notably, the association

TABLE 2 Genotype distribution among the SBWC goats, GB goats, and NB goats.

CNV Loci	Breeds	Size	Genotypic frequencies			χ^2	*P*-value
			Loss	Median	Gain		
CNV1	SBWC	276	16	24	236	14.012	**0.007**
	GB	48	0	0	48		
	NB	38	0	0	38		
CNV2	SBWC	291	0	2	289	0.587	0.746
	GB	47	0	0	47		
	NB	38	0	0	38		
CNV3	SBWC	363	89	12	262	30.873	**3.000E-6**
	GB	48	0	0	48		
	NB	38	0	0	38		
CNV4	SBWC	281	48	89	144	44.819	**4.335E-9**
	GB	46	0	9	37		
	NB	38	0	0	38		
CNV5	SBWC	372	44	117	211	23.749	**9.000E-5**
	GB	47	0	24	23		
	NB	38	0	6	32		

SBWC, Shaanbei white cashmere goats; GB, means Guizhou black goats; NB, Nubian goats. The bold values indicate the value of $P < 0.05$.

analysis showed that four CNVs were observably associated with chest width, body length, body height, cannon circumference, and chest circumference ($P < 0.05$) in SBWC goats, which supports our conjecture. Moreover, we found that in CNV1 individuals, the gain and loss genotypes were superior to those with the median genotypes in terms of growth traits, but in CNV4 and CNV5 individuals, the median genotypes were better than the loss and gain genotypes. In addition, in the CNV3, the gain genotype performed better growth traits, which could be due to mutation, selection, gene recombination, and genetic drift migration (42). These outcomes suggest that the gain/loss genotype of CNV1, the gain genotype of CNV3, and the median genotype of CNV4 and CNV5 have a positive effect on growth traits (43).

In this study, we found that the CNVs of *SNX29* were associated with the growth traits of goats, which is consistent with the function of *SNX29* in previous studies associated with growth. A genome-wide scan identified the growth-related SNP markers of *SNX29* in Chinese Wagyu cattle (44). Genome-wide association analysis showed that CNV27 of the *SNX29* gene was associated with growth traits of African goats (45), and also two InDels within this gene are significantly correlated with chest width, hip width, and other growth traits in goats (46). In addition, this gene has shown growth-related functions in different species. In York pigs, genome-wide association analysis of five meat quality traits found that 12 intron SNPs of the *SNX29* gene were associated with intramuscular fat content (47). Therefore, the *SNX29* has been identified as a candidate gene associated with growth traits, whose CNVs can also act as an influence on the growth traits of livestock. We will continue to explore the molecular mechanism between this gene and growth traits in further studies.

Conclusion

In this study, the growth effect of the *SNX29* gene was elucidated from the aspects of protein structure, physicochemical properties, and DNA variation. The protein encoded by *SNX29* was a non-secreted protein, whose five CNVs were identified in SBWC goats, GB goats, and NB goats. Moreover, CNVs were found to be associated with growth traits in SBWC goats. The CNV1, CNV3, CNV4, and CNV5 were significantly associated with the SBWC goats, GB goats, and NB goats ($P < 0.01$). Thus, the *SNX29* gene may be an essential functional candidate gene for growth traits.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The animal study was reviewed and approved by Northwest A&F University. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

XS, HZ, XW, ZG, and CP: sample collection. QW and YB: experimental operation. QW, YB, HZ, XW, and ZG: data collation and analysis. QW: article writing. QW, YB, and CP: manuscript revision and editing. ML and CP: project management. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (No.32002166).

Acknowledgments

The authors sincerely thank the Shaanxi Yulin goat farm, Guizhou Bijie goat farm, and Fujian ZhangZhou goat cooperative for providing them with samples. The authors would also like to thank Lei Qu, Hailong Yan, and Jinwang Liu from Yulin University for their help in sample collection, and XW for her help at the Institute of Animal Husbandry and Veterinary, Fujian Academy of Agricultural Sciences.

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/fvets.2023.1132833/full#supplementary-material

References

1. Hao X, Wang Y, Ren F, Zhu S, Ren Y, Jia B, et al. SNX25 regulates TGF-β signaling by enhancing the receptor degradation. *Cell Signal.* (2011) 23:935–46. doi: 10.1016/j.cellsig.2011.01.022

2. Lin YJ, Chang JS, Liu X, Lin TH, Huang SM, Liao CC, et al. Sorting nexin 24 genetic variation associates with coronary artery aneurysm severity in Kawasaki disease patients. *Cell Biosci.* (2013) 3:44. doi: 10.1186/2045-3701-3-44

3. Amatya B, Lee H, Asico LD, Konkalmatt P, Armando I, Felder RA, et al. Subfamily of SNXs in the regulation of receptor-mediated signaling and membrane trafficking. *Int J Mol Sci.* (2021) 22:2319. doi: 10.3390/ijms22052319

4. Xia L, Ou J, Li K, Guo H, Hu Z, Bai T, et al. Genome-wide association analysis of autism identified multiple loci that have been reported as strong signals for neuropsychiatric disorders. *Autism Res.* (2020) 13:382–96. doi: 10.1002/aur.2229

5. Chen JH, Zhao Y, Khan RAW, Li ZQ, Zhou J, Shen JW, et al. SNX29, a new susceptibility gene shared with major mental disorders in Han Chinese population. *World J Biol Psychiatry.* (2021) 22:526–34. doi: 10.1080/15622975.2020.1845793

6. Twa DD, Mottok A, Chan FC, Ben-Neriah S, Woolcock BW, Tan KL, et al. Recurrent genomic rearrangements in primary testicular lymphoma. *J Pathol.* (2015) 236:136–41. doi: 10.1002/path.4522

7. Zhu L, Hu Z, Liu J, Gao J, Lin B. Gene expression profile analysis identifies metastasis and chemoresistance-associated genes in epithelial ovarian carcinoma cells. *Med Oncol.* (2015) 32:426. doi: 10.1007/s12032-014-0426-5

8. Sparks AM, Watt K, Sinclair R, Pilkington JG, Pemberton JM, McNeilly TN, et al. The genetic architecture of helminth-specific immune responses in a wild population of Soay sheep (Ovis aries). *PLoS Genet.* (2019) 15:e1008461. doi: 10.1371/journal.pgen.1008461

9. Peng S, Song C, Li H, Cao X, Ma Y, Wang X, et al. Circular RNA SNX29 sponges miR-744 to regulate proliferation and differentiation of myoblasts by activating the Wnt5a/Ca2+ signaling pathway. *Mol Therapy Nucl Acids.* (2019) 16:481–93. doi: 10.1016/j.omtn.2019.03.009

10. Chen Y, Yang L, Lin X, Peng P, Shen W, Tang S, et al. Effects of genetic variation of the sorting nexin 29 (SNX29) gene on growth traits of xiangdong black goat. *Animals.* (2022) 12:3461. doi: 10.3390/ani12243461

11. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, Vieaud A., et al. (2009). Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. iPLoS Geneti. 5, e1000512. doi: 10.1371/journal.pgen.1000512

12. Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, et al. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLoS Genet.* (2019) 15:e1008536. doi: 10.1371/journal.pgen.1008536

13. Zhang RQ, Wang JJ, Zhang T, Zhai HL, Shen W. Copy-number variation in goat genome sequence: a comparative analysis of the different litter size trait groups. *Gene.* (2019) 696:40–6. doi: 10.1016/j.gene.2019.02.027

14. Wang, L., Xu, L., Liu, X., Zhang, T., Li, N., Hay el, H., et al. Copy number variation-based genome wide association study reveals additional variants contributing to meat quality in Swine. *Sci Rep.* (2015) 5:12535. doi: 10.1038/srep12535

15. Kang X, Li M, Liu M, Liu S, Pan MG, Wiggans GR, et al. Copy number variation analysis reveals variants associated with milk production traits in dairy goats. *Genomics.* (2020) 112:4934–7. doi: 10.1016/j.ygeno.2020.09.007

16. Fernandes AC, da Silva VH, Goes CP, Moreira GC, Godoy TF, Ibelli AM, et al. Genome-wide detection of CNVs and their association with performance traits in broilers. *BMC Genom.* (2021) 22:354. doi: 10.1186/s12864-021-07676-1

17. Strillacci MG, Gorla E, Ríos-Utrera A, Vega-Murillo VE, Montaño-Bermudez M, Garcia-Ruiz A, et al. Copy number variation mapping and genomic variation of autochthonous and commercial turkey populations. *Front Genet.* (2019) 10:982. doi: 10.3389/fgene.2019.00982

18. Pös O, Radvanszky J, Buglyó G, Pös Z, Rusnakova D, Nagy B, et al. Copy number variation: main characteristics, evolutionary significance, and pathological aspects. *Biomed J.* (2021) 44:548–59. doi: 10.1016/j.bj.2021.02.003

19. Bi Y, Feng B, Wang Z, Zhu H, Qu L, Lan X, et al. Myostatin (MSTN) gene indel variation and its associations with body traits in Shaanbei White Cashmere Goat. *Animals.* (2020) 10:168. doi: 10.3390/ani10010168

20. Wei Z, Wang K, Wu H, Wang Z, Pan C, Chen H, et al. Detection of 15-bp deletion mutation within PLAG1 gene and its effects on growth traits in goats. *Animals.* (2021) 11:2064. doi: 10.3390/ani11072064

21. Cai HF, Chen Z, Luo WX. Associations between polymorphisms of the GFI1B gene and growth traits of indigenous Chinese goats. *Genet Mol Res.* (2014) 13:872–80. doi: 10.4238/2014.February.13.5

22. Kholif AE, Gouda GA, Hamdon HA. Performance and milk composition of Nubian goats as affected by increasing level of nannochloropsis oculata microalgae. *Animals.* (2020) 10:2453. doi: 10.3390/ani10122453

23. Li X, Ding X, Liu L, Yang P, Yao Z, Lei C, et al. Copy number variation of bovine DYNC1I2 gene is associated with body conformation traits in Chinese beef cattle. *Gene.* (2022) 810:146060. doi: 10.1016/j.gene.2021.146060

24. Aljanabi SM, Martinez I. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res.* (1997) 25:4692–3. doi: 10.1093/nar/25.22.4692

25. Cui Y, Chen R, Lv X, Pan C. Detection of coding sequence, mRNA expression and three insertions/deletions (indels) of KDM6A gene in male pig. *Theriogenology.* (2019) 133:10–21. doi: 10.1016/j.theriogenology.2019.04.023

26. Zhang X, Yu S, Yang Q, Wang K, Zhang S, Pan C, et al. Goat boule: isoforms identification, mRNA expression in testis and functional study and promoter methylation profiles. *Theriogenology.* (2018) 116:53–63. doi: 10.1016/j.theriogenology.2018.05.002

27. Fu W, Wang R, Yu J, Hu D, Cai Y, Shao J, et al. A goat genome variation database for tracking the dynamic evolutionary process of selective signatures and ancient introgressions. *J Genet Genom.* (2021) 48:248–56. doi: 10.1016/j.jgg.2021.03.003

28. Wang Q, Bi Y, Wang Z, Zhu H, Liu M, Wu X, et al. Goat SNX29: mRNA expression, indel and CNV detection, and their associations with litter size. *Front Vet Sci.* (2022) 9:981315. doi: 10.3389/fvets.2022.981315

29. Li J, Zhang S, Erdenee S, Sun X, Dang R, Huang Y, et al. Nucleotide variants in prion-related protein (testis-specific) gene (PRNT) and effects on Chinese and Mongolian sheep phenotypes. *Prion.* (2018) 12:185–96. doi: 10.1080/19336896.2018.1467193

30. Yang Q, Zhang S, Li J, Wang X, Peng K, Lan X, et al. Development of a touch-down multiplex PCR method for simultaneously rapidly detecting three novel insertion/deletions (indels) within one gene: an example for goat GHR gene. *Anim Biotechnol.* (2019) 30:366–71. doi: 10.1080/10495398.2018.1517770

31. Bi Y, Feng W, Kang Y, Wang K, Yang Y, Qu L, et al. Detection of mRNA expression and copy number variations within the goat FecB gene associated with litter size. *Front Vet Sci.* (2021) 8:758705. doi: 10.3389/fvets.2021.758705

32. Yang Y, Hu H, Mao C, Jiang F, Lu X, Han X, et al. Detection of the 23-bp nucleotide sequence mutation in retinoid acid receptor related orphan receptor alpha (RORA) gene and its effect on sheep litter size. *Anim Biotechnol.* (2020) 33:70–8. doi: 10.1080/10495398.2020.1770273

33. Liu H, Xu H, Lan X, Cao X, Pan C. The InDel variants of sheep IGF2BP1 gene are associated with growth traits. *Anim Biotechnol.* (2021) 13:1–9. doi: 10.1080/10495398.2021.1942029

34. Edea Z, Hong JK, Jung JH, Kim DW, Kim YM, Kim ES, et al. Detecting selection signatures between Duroc and Duroc synthetic pig populations using high-density SNP chip. *Anim Genet.* (2017) 48:473–7. doi: 10.1111/age.12559

35. Lin S, Zhang H, Hou Y, Liu L, Li W, Jiang J, et al. discovery and functional candidate gene identification for milk composition based on whole genome resequencing of Holstein bulls with extremely high and low breeding values. *PLoS ONE.* (2019) 14:e0220629. doi: 10.1371/journal.pone.0220629

36. Muirhead G, Dev KK. The expression of neuronal sorting nexin 8 (SNX8) exacerbates abnormal cholesterol levels. *J Mol Neurosci.* (2014) 53:125–34. doi: 10.1007/s12031-013-0209-z

37. An B, Xu L, Xia J, Wang X, Miao J, Chang T, et al. Multiple association analysis of loci and candidate genes that regulate body size at three growth stages in Simmental beef cattle. *BMC Genet.* (2020) 21:32. doi: 10.1186/s12863-020-0837-6

38. Castillejo-Lopez C, Pjanic M, Pirona AC, Hetty S, Wabitsch M, Wadelius C, et al. Detailed functional characterization of a waist-hip ratio locus in 7p15.2 defines an enhancer controlling adipocyte differentiation. *iScience.* (2019) 20:42–59. doi: 10.1016/j.isci.2019.09.006

39. Guo L, Sun H, Zhao Q, Xu Z, Zhang Z, Liu D, et al. Positive selection signatures in Anqing six-end-white pig population based on reduced-representation genome sequencing data. *Anim Genet.* (2021) 52:143–54. doi: 10.1111/age.13034

40. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* (2000) 28:231–4. doi: 10.1093/nar/28.1.231

41. Zhang Y, Wang K, Liu J, Zhu H, Qu L, Chen H, et al. An 11-bp Indel polymorphism within the CSN1S1 gene is associated with milk performance and body measurement traits in Chinese Goats. *Animals.* (2019) 9:1114. doi: 10.3390/ani9121114

42. Huang Y, Su P, Akhatayeva Z, Pan C, Zhang Q, Lan X, et al. Novel InDel variations of the Cry2 gene are associated with litter size in Australian White sheep. *Theriogenology.* (2022) 179:155–61. doi: 10.1016/j.theriogenology.2021.11.023

43. Zhou T, Wei H, Li D, Yang W, Cui Y, Gao J, et al. Novel missense mutation within the domain of lysine demethylase 4D (KDM4D) gene is strongly associated with testis morphology traits in pigs. *Anim Biotechnol.* (2020) 31:52–8. doi: 10.1080/10495398.2018.1531880

44. Wang Z, Ma H, Xu L, Zhu B, Liu Y, Bordbar F, et al. Genome-wide scan identifies selection signatures in Chinese Wagyu cattle using a high-density SNP array. *Animals.* (2019) 9:296. doi: 10.3390/ani9060296

45. Liu M, Woodward-Greene J, Kang X, Pan MG, Rosen B, Van Tassell CP, et al. Genome-wide CNV analysis revealed variants associated with growth traits in African indigenous goats. *Genomics.* (2020) 112:1477–80. doi: 10.1016/j.ygeno.2019.08.018

46. Bi Y, Chen Y, Xin D, Liu T, He L, Kang Y, et al. Effect of indel variants within the sorting nexin 29 (SNX29) gene on growth traits of goats. *Anim Biotechnol.* (2020) 19:1–6. doi: 10.1080/10495398.2020.1846547

47. Dong Q, Liu H, Li X, Wei W, Zhao S, Cao JA, et al. genome-wide association study of five meat quality traits in Yorkshire pigs. *Front Agric Sci Eng.* (2014) 1:137–43. doi: 10.15302/J-FASE-2014014

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/fgene.2023.1113411

OPEN ACCESS

EDITED BY
Ran Di,
Chinese Academy of Agricultural Sciences, China

REVIEWED BY
Cuijuan Han,
Jackson Laboratory, United States
Zengkui Lu,
Chinese Academy of Agricultural Sciences, China

*CORRESPONDENCE
Xiaoxue Zhang,
zhangxx@gsau.edu.cn
Wenxin Zheng,
zwx2020@126.com

RECEIVED 01 December 2022
ACCEPTED 09 October 2023
PUBLISHED 19 October 2023

CITATION
Lin C, Wang W, Zhang D, Huang K, Zhang Y, Li X, Zhao Y, Zhao L, Wang J, Zhou B, Cheng J, Xu D, Li W, Zhang X and Zheng W (2023), Analysis of liver miRNA in Hu sheep with different residual feed intake.
Front. Genet. 14:1113411.
doi: 10.3389/fgene.2023.1113411

Analysis of liver miRNA in Hu sheep with different residual feed intake

Changchun Lin[1,2], Weimin Wang[3], Deyin Zhang[3], Kai Huang[3], Yukun Zhang[3], Xiaolong Li[3], Yuan Zhao[3], Liming Zhao[3], Jianghui Wang[1], Bubo Zhou[1], Jiangbo Cheng[3], Dan Xu[1], Wenxin Li[1], Xiaoxue Zhang[1]* and Wenxin Zheng[2]*

[1]College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, Gansu, China, [2]Institute of Animal Husbandry Quality Standards, Xinjiang Academy of Animal Sciences, Urumqi, Xinjiang, China, [3]The State Key Laboratory of Grassland Agro-ecosystems, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, Gansu, China

Feed efficiency (FE), an important economic trait in sheep production, is indirectly assessed by residual feed intake (RFI). However, RFI in sheep is varied, and the molecular processes that regulate RFI are unclear. It is thus vital to investigate the molecular mechanism of RFI to developing a feed-efficient sheep. The miRNA-sequencing (RNA-Seq) was utilized to investigate miRNAs in liver tissue of 6 out of 137 sheep with extreme RFI phenotypic values. In these animals, as a typical metric of FE, RFI was used to distinguish differentially expressed miRNAs (DE_miRNAs) between animals with high (n = 3) and low (n = 3) phenotypic values. A total of 247 miRNAs were discovered in sheep, with four differentially expressed miRNAs (DE_miRNAs) detected. Among these DE_miRNAs, three were found to be upregulated and one was downregulated in animals with low residual feed intake (Low_RFI) compared to those with high residual feed intake (High_RFI). The target genes of DE_miRNAs were primarily associated with metabolic processes and biosynthetic process regulation. Furthermore, they were also considerably enriched in the FE related to glycolysis, protein synthesis and degradation, and amino acid biosynthesis pathways. Six genes were identified by co-expression analysis of DE_miRNAs target with DE_mRNAs. These results provide a theoretical basis for us to understand the sheep liver miRNAs in RFI molecular regulation.

KEYWORDS

miRNA, residual feed intake, gene interactions, liver, sheep

1 Introduction

Feed efficiency (FE), an important economic trait in sheep production, is indirectly assessed by residual feed intake (RFI) and feed conversion ratio (FCR) (Carberry et al., 2012; Zhang et al., 2017a; Claffey et al., 2018; McGovern et al., 2018). RFI is defined as the discrepancy between the amount of feed actually consumed and amount anticipated to be needed for maintenance and growth (Mebratie et al., 2019). Improved FE has the potential to reduce meat production costs, with feed and feeding-related expenses accounting for 75% of total variable production costs in beef cattle farming (Ahola and Hill, 2012). Zhang et al. (2017b) showed in indoor sheep husbandry, feed expenditures account for 65%–70% of overall costs. In addition, research has been shown that enhancing ruminant FE may

effectively mitigate greenhouse gas emissions and provide positive environmental outcomes (Nkrumah et al., 2006; Hegarty et al., 2007; Deng et al., 2018). Thus, livestock producers have a keen interested in the domain of genetic selection and breeding, particularly with regard to enhancing FE of their animals. However, the precise definition of FE in animals is currently being disputed, due to imperfect quality of ratios such as FCR (Gunsett, 1984). Therefore, RFI serves only as a metric for evaluating FE within the context of animal production. RFI is influenced by a variety of internal and external environmental variables such as body composition, nutrition digestion and metabolism, energy expenditure, physical activity, and control of body temperature (Zhang et al., 2017b). Recently, there has been a growing interest in the subject of FE within the context of livestock and poultry production, and researches on RFI-related genes have mainly focused on *swine* (Do et al., 2014; Jing et al., 2015; Horodyska et al., 2017; Messad et al., 2019), cattle (Santana et al., 2014; Salleh et al., 2018) and poultry (Yi et al., 2015). Animals with low_RFI exhibit reduced feed intake, and resulting in decreasedless waste and generation, which do not affect the body size, productivity, or weight of the animals (Koch et al., 1963). Therefore, studying the mechanisms of Low_RFI animals will not only reduce costs but also benefit the environment. The liver, being a vital digestive gland and metabolic organ (Xing et al., 2019), plays an important part in the metabolism of lipids, carbohydrates, and glucose metabolism, and has crucial physiological roles in oxidation, metabolism and reduction (El-Badawy et al., 2019; Cigrovski Berkovic et al., 2020; Ndiaye et al., 2020). Given the essential function played by the liver in the metabolic processes of livestock and poultry, it was selected as the sample in this present research.

MicroRNAs (miRNAs) are a class of small (~22 nucleotides) endogenous non-coding RNAs that exhibit a high degree of conservation across different species (Nelson et al., 2011). The miRNAs have been discovered in several physiological fluids, tissues, and cell types, where they play a crucial role in regulating gene expression at the post-transcriptional level, and they are associated with a wide range of important biological processes (Halushka et al., 2018). Previous studies has shown that miRNA exert control over gene expression by their binding to particular messenger RNA (mRNA), which ultimately leads to the subsequent destruction or inhibition of the targeted transcript. miRNA is associated with the regulation of almost all cellular and developmental processes in eukaryotes (Nejad et al., 2018). For instance, it has been shown that miR-1, miR-133a, miR-133b, and miR-206 exhibit increased expression throughout the advanced phases of human of human fetal muscle development (Koutsoulidou et al., 2011). The miR-33 has inhibitory effects on the process of fatty acid breakdown by targeting several genes involved in fatty acid β-oxidation (Gerin et al., 2010). Moreover, miRNAs has been demonstrated to be essential for the development of brain structures and to support critical systems that, if disturbed, may lead to or cause neurodevelopmental disorders (Hollins et al., 2014). Previous studies have shown that a number of miRNAs in the liver tissues of various livestock play an important role in influencing FE. For instance, miR-338 influences fatty acid synthase (Xing et al., 2019), miR-185 affects glucose and lipid metabolism (Li et al., 2016), and miR-545-3p in pig liver affects fat deposition (Chu et al., 2017). In the liver of cattle, miR-19b regulates lipid metabolism of fat, miR-122-3p influences hepatic cholesterol and lipid metabolism, and miR-143 affects insulin signaling and glucose homeostasis (Al-Husseini et al., 2016). However, the precise processes via which miRNAs regulate RFI in sheep have yet to be fully unclear.

Thus, the present study aimed to identify candidate miRNAs that regulate FE to breed a Low_RFI Hu sheep population. We used sequencing to determine transcription differences in liver tissue of sheep with extreme RFI phenotypes.

2 Materials and methods

2.1 Ethical statement

The animal studies were done in accordance with the regulations and guidelines set out by the government of Gansu Province, as well as with the approval of the Animal Health and Ethics Committee of Gansu Agricultural University (Animal Experimentation License No. 2012-2-159).

2.2 Experimental animals and daily management

The experimental animals used in this study have been comprehensively detailed, together with their corresponding management regimens, in previous publications (Zhang et al., 2017b; Zhang et al., 2019). To put it simply, a total of 137 male Hu sheep were obtained from Jinchang Zhongtian Sheep Co., Ltd. (Jinchang China) and transported to Minqin Zhongtian Sheep Co., Ltd. (Minqin China) during the same time frame for the purpose of breeding. The process of weaning was established when the lambs reached 56 days of age. Subsequently, during the initial phases of the experimental study, only lambs displaying resilient development and overall outstanding health were selected as candidates. The lambs were supplied with nourishment in a standardized single pen (0.8 × 1 m), whereby they access to fresh water and food every day until the end of the experiment (180 days of age). The lambs attained an ideal age of 80 days for the start of the official experiment, which was recorded as day one. The performance experiment is concluded when the lambs reach 180 days, hence rendering the official duration of the trial as 100 days. Consequently, all lambs participating in this study had a 2-week transitional period prior before a 10 days pre-feeding phase. Throughout the transitional phase, a regular percentage adjustment was made to the form of feed utilized each day. Additionally, the whole pelleted feed was consumed on the initial day of the pre-fed phase. During the first 10-day pre-fed period and ensuing 100-day official trial, the pellet feed was purchased from Gansu Sanyang Jinyuan Animal Husbandry Co., Ltd., (Gansu China).

2.3 Phenotypic measurements and RFI calculation methods

Lambs were treated to a feeding regimen structured in 20-day cycles until the completion of the feeding experiment (180 days of age), with their initial weight being measured on the first day of the

specified period (80 days of age). The lambs underwent daily weighing before to feeding, using a calibrated electronic scale. No modifications were made to the participants or the equipment utilized over the whole period of the experiment. Furthermore, each sheep was weighed for remaining feed before each weighing period, which was used to calculate feed intake and RFI. For the computational model used in this study, the main reference was the formula of Zhang et al. (2017b). The specific formula used in this particular instance was as follows:

$$Y_k = \beta_0 + \beta_1 MBW_k + \beta_2 ADG_k + e_k,$$
$$MBW = [(BW_i + BW_f)/2]^{0.75},$$
$$ADG = (BW_f\text{-}BW_i)/N,$$

where Y_k represents the average daily feed intake of the *i*th individual; β_0 regres-sion intercept; β_1 regression coefficient for mid-test metabolic body weight (MBW); β_2 regression coefficient for average daily gain (ADG); e_k represents uncontrolled error of the *i*th individual; BW_f represents final body weight; BW_i represents initial body weight; and N, experimental period (days).

2.4 Liver tissue collection and total RNA extraction

The methodologies used for the collection and processing methods of the tissues were cited from previous scholarly investigations (Zhang et al., 2017b). The methodology may be concisely summarized in the following manner: all lambs are uniformly transferred to a professional slaughterhouse after the end of the measurement. The RFI values of the six sheep used in this study are provided fully, concerning prior research conducted by our research group. The lambs were subjected to a 24 h of fasting period before to being weighed and then executed in standard procedures. Each liver sample was immediately collected after slaughter process and then preserved in liquid nitrogen for temporary storage. Following the process of slaughter, it was transferred to −80°C for long-term storage until RNA was extracted. As described in previous study, we selected 3 High_RFI and 3 Low_RFI sheep from 137 male Hu lambs for total RNA extraction (Zhang et al., 2017b; Zhang et al., 2019). The total RNA extraction was performed using the TRIzol Reagent (Invitrogen, Waltham, MA, United States) method as per the provided instructions. The NanoPhotometer® spectrophotometer (IMPLEN, CA, United States) was used to quantify the purity of RNA, while the integrity of RNA was performed using the Agilent Bioanalyzer 2,100 system's RNA Nano 6000 Assay Kit (Agilent Technologies, CA, United States).

2.5 Library preparation and small RNA sequencing

The small RNA library preparation kits were used to produce small RNA sequencing libraries (Galina-Pantoja et al., 2006). There are many steps that involved, as seen below: firstly, whole RNA molecule was used as a template to directly connect the 3′ and 5′ ends of the small RNA with the synthetic adaptors. The synthesis cDNA from total RNA was conducted with M-MuLV reverse transcriptase (NEB, United States), followed by amplification of the resulting cDNA in accordance with the recommended protocols provided by Illumina. The PCR products underwent purification and recovery processes using 8% polyacrylamide gels. These gels exhibited the capability to effectively separate DNA fragments with sizes 140 to 160 base pairs. Following that, the DNA fragments that had undergone purification were dissolved in 8 μL of elution solution. In the end, the evaluation of the library's integrity on the Agilent Bioanalyzer 2,100 system may be conducted by using DNA high-sensitivity chip. Once the library has undergone qualification, the product was subjected to sequencing on the Illumina HiSeq 2,500 platform, resulting in the generation of a 50 bp single-ended read.

2.6 Bioinformatics sequence data processing and miRNA expression profiling

The raw data collected by the sequencing equipment was then given a Base Calling analysis with the intention of producing FASTQ files. The acquisition of clean data included the elimination of extraneous information from the raw data via the use of a customized Perl script. Concurrently, Q20, Q30, and GC-contents of the raw data were obtained. Then, for all subsequent analysis, choose certain range of length from clean reads. The Bowtie (Langmead et al., 2009) method was used for the purpose of aligning small RNA tags with the reference sequence, facilitating the assessment of their expression levels on the reference. Then used miRbase 20.0 as a reference to mapping known miRNAs. In order to exclude protein-coding genes, repetitive sequences, rRNA, tRNA, snRNA, and snoRNA, custom scripts are used to extract miRNAs of predetermined length. Following, the software tools miREvo (Wen et al., 2012) and mirdeep2 (Friedländer et al., 2012) were used to map sequences with the sheep reference genome (*Oar_v1.0*) to provide predictions about novel miRNAs. The use of the whole rRNA ratio served as an indicator of sample quality. In animal samples, this ratio should ideally be below 40%. Additionally, the cumulative *p* values for RNA folding were used as a metric for output measure.

2.7 Differential expression analysis and prediction of target genes of miRNAs

The miRNA expression levels were estimated by TPM (transcript per million) (Zhou et al., 2010). The processing procedure differs according on the sample's quality. To perform differential expression analysis on Low_RFI and High_RFI samples with biological recurrence, use the DESeq R package (version 1.8.3). The filtering conditions for differential expression of miRNAs in this study were *p*-value <0.05 and |log2 (foldchange)|≤ 0.5 (Tang et al., 2007). The threshold for substantial differential expression is set to the default value. The target gene of miRNA was then predicted for animals using miRanda (Enright et al., 2003). Statistics calculations were carried out to analyze the expression levels of the most prevalent known miRNAs and novel miRNAs in both experimental groups. Additionally, the most common miRNA or

TABLE 1 Summary of clean reads mapped to the *Ovis aries* reference genome.

Sample	LR1	LR2	LR3	HR1	HR2	HR3
Raw Reads	11,585,179	10,414,448	11,147,151	9,959,923	10,728,970	15,885,959
Clean Reads	11,120,503	10,206,637	10,509,901	9,583,661	10,176,841	15,142,493
Q30 (%)	97.98	98.62	97.82	97.97	97.83	97.81
GC Content (%)	48.91	48.36	49.19	49.19	48.92	48.99
Raw Reads	11,585,179	10,414,448	11,147,151	9,959,923	10,728,970	15,885,959
Total Mapped	10,278,758 (92.43%)	10,014,172 (98.11%)	9,533,396 (90.71%)	8,699,934 (90.78%)	9,391,798 (92.29%)	13,765,901 (90.91%)

Notes: LR1: Low_RFI, No. 1 Hu sheep; LR2: Low_RFI, No. 2 Hu sheep; LR3: Low_RFI, No. 3 Hu sheep; HR1: High_RFI, No. 1 Hu sheep: HR2: High_RFI, No. 2 Hu sheep; HR3: High_RFI, No. 3 Hu sheep; Q30: (Percentage of bases with phred values greater than 30 in the total number of bases); GC, Content: Calculate the sum of the number of bases G and C as a percentage of the overall number of bases.

DE_miRNA expression was estimated in relation to the anticipated mRNA expression of its target gene.

2.8 GO and KEGG enrichment analysis

The target gene candidates of DE_miRNAs were analyzed using Gene Ontology (GO) enrichment analysis ("target gene candidates" in the follows). For GO enrichment analysis, a GOseq-based Wallenius non-central hyper-geometric distribution (Young et al., 2010) was used, which might correct for gene length bias. Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2008) was a database used to understand the advanced functions and benefits of biological systems based on molecular-level information, especially genome sequencing and other highly experimental technologies (http://www.genome.jp/kegg/). To assess the statistical enrichment of target gene candidates in KEGG pathways, we employed the KOBAS (Mao et al., 2005) tool.

2.9 Integral miRNA–mRNA networks analysis

Investigate the possible association between DE_miRNA found in this study and Zhang et al. describes DE_mRNA (Cigrovski Berkovic et al., 2020). The construction of a miRNA-mRNA interaction network was facilitated by using Cytoscape software (Shannon et al., 2003). This network was established by including DE_miRNAs and DE_mRNAs based on their specific roles. Specifically, mRNAs exhibiting discernible association with miRNAs were integrated into the miRNA-mRNA interaction network. The same shape with various colors shows types that are up-or downregulated in DE_miRNAs and DE_mRNAs.

3 Results

3.1 miRNA sequence data and mapping quality

3.2 Illumina sequencing generated more than 10 M (million) high quality raw reads for each of the two groups of Hu sheep (except for High_RFI No.1 Hu sheep) (Table 1). The raw data obtained from sequencing is given to a filtering process so as to generate clean reads. Reads with more than 10% N content were removed first. An average of 124 reads was removed in the Low_RFI group, less than 0.01%, and similarly an average of 133 reads was removed in the High_RFI group, less than 0.01% (Supplementary Table S1). Further, after removing low-quality readings, the Low_RFI group averaged 0.32% and the High_RFI group 0.35% being removed (Supplementary Table S1). The number of deletions resulting from the existence of 5 adapter contamination and the presence of ployA/T/G/C was low, with an average 0.00% and 0.03% in the Low_RFI group 0.00% and 0.06% in the High_RFI group, respectively (Supplementary Table S1). More reads were deleted due to the absence of 3 adapter null or insert null, about 0.97% and 0.95%. All samples were filtered to retain clean reads of 98% or more. Finally, the clean reads of each sample were screened for sRNAs within a certain length range (18~35 nt) for subsequent analysis (Table 1). The length-screened sRNAs were localized to the *Ovis aries* reference genome to analyze the distribution of small RNAs (Table 1). From the clean data, a total of 10,278,758, 10,014,172, 9,533,396, 8,699,934, 9,391,798, and 13,765,901 mapped reads from the LR1 (Low_RFI No.1 Hu sheep), LR2 (Low_RFI No.2 Hu sheep), LR3 (Low_RFI No.3 Hu sheep), HR1 (High_RFI No.1 Hu sheep), HR2 (High_RFI No.2 Hu sheep), and HR3 (High_RFI No.3 Hu sheep) libraries were retrieved, with over 90% mapping to the *Ovis aries* reference genome (Table 1). In terms of miRNA expression level, our research results indicate that genes with a TPM <60 retrieved from RNA-seq accounted for about 75% of the total, whereas high-expressed genes, that is, genes with TPM >60, accounted for around 25% (Table 2). However, HR2 accounted for only 7.81%. Screening miRNAs ranged from 18 to 35 nt in length (Figures 1A, B). Furthermore, a variety of non-coding RNAs (ncRNAs) were discovered, including transfer RNAs (tRNAs), snRNAs and miRNAs (Figures 1C, D). Among the identified ncRNAs, a minute fraction constituted recently discovered miRNAs.

3.2 Known miRNA expression and novel miRNA profiles

We identified 121, 120, and 122 known miRNAs in the High_RFI samples, and 119, 83, and 128 known miRNAs in the Low_RFI

TABLE 2 Analysis of miRNA expression levels.

Sample	LR1	LR2	LR3	HR1	HR2	HR3
0–0.1	62 (24.22%)	46 (17.97%)	37 (14.45%)	58 (22.66%)	145 (56.64%)	38 (14.84%)
0.1–0.3	35 (13.67%)	36 (14.06%)	24 (9.38%)	25 (9.77%)	29 (11.33%)	0 (0.00%)
0.3–3.57	55 (21.48%)	66 (25.78%)	85 (33.20%)	67 (26.17%)	36 (14.06%)	99 (38.67%)
3.57–15	30 (11.72%)	28 (10.94%)	26 (10.16%)	32 (12.50%)	16 (6.25%)	28 (10.94%)
15–60	17 (6.64%)	21 (8.20%)	23 (8.98%)	18 (7.03%)	10 (3.91%)	24 (9.38%)
>60	57 (22.27%)	59 (23.05%)	61 (23.83%)	56 (21.88%)	20 (7.81%)	67 (26.17%)

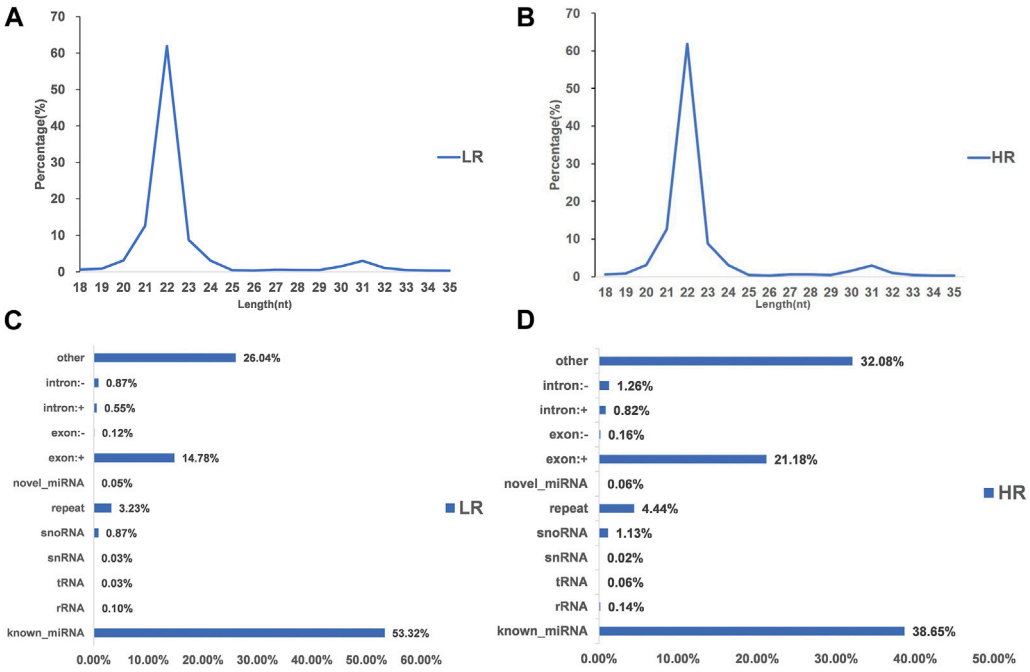


FIGURE 1
Characterization of microRNA (miRNA) profiling and the percentage of detected miRNAs from ncRNAs. **(A, B)** Length distribution of clean reads from identified miRNA fragments. **(C, D)** Categories of identified non-coding RNAs (ncRNAs) via sequencing in Low_FRI and High_FRI. Note "LR: Low_RFI, HR: High_RFI" (the following figures are identical).

TABLE 3 Number and ratio of identified miRNA matrices.

Types	Total	HR1	HR2	HR3	LR1	LR2	LR3
know	139	121	120	122	119	83	128
ratio	100%	100%	87%	86%	88%	86%	60%
novel	117	73	90	97	79	28	90
ratio	100%	62%	77%	83%	68%	24%	77%

samples (Table 3). In all of these reads, approximately 59% of known miRNAs were detected in all samples (Supplementary Table S2). The miRNA the highest abundance across all samples was oar-miR-148a, with mean paired reads from HR1, HR2, HR3, LR1, LR2, and LR3 of 2,885,255, 1,771,013, 2,178,809, 3,049,001, 7,921,051, and 1,323,994, respectively. Among the top 20 expressed miRNAs in each group, the expression levels of 10 miRNAs such as oar-miR-148a, oar-let-7f, oar-miR-143, oar-miR-30a-5p, oar-miR-26a, oar-miR-21, oar-let-7g, oar-let-7i, oar-miR-30d and oar-miR-99a accounted for an average of more than 94% (Supplementary Table S3). The expression analysis shows the top 20 highly expressed miRNAs in the study samples from each respective group (Supplementary Figure S1).

The hairpin structure that is characteristic of miRNA precursors may be used as a means to forecast the existence of novel miRNAs. We identified 73, 90, and 97 novel miRNAs in the High_RFI samples, and 79, 28, and 90 novel miRNAs in the Low_RFI samples (Table 3). Among the novel that were identified, only 23 were identified in all samples (Supplementary Table S3). Of the 117 unique novel miRNAs, novel_31, novel_50 and novel_41were the most expressed in the samples with an average of 1,826, 1,246 and 1,173 reads aligned to these miRNAs, respectively (Supplementary Table S3). All detected new miRNAs were supplied (Supplementary Table S3), and the top 20 expressed novel miRNAs for each group were presented (Supplementary Table S4).

TABLE 4 Four differentially expressed miRNAs in Hu sheep with Low and High_RFI.

miRNA	log2FoldChange	*p*-value	Mature sequence
novel-171	0.97	0.0104	aaucaguaucugucuggguaga
novel-41	0.78	0.0421	ucacugggcauccucugcuuu
oar-miR-485-3p	0.77	0.0459	gucauacacggcucuccucucu
novel-115	−0.82	0.0068	uugcacaacucuagaagacaug

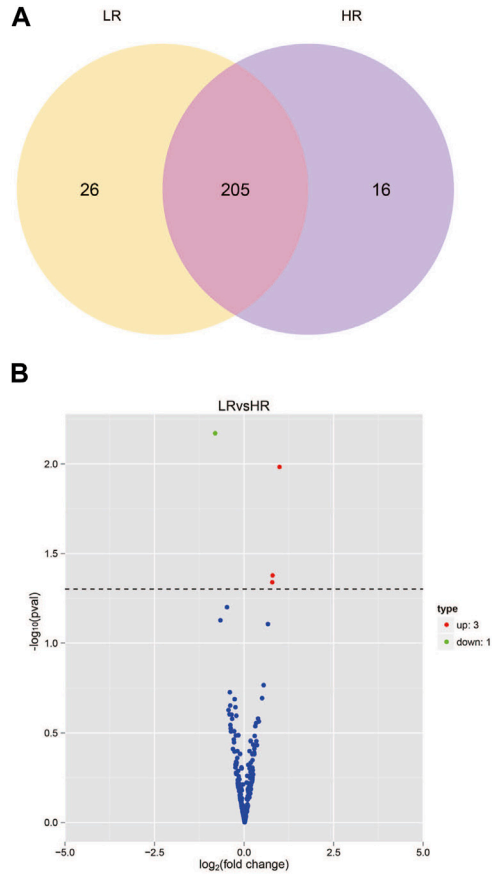


FIGURE 2
miRNA expression in the liver of sheep with different RFI. **(A)** Venn diagram showing the total number of miRNAs expressed in each group individually and in both. **(B)** Volcano map of differentially expressed miRNAs. In the volcano plot, significant downregulated genes are indicated by "green" dots, while significant upregulated genes are indicated by "red" dots.

3.3 miRNA differential expression

The sequencing data has been submitted to the NCBI Sequence Read Archive (SRA) database under the biological project PRJNA813431. Differential miRNA expression analysis between Low_RFI and High_RFI groups from the same population of sheep with different phenotypic values. The relevant phenotypic data for the selected sheep were detailed in previous studies (Zhang et al., 2022). Phenotypic differences were not significant except for FCR, RFI and feed intake (FI). In total, an average of 11.62 million raw read were obtained from each sample. A total of 247 miRNAs were detected in 6 liver samples, of which four miRNAs (one known miRNA and three novel miRNAs) were identified as differentially expressed ($p < 0.05$) (Table 4). The Venn diagrams show miRNAs that are uniquely expressed or co-expressed in different groups, with 205 miRNAs co-expressed in both groups (Figure 2A). Four miRNAs were differentially expressed in the Low_RFI group compared to the High_RFI group, including three upregulated and a downregulated ($p < 0.05$) (Table 4) (Figure 2B). Of all the DE_miRNAs identified, novel_41 and novel_115 were expressed in all samples. The novel_41 was upregulated in Low_RFI group, while novel_115 showed downregulated compared to the High_RFI group (Table 4).

3.4 Target gene prediction and functional enrichment analyses for the most abundant known and novel miRNAs

The target genes of ten most highly expressed miRNAs (seven known and three novel) were predicted in the two groups for further functional analysis (Supplementary Table S5). The majority of these target genes mainly involved in various biological processes, including glycolysis, protein synthesis and catabolism, cell growth and proliferation, scavenging of free radicals, as well as cell death and survival (Supplementary Figure S2). In terms of glycolytic processes, the target genes were mainly involved in nucleoside diphosphate kinase activity, nucleoside kinase activity, adenylate kinase activity, hexokinase activity, and other glucose binding activities. For protein synthesis and catabolism, target genes were involved in the positive regulation of protein secretion, protein polymerization and protein deubiquitination, among other roles. For cell growth and proliferation, target genes were involved in platelet alpha granulation, cell cortex proliferation, and the proliferation and development of microtubule cell ribosomes (Supplementary Figure S2).

3.5 DE_miRNAs target gene prediction

To further understand the biological functions and roles of these four DE_miRNAs, target genes were identified (miRDB) for highly differentiated miRNAs between Low_RFI and High_RFI gruops. Upregulated miRNAs in Low_RFI group were integrated to several target genes: top ranking genes were *DZANK1*, *CYP26B1*, *IDH3G*, *TRAC*, *IL36RN*, *UBE2Z*, *ZMYND12*, *PDGFD*, *VSIG2,* and *ELK4*, whereas top-ranking target genes with downregulated miRNAs were *CFAP221*, *HLA-DOB*, *TOP2B*, *ACSL4*, *FRYL*, *RAB3GAP2*, *TSPAN9*, *ILKAP*, *USP13,* and *PRR36*. The provided diagram illustrates the top 20 anticipated target genes for the four DE_miRNAs (Figure 3). To further elucidate the functions of the DE_miRNAs, we performed enrichment analysis of their candidate target genes. GO enrichment results showed that these target genes were mainly related to metabolism and binding: Biological processes: metabolism, organic metabolism, primary metabolism and macromolecular

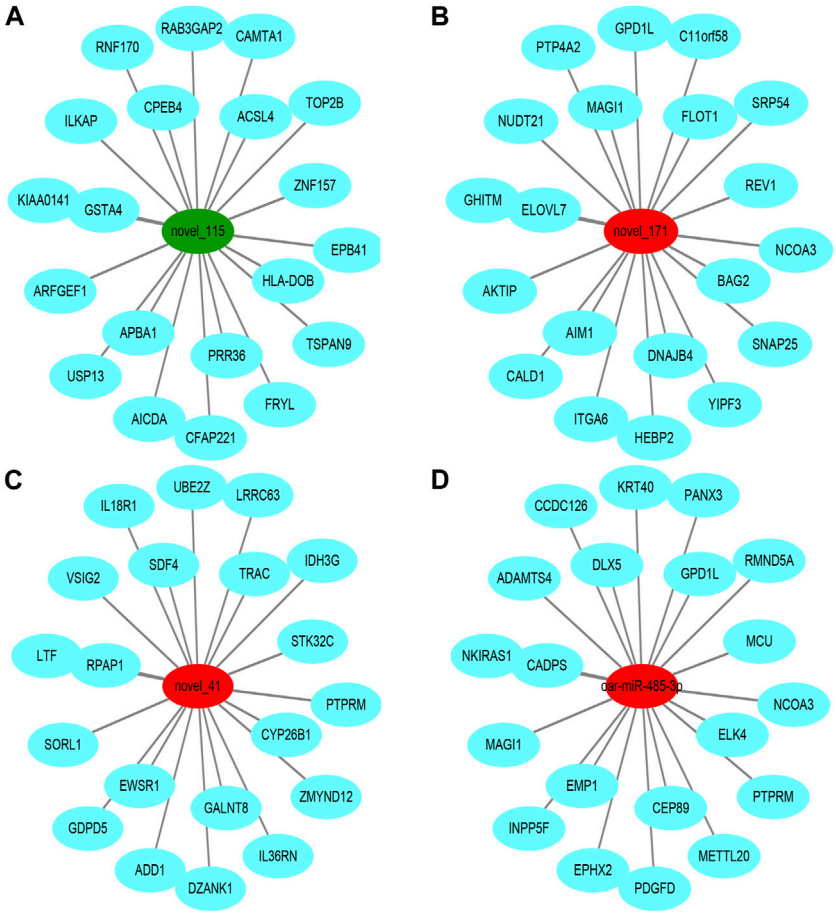


FIGURE 3
Top ranked (based on total target score of miRDB) DE_miRNAs for target genes. In the network plot, significant downregulated genes are indicated by green, while significant upregulated genes are indicated by red. The target gene predicted by psRobottar is shown in blue.

metabolism. Cellular components: intracellular, intracellular fractions, organelles and membrane-bound organelles. Molecular functions: binding, protein binding, catalytic activity and Hydrolytic enzyme activity (Figure 4A). It was shown that the metabolism in the liver plays an important role in the efficiency of animal feed. KEGG pathway showed DE_miRNAs target genes were mainly enriched in transcriptional dysregulation in herpes simplex virus infection, leishmaniasis, and biosynthesis of amino acids (Figure 4B).

3.6 Target gene matching between previously identified DE_mRNAs and the DE_miRNAs prediction

To fully understand the potential RFI effects of miRNA, we used DE_miRNA and their targets genes to create an interactive population network. A total of 1423 DE_miRNAs target genes were identified. We previously discovered 101 DE_mRNAs between the low and high RFI sheep groups using the same objective as the present investigation (Zhang et al., 2022). Among these, seven miRNAs were co-expressed (Figure 5A). However, some target genes were predicted to be the target genes of a single miRNA, and it was observed that upregulation of DE_miRNAs did not always result in downregulation of DE_mRNAs in liver tissue obtained from animals with Low_RFI. Most DE_miRNAs were predicted to primarily target a single differential target gene. As shown in figure, the upregulated miRNA novel_171 targets the upregulated target gene *RTP4*, and the upregulated miRNA novel_41 targets the downregulated target gene *CD274* (Figure 5B). However, the upregulated miRNA oar-miR-485-3p targeted three different genes, the downregulated target gene *OAS1* and the two upregulated target genes *SHISA3* and *PLEKHH2*.

4 Discussion

RNA sequencing can serve as a powerful miRNAs expression profiling tool to identify the DE_miRNAs(Motameny et al., 2010), even at low expression levels in all cells, as well as allows for the parallel analysis of known miRNAs and the identification of miRNAs (Pritchard et al., 2012). Along with the simultaneous analysis of known miRNAs, the examination of novel miRNAs also becomes feasible. Furthermore, the use of mature miRNA sequences may facilitate the identification of prospective target genes for both known and undiscovered

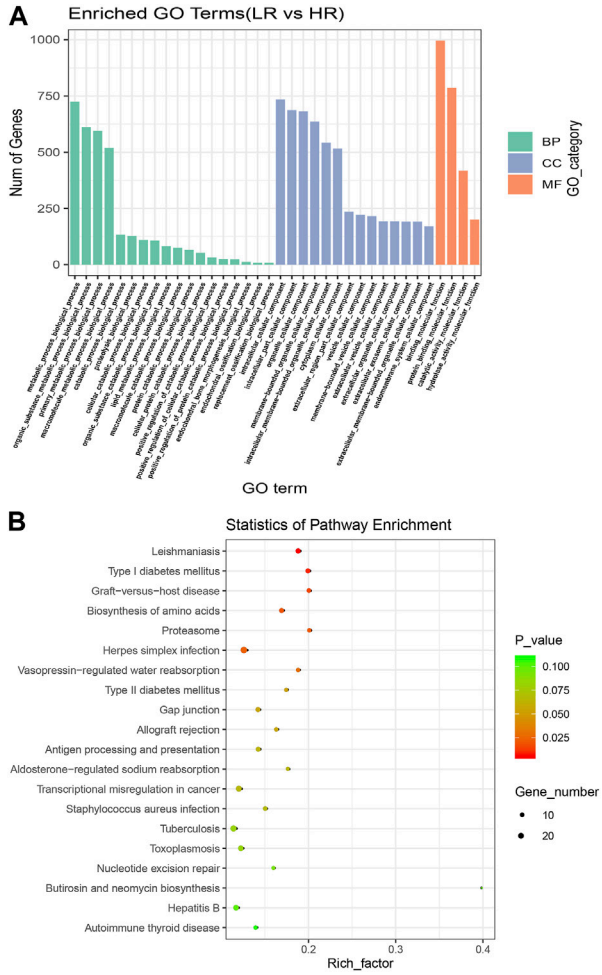


FIGURE 4
Enrichment analysis of high and Low_RFI differential miRNAs. **(A)** Using GO (Gene Ontology) enrichment analysis, BP indicates biological process, CC indicates cellular component, MF indicates molecular function. **(B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) differentially expressed miRNAs target gene enrichment.

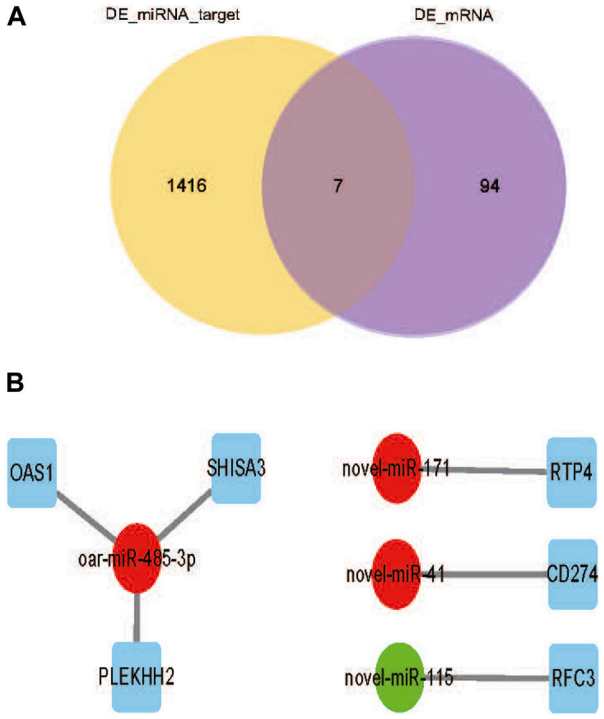


FIGURE 5
(A) Venn diagram showing the number of differential mRNA target genes in yellow, the number of differential miRNA candidate target genes in purple, and the overlapping part indicates the co-expressed part. **(B)** DE_miRNAs are indicated by circles, significant downregulated genes are indicated by green, while significant upregulated genes are indicated by red. And squares indicate DE_miRNAs co-expressing target genes with DE_mRNA.s.

miRNAs. In this study, miRNAs sequencing was used to identify miRNA expression profiles in liver tissue from 6 Hu sheep with extreme RFI from the same farm. Sequencing results showed that sequencing data were of high quality with an average Q30 value of 94%. In addition, after quality control processing of raw sequencing data, the read sequences was an average length of 22 bp, while the majority of reads f were ranged between 20 and 24 bp length from both Low_RFI and High_RFI gruops, providing a high quality and reliable data for subsequent analysis (Friedländer et al., 2012) (Figures 1A, B). The observed average alignment rate, above 90%, indicates a strong agreement between the identified miRNAs and the liver samples. Furthermore, roughly 83% of these miRNAs were found to be expressed in all liver samples, which aligns with the findings of a previous research on miRNAs in bovine liver (Mukiibi et al., 2018). This observation implies that miRNAs exhibit a high degree of conservation within a given population. Among the miRNAs that have been identified, ten highly expressed miRNAs, including oar-miR-148a, oar-let-7f, oar-miR-143, oar-miR-30a-5p, oar-miR-26a, oar-miR-21, oar-let-7g, oar-let-7i, oar-miR-30d and oar-miR-99a, accounted for an average of 92.14% and 95.92% of the total aligned sequence reads in the High and Low_RFI groups, respectively. According to a publication, let-7 miRNA has been detected in various animal species, including humans (Lee et al., 2016). In the present study, it was shown that oar-let-7f, oar-let-7g and oar-let-7i which are members of the let-7 family in sheep, had highly expressed levels in the liver of both groups of FRI sheep. This finding suggests that let-7 family of miRNAs was substantially conserved. The description was consistent with the previously reported results (Friedman et al., 2009). Therefore, from this we speculate that let-7 family miRNAs have the same trend in the same species. Interestingly, oar-miR-148a was the most highly expressed miRNA in all samples, and it belongs to the miR-148/152 family, whose homologous members are involved in a variety of biological functions and diseases in different species. For example, it has been reported that overexpression of miR-148 significantly promotes myogenic differentiation in C2C12-derived myoblasts and primary myoblasts (Zhang et al., 2012). In sheep, miR-148a have been reported to accelerate lipogenic differentiation of sheep preadipocytes and inhibit the proliferation of sheep preadipocytes by inhibiting *PTEN* expression (Jin et al., 2021). Furthermore, it had an inhibitory

effect on the proliferation of Hu sheep hair papilla cells and was associated with hair follicle growth and development (Lv et al., 2019).

To explore the biological significance of sheep-associated DE_miRNAs with varying RFI characteristics. We performed target gene prediction for ten miRNAs that were highly expressed in two groups of RFI sheep. Among these target genes, the main biological functions involved include: negative term regulation of the apoptotic process, cell growth and proliferation, apoptosis and survival, and adipocyte differentiation. Some miRNAs with higher abundance have been discovered as significant regulators of animal cell proliferation and development, apoptosis, and regeneration, which is consistent with our results (Wang et al., 2009). As an example, the second highest expressed miRNA in our research, namely, oar-miR-30a-5p, has been previously associated to lipid and insulin metabolism in mice (Sud et al., 2017; Kim et al., 2019). miR-26a and miR-143 are involved in the regulation of mouse hepatocyte proliferation, a significant aspect in liver tissue regeneration (Geng et al., 2016; Zhou et al., 2019). miR-99a and miR-148a (Gailhouste et al., 2013) were identified as regulators hepatic detoxification in liver tissues of mice and human animals. Based on the observed of miRNA-mRNA interactions across mammalian species and our results of our study, we hypothesized that these miRNAs highly expressed in sheep liver may perform similar biological functions to other species. Moreover, since these highly expressed miRNAs are in a state of continuous self-regeneration or regeneration, it explains their involvement in proliferation as well as apoptosis and regeneration of different cells.

The liver, being the biggest internal organ, plays crucial functions in several physiological metabolic processes, including detoxification (He et al., 2020; Wang et al., 2020). It also serves as a central regulator of energy metabolism, with glycation as a fundamental feature, and is an important coordinator of metabolism and a key site for maintaining metabolic homeostasis (He et al., 2017; Matz et al., 2017; Xue et al., 2019; Moscoso and Steer, 2020). It has been reported that miRNAs are involved in almost every aspect of cell biology (Chen and Verfaillie, 2014). miRNAs play crucial biological roles in cell differentiation, proliferation, metabolism and apoptosis, as well as in viral infection (Kim et al., 2009). Hence, the variable expression of liver miRNAs in Low_RFI and High_RFI sheep might potentially lead to molecular differences in FE. In this study, one known and three novel miRNAs were identified between Low_RFI and High_RFI gruops. However, most of the detected DE_miRNAs (50%) were conservative, which is consistent with the conclusion that miRNAs are conservative (Friedman et al., 2009). For this study to validate the present findings, it would be necessary to conduct more investigations including bigger cohorts of sheep and more broad range of phenotypic animal populations, given that a lower threshold of DE_miRNA screening ($p < 0.05$) was used. In the present study, more than 75% of the DE_miRNAs were upregulated in Low_RFI animals, which was consistent with the results of differential expression analysis of miRNAs in beef cattle with different FE phenotypes (Mukiibi et al., 2020). Thus, this suggests that reduced expression of target genes for these miRNAs is expected.

To investigate the potential biological role of RFI-associated DE_miRNAs in sheep, we predicted their target genes. The target genes are associated with many crucial biological activities, such as metabolic processes, organic metabolism, cell assembly and structure, lipid metabolism, protein breakdown, protein binding, protein metabolism, catalytic activity, and hydrolytic enzyme activity. Among these functions, lipid metabolism and protein synthesis have been reported to be relevant in other species (Chen et al., 2011; Alexandre et al., 2015; Tizioto et al., 2015; Mukiibi et al., 2018). To further understand how DE_miRNAs interact with the 101 DE_mRNAs identified in previous studies (Zhang et al., 2019). Only seven DE_mRNAs (annotated as *RTP4, CD274, OAS1, PLEKHH2, SHISA3*, and *RFC3*) were identified as target genes for DE_miRNAs. These target DE_mRNAs play important roles in innate antiviral response, protein-coupled receptor trafficking, immunity, cell mobility, intercellular signaling and connections, cell death, cell development, differentiation, and gene regulation. Meanwhile, the *RTP4* gene has been shown to be associated with RFI in sheep (Zhang et al., 2022). Certain DE_miRNAs have the potential to impact hepatic functional efficiency FE via their distinct regulatory effects on several biological processes inside the liver. According to the DE_miRNAs- mRNAs interaction network (Figure 5), some single miRNAs were predicted to be targets of single or multiple DE_mRNAs. Because a single miRNA using its seed region may bind to multiple sites in the 3′-UTR of distinct genes (mRNAs), and one target can have multiple binding sites for one or more miRNAs, miRNAs can modulate multiple biological processes even if they are few in number compared to mRNAs they regulate (Ambros, 2004; Bartel, 2004; Brennecke et al., 2005).

Overall, the comparison of DE_miRNAs and DE_mRNAs expression patterns in liver tissue that we identified was consistent with expectation. This may be attributed to the fact that miRNAs accelerate degradation of target genes by promoting the deadenylation of their target transcripts (Stroynowska-Czerwinska et al., 2014). Consequently, we have observe different patterns of DE_miRNA targeting of DE_mRNAs, perhaps attributable to variations in the regulatory mechanisms governing mRNA degradation. To better understand the relationship between miRNAs and mRNAs, further studies at the cellular level are needed to verify these interactions.

5 Conclusion

In the present study, we employed RNA-seq to analyze liver miRNAs in sheep populations. Among these miRNAs, oar-miR-148a, oar-let-7f, oar-miR-143, oar-miR-30a-5p, oar-miR-26a, oar-miR-21, oar-let-7g, oar-let-7i, oar-miR-30d and oar miR-99a had the highest expression levels in all samples. By differential miR-mRNA expression analysis, four miRNAs associated to RFI were discovered, including three novel miRNAs (novel_41, novel_115, and novel_171). Only two miRNAs (novel_41 and novel_115) were expressed in all samples, indicating that most DE_miRNAs were distinct. The predicted target genes of identified DE_miRNAs are involved in a variety of cellular and molecular functions. In addition, only 6.30% of the identified common target genes were found in the liver tissue of the

same subjects. These target genes primarily regulate lipid metabolism, molecular transport, intercellular communication and connections, cell death, and survival. These results provide a theoretical basis for us to understand miRNA expression profile and the molecular mechanisms of miRNA related to FE in sheep liver.

Data availability statement

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

Ethics statement

All animal experiments were conducted out in compliance with the rules and recommendations of Gansu Province's NPC government and were authorized by Gansu Agricultural University's Animal Health and Ethics Committee. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

XZ, CL, and WZ designed the study. XL, YuZ, JW, JC, DX, WL, BZ, and LZ involved in animal husbandry and liver sample collection. YkZ, DZ, KH, and WW correct the manuscript. CL and XZ analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key R&D Program of China (2022YFD1302000), the National for joint research on improved breeds of livestock and poultry (19210365), the West Light Foundation of the Chinese Academy of Sciences (CN), and the China Agriculture Research System (CARS-39).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1113411/full#supplementary-material

SUPPLEMENTARY TABLE S1
Raw data filtering information.

SUPPLEMENTARY TABLE S2
Identification of all known miRNAs information.

SUPPLEMENTARY TABLE S3
Identification of all novel miRNAs information.

SUPPLEMENTARY TABLE S4
Top 20 expressed novel miRNAs information.

SUPPLEMENTARY TABLE S5
Top 10 miRNAs target gene prediction.

SUPPLEMENTARY FIGURE S1
Top 20 expressed known miRNAs information.

SUPPLEMENTARY FIGURE S2
Top 10 miRNAs target gene GO enrichment.

References

Ahola, J. K., and Hill, R. A. (2012). *Input factors affecting profitability: a changing paradigm and a challenging time: feed efficiency in the beef industry*.

Alexandre, P. A., Kogelman, L. J., Santana, M. H., Passarelli, D., Pulz, L. H., Fantinato-Neto, P., et al. (2015). Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. *BMC genomics* 16, 1073. doi:10.1186/s12864-015-2292-8

Al-Husseini, W., Chen, Y., Gondro, C., Herd, R. M., Gibson, J. P., and Arthur, P. F. (2016). Characterization and profiling of liver microRNAs by RNA-sequencing in cattle divergently selected for residual feed intake. *Asian-Australasian J. animal Sci.* 29 (10), 1371–1382. doi:10.5713/ajas.15.0605

Ambros, V. (2004). The functions of animal microRNAs. *Nature* 431 (7006), 350–355. doi:10.1038/nature02871

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 116 (2), 281–297. doi:10.1016/s0092-8674(04)00045-5

Brennecke, J., Stark, A., Russell, R. B., and Cohen, S. M. (2005). Principles of microRNA-target recognition. *PLoS Biol.* 3 (3), e85. doi:10.1371/journal.pbio.0030085

Carberry, C. A., Kenny, D. A., Han, S., McCabe, M. S., and Waters, S. M. (2012). Effect of phenotypic residual feed intake and dietary forage content on the rumen microbial community of beef cattle. *Appl. Environ. Microbiol.* 78 (14), 4949–4958. doi:10.1128/AEM.07759-11

Chen, Y., and Verfaillie, C. M. (2014). MicroRNAs: the fine modulators of liver development and function. *Liver Int.* 34 (7), 976–990. official journal of the International Association for the Study of the Liver. doi:10.1111/liv.12496

Chen, Y., Gondro, C., Quinn, K., Herd, R. M., Parnell, P. F., and Vanselow, B. (2011). Global gene expression profiling reveals genes expressed differentially in cattle with high and low residual feed intake. *Anim. Genet.* 42 (5), 475–490. doi:10.1111/j.1365-2052.2011.02182.x

Chu, A. Y., Deng, X., Fisher, V. A., Drong, A., Zhang, Y., Feitosa, M. F., et al. (2017). Multiethnic genome-wide meta-analysis of ectopic fat depots identifies loci associated with adipocyte development and differentiation. *Nat. Genet.* 49 (1), 125–130. doi:10.1038/ng.3738

Cigrovski Berkovic, M., Virovic-Jukic, L., Bilic-Curcic, I., and Mrzljak, A. (2020). Post-transplant diabetes mellitus and preexisting liver disease - a bidirectional

relationship affecting treatment and management. *World J. Gastroenterol.* 26 (21), 2740–2757. doi:10.3748/wjg.v26.i21.2740

Claffey, N. A., Fahey, A. G., Gkarane, V., Moloney, A. P., Monahan, F. J., and Diskin, M. G. (2018). Effect of breed and castration on production and carcass traits of male lambs following an intensive finishing period. *Transl. animal Sci.* 2 (4), 407–418. doi:10.1093/tas/txy070

Deng, K. D., Xiao, Y., Ma, T., Tu, Y., Diao, Q. Y., Chen, Y. H., et al. (2018). Ruminal fermentation, nutrient metabolism, and methane emissions of sheep in response to dietary supplementation with Bacillus licheniformis. *Animal Feed Sci. Technol.*, S0377840117313950. doi:10.1016/j.anifeedsci.2018.04.014

Do, D. N., Strathe, A. B., Ostersen, T., Pant, S. D., and Kadarmideen, H. N. (2014). Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. *Front. Genet.* 5, 307. doi:10.3389/fgene.2014.00307

El-Badawy, R. E., Ibrahim, K. A., Hassan, N. S., and El-Sayed, W. M. (2019). Pterocarpus santalinus ameliorates streptozotocin-induced diabetes mellitus via anti-inflammatory pathways and enhancement of insulin function. *Iran. J. basic Med. Sci.* 22 (8), 932–939. doi:10.22038/ijbms.2019.34998.8325

Enright, A., John, B., Gaul, U., Tuschl, T., Biology, CSJG, and Marks, D. S. (2003). MicroRNA targets in Drosophila. *MicroRNA targets Drosophila* 5 (11), R1. doi:10.1186/gb-2003-5-1-r1

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids Res.* 40 (1), 37–52. doi:10.1093/nar/gkr688

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19 (1), 92–105. doi:10.1101/gr.082701.108

Gailhouste, L., Gomez-Santos, L., Hagiwara, K., Hatada, I., Kitagawa, N., Kawaharada, K., et al. (2013). miR-148a plays a pivotal role in the liver by promoting the hepatospecific phenotype and suppressing the invasiveness of transformed cells. *Hepatol. Baltim. Md)* 58 (3), 1153–1165. doi:10.1002/hep.26422

Galina-Pantoja, L., Mellencamp, M. A., Bastiaansen, J., Cabrera, R., Solano-Aguilar, G., and Lunney, J. K. (2006). Relationship between immune cell phenotypes and pig growth in a commercial farm. *Anim. Biotechnol.* 17 (1), 81–98. doi:10.1080/10495390500461146

Geng, X., Chang, C., Zang, X., Sun, J., Li, P., Guo, J., et al. (2016). Integrative proteomic and microRNA analysis of the priming phase during rat liver regeneration. *Gene* 575 (2), 224–232. doi:10.1016/j.gene.2015.08.066

Gerin, I., Clerbaux, L. A., Haumont, O., Lanthier, N., Das, A. K., Burant, C. F., et al. (2010). Expression of miR-33 from an SREBP2 intron inhibits cholesterol export and fatty acid oxidation. *J. Biol. Chem.* 285 (44), 33652–33661. doi:10.1074/jbc.M110.152090

Gunsett, F. C. (1984). Linear index selection to improve traits defined as ratios. *J. animal Sci.* 59 (5), 1185–1193. doi:10.2527/jas1984.5951185x

Halushka, M. K., Fromm, B., Peterson, K. J., and McCall, M. N. (2018). Big strides in cellular MicroRNA expression. *Trends Genet. TIG* 34 (3), 165–167. doi:10.1016/j.tig.2017.12.015

He, X., Li, L., Fang, Y., Shi, W., Li, X., and Ma, H. (2017). *In vivo* imaging of leucine aminopeptidase activity in drug-induced liver injury and liver cancer via a near-infrared fluorescent probe. *Chem. Sci.* 8 (5), 3479–3483. doi:10.1039/c6sc05712h

He, B., Shi, J., Wang, X., Jiang, H., and Zhu, H. J. (2020). Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol.* 18 (1), 97. doi:10.1186/s12915-020-00830-3

Hegarty, R. S., Goopy, J. P., Herd, R. M., and McCorkell, B. (2007). Cattle selected for lower residual feed intake have reduced daily methane production. *J. animal Sci.* 85 (6), 1479–1486. doi:10.2527/jas.2006-236

Hollins, S. L., Goldie, B. J., Carroll, A. P., Mason, E. A., Walker, F. R., Eyles, D. W., et al. (2014). Ontogeny of small RNA in the regulation of mammalian brain development. *BMC genomics* 15 (1), 777. doi:10.1186/1471-2164-15-777

Horodyska, J., Hamill, R. M., Varley, P. F., Reyer, H., and Wimmers, K. (2017). Genome-wide association analysis and functional annotation of positional candidate genes for feed conversion efficiency and growth rate in pigs. *PloS one* 12 (6), e0173482. doi:10.1371/journal.pone.0173482

Jin, X., Hao, Z., Zhao, M., Shen, J., Ke, N., Song, Y., et al. (2021). MicroRNA-148a regulates the proliferation and differentiation of ovine preadipocytes by targeting PTEN. Animals: an open access. *J. MDPI* 11 (3), 820. doi:10.3390/ani11030820

Jing, L., Hou, Y., Wu, H., Miao, Y., Li, X., Cao, J., et al. (2015). Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs. *Sci. Rep.* 5, 11953. doi:10.1038/srep11953

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi:10.1093/nar/gkm882

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell. Biol.* 10 (2), 126–139. doi:10.1038/nrm2632

Kim, J. Y., Jun, J. H., Park, S. Y., Yang, S. W., Bae, S. H., and Kim, G. J. (2019). Dynamic regulation of miRNA expression by functionally enhanced placental mesenchymal stem cells PromotesHepatic regeneration in a rat model with bile duct ligation. *Int. J. Mol. Sci.* 20 (21), 5299. doi:10.3390/ijms20215299

Koch, R. M., Swiger, L. A., Chambers, D., and Gregory, K. E. (1963). Efficiency of feed use in beef cattle. *J. animal Sci.* 22 (2), 486–494. doi:10.2527/jas1963.222486x

Koutsoulidou, A., Mastroyiannopoulos, N. P., Furling, D., Uney, J. B., and Phylactou, L. A. (2011). Expression of miR-1, miR-133a, miR-133b and miR-206 increases during development of human skeletal muscle. *BMC Dev. Biol.* 11, 34. doi:10.1186/1471-213X-11-34

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10 (3), R25. doi:10.1186/gb-2009-10-3-r25

Lee, H., Han, S., Kwon, C. S., and Lee, D. (2016). Biogenesis and regulation of the let-7 miRNAs and their functional implications. *Protein and Cell.* 7 (2), 100–113. doi:10.1007/s13238-015-0212-y

Li, Y., Li, X., Sun, W. K., Cheng, C., Chen, Y. H., Zeng, K., et al. (2016). Comparison of liver microRNA transcriptomes of Tibetan and Yorkshire pigs by deep sequencing. *Gene* 577 (2), 244–250. doi:10.1016/j.gene.2015.12.003

Li, P., Fan, C., Cai, Y., Fang, S., Zeng, Y., Zhang, Y., et al. (2020). Transplantation of brown adipose tissue up-regulates miR-99a to ameliorate liver metabolic disorders in diabetic mice by targeting NOX4. *Adipocyte* 9 (1), 57–67. doi:10.1080/21623945.2020.1721970

Lv, X., Gao, W., Jin, C., Wang, L., Wang, Y., Chen, W., et al. (2019). Preliminary study on microR-148a and microR-10a in dermal papilla cells of Hu sheep. *BMC Genet.* 20 (1), 70. doi:10.1186/s12863-019-0770-8

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21 (19), 3787–3793. doi:10.1093/bioinformatics/bti430

Matz, P., Wruck, W., Fauler, B., Herebian, D., Mielke, T., and Adjaye, J. (2017). Footprint-free human fetal foreskin derived iPSCs: a tool for modeling hepatogenesis associated gene regulatory networks. *Sci. Rep.* 7 (1), 6294. doi:10.1038/s41598-017-06546-9

McGovern, E., Kenny, D. A., McCabe, M. S., Fitzsimons, C., McGee, M., Kelly, A. K., et al. (2018). 16S rRNA sequencing reveals relationship between potent cellulolytic genera and feed efficiency in the rumen of bulls. *Front. Microbiol.* 9, 1842. doi:10.3389/fmicb.2018.01842

Mebratie, W., Madsen, P., Hawken, R., Romé, H., Marois, D., Henshall, J., et al. (2019). Genetic parameters for body weight and different definitions of residual feed intake in broiler chickens. *Genet. Sel. Evol.* 51 (1), 53. doi:10.1186/s12711-019-0494-2

Messad, F., Louveau, I., Koffi, B., Gilbert, H., and Gondret, F. (2019). Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. *BMC genomics* 20 (1), 659. doi:10.1186/s12864-019-6010-9

Moscoso, C. G., and Steer, C. J. (2020). The evolution of gene therapy in the treatment of metabolic liver diseases. *Genes* 11 (8), 915. doi:10.3390/genes11080915

Motameny, S., Wolters, S., Nürnberg, P., and Schumacher, B. (2010). Next generation sequencing of miRNAs - strategies, resources and methods. *Genes* 1 (1), 70–84. doi:10.3390/genes1010070

Mukiibi, R., Vinsky, M., Keogh, K. A., Fitzsimmons, C., Stothard, P., Waters, S. M., et al. (2018). Transcriptome analyses reveal reduced hepatic lipid synthesis and accumulation in more feed efficient beef cattle. *Sci. Rep.* 8 (1), 7303. doi:10.1038/s41598-018-25605-3

Mukiibi, R., Johnston, D., Vinsky, M., Fitzsimmons, C., Stothard, P., Waters, S. M., et al. (2020). Bovine hepatic miRNAome profiling and differential miRNA expression analyses between beef steers with divergent feed efficiency phenotypes. *Sci. Rep.* 10 (1), 19309. doi:10.1038/s41598-020-73885-5

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). doi:10.1042/BSR20200219

Nejad, C., Stunden, H. J., and Gantier, M. P. (2018). A guide to miRNAs in inflammation and innate immune responses. *Febs J.* 285 (20), 3695–3716. doi:10.1111/febs.14482

Nelson, P. T., Wang, W. X., Mao, G., Wilfred, B. R., Xie, K., Jennings, M. H., et al. (2011). Specific sequence determinants of miR-15/107 microRNA gene group targets. *Nucleic Acids Res.* 39 (18), 8163–8172. doi:10.1093/nar/gkr532

Nkrumah, J. D., Okine, E. K., Mathison, G. W., Schmid, K., Li, C., Basarab, J. A., et al. (2006). Relationships of feedlot feed efficiency, performance, and feeding behavior with metabolic rate, methane production, and energy partitioning in beef cattle. *J. animal Sci.* 84 (1), 145–153. doi:10.2527/2006.841145x

Pritchard, C. C., Cheng, H. H., and Tewari, M. (2012). MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* 13 (5), 358–369. doi:10.1038/nrg3198

Salleh, S. M., Mazzoni, G., Løvendahl, P., and Kadarmideen, H. N. (2018). Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and

pathways affecting feed efficiency. *BMC Bioinforma.* 19 (1), 513. doi:10.1186/s12859-018-2553-z

Santana, M. H., Utsunomiya, Y. T., Neves, H. H., Gomes, R. C., Garcia, J. F., Fukumasu, H., et al. (2014). Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. *BMC Genet.* 15, 21. doi:10.1186/1471-2156-15-21

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Stroynowska-Czerwinska, A., Fiszer, A., and Krzyzosiak, W. J. (2014). The panorama of miRNA-mediated mechanisms in mammalian cells. *Cell. Mol. life Sci. CMLS* 71 (12), 2253–2270. doi:10.1007/s00018-013-1551-6

Sud, N., Zhang, H., Pan, K., Cheng, X., Cui, J., and Su, Q. (2017). Aberrant expression of microRNA induced by high-fructose diet: implications in the pathogenesis of hyperlipidemia and hepatic insulin resistance. *J. Nutr. Biochem.* 43, 125–131. doi:10.1016/j.jnutbio.2017.02.003

Tang, Y., Ghosal, S., and Roy, A. (2007). Nonparametric bayesian estimation of positive false discovery rates. *Biometrics* 63 (4), 1126–1134. doi:10.1111/j.1541-0420.2007.00819.x

Tizioto, P. C., Coutinho, L. L., Decker, J. E., Schnabel, R. D., Rosa, K. O., Oliveira, P. S., et al. (2015). Global liver gene expression differences in Nelore steers with divergent residual feed intake phenotypes. *BMC genomics* 16 (1), 242. doi:10.1186/s12864-015-1464-x

Wang, Y., Rathinam, R., Walch, A., and Alahari, S. K. (2009). ST14 (suppression of tumorigenicity 14) gene is a target for miR-27b, and the inhibitory effect of ST14 on cell growth is independent of miR-27b regulation. *J. Biol. Chem.* 284 (34), 23094–23106. doi:10.1074/jbc.M109.012617

Wang, C., Li, Y., Li, H., Zhang, Y., Ying, Z., Wang, X., et al. (2020). Disruption of FGF signaling ameliorates inflammatory response in hepatic stellate cells. *Front. Cell. Dev. Biol.* 8, 601. doi:10.3389/fcell.2020.00601

Wen, M., Shen, Y., Shi, S., and Tang, T. (2012). miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC Bioinforma.* 13, 140. doi:10.1186/1471-2105-13-140

Xing, K., Zhao, X., Ao, H., Chen, S., Yang, T., Tan, Z., et al. (2019). Transcriptome analysis of miRNA and mRNA in the livers of pigs with highly diverged backfat thickness. *Sci. Rep.* 9 (1), 16740. doi:10.1038/s41598-019-53377-x

Xue, R., Zhu, X., Jia, L., Wu, J., Yang, J., Zhu, Y., et al. (2019). Mitofusin2, a rising star in acute-on-chronic liver failure, triggers macroautophagy via the mTOR signalling pathway. *J. Cell. Mol. Med.* 23 (11), 7810–7818. doi:10.1111/jcmm.14658

Yi, G., Yuan, J., Bi, H., Yan, W., Yang, N., and Qu, L. (2015). In-depth duodenal transcriptome survey in chickens with divergent feed efficiency using RNA-seq. *PloS one* 10 (9), e0136765. doi:10.1371/journal.pone.0136765

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11 (2), R14. doi:10.1186/gb-2010-11-2-r14

Zhang, J., Ying, Z. Z., Tang, Z. L., Long, L. Q., and Li, K. (2012). MicroRNA-148a promotes myogenic differentiation by targeting the ROCK1 gene. *J. Biol. Chem.* 287 (25), 21093–21101. doi:10.1074/jbc.M111.330381

Zhang, J., Shi, H., Wang, Y., Li, S., Cao, Z., Ji, S., et al. (2017a). Effect of dietary forage to concentrate ratios on dynamic profile changes and interactions of ruminal microbiota and metabolites in holstein heifers. *Front. Microbiol.* 8, 2206. doi:10.3389/fmicb.2017.02206

Zhang, X., Wang, W., Mo, F., La, Y., Li, C., and Li, F. (2017b). Association of residual feed intake with growth and slaughtering performance, blood metabolism, and body composition in growing lambs. *Sci. Rep.* 7 (1), 12681. doi:10.1038/s41598-017-13042-7

Zhang, D., Zhang, X., Li, F., Li, C., La, Y., Mo, F., et al. (2019). Transcriptome analysis identifies candidate genes and pathways associated with feed efficiency in Hu sheep. *Front. Genet.* 10, 1183. doi:10.3389/fgene.2019.01183

Zhang, D., Zhang, X., Li, F., Li, X., Zhao, Y., Zhang, Y., et al. (2022). Identification and characterization of circular RNAs in association with the feed efficiency in Hu lambs. *BMC Genomics* 23 (1), 288. doi:10.1186/s12864-022-08517-5

Zhou, L., Chen, J., Li, Z., Li, X., Hu, X., Huang, Y., et al. (2010). Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. *PloS one* 5 (12), e15224. doi:10.1371/journal.pone.0015224

Zhou, J., Li, Z., Huang, Y., Ju, W., Wang, D., Zhu, X., et al. (2019). MicroRNA-26a targets the mdm2/p53 loop directly in response to liver regeneration. *Int. J. Mol. Med.* 44 (4), 1505–1514. doi:10.3892/ijmm.2019.4282

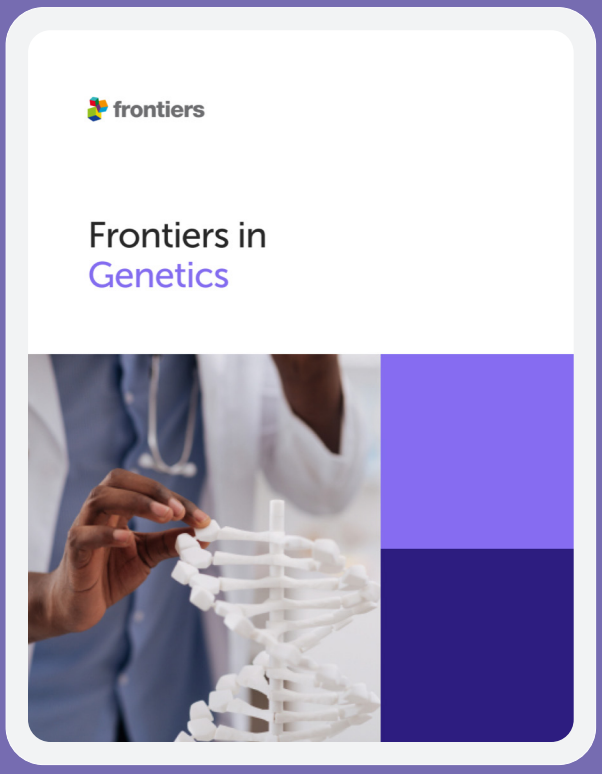
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact