frontiers | Research Topics

From agriculture genome to phenome: Genome-wide association, prediction and selection

Edited by
Kefei Chen and Li Ma

Published in
Frontiers in Genetics
Frontiers in Veterinary Science

ISSN 1664-8714
ISBN 978-2-8325-4061-9
DOI 10.3389/978-2-8325-4061-9

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

From agriculture genome to phenome: Genome-wide association, prediction and selection

Topic editors
Kefei Chen — Curtin University, Australia
Li Ma — University of Maryland, College Park, United States

Citation
Chen, K., Ma, L., eds. (2023). *From agriculture genome to phenome: Genome-wide association, prediction and selection*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4061-9

Table of contents

ORIGINAL RESEARCH
published: 13 June 2022
doi: 10.3389/fvets.2022.911166

Deep Small RNA Sequencing Reveals Important miRNAs Related to Muscle Development and Intramuscular Fat Deposition in *Longissimus dorsi* Muscle From Different Goat Breeds

Jiyuan Shen, Zhiyun Hao, Yuzhu Luo, Huimin Zhen, Yan Liu, Jiqing Wang*, Jiang Hu, Xiu Liu, Shaobin Li, Zhidong Zhao, Yuan Liu, Shutong Yang and Longbin Wang

Gansu Key Laboratory of Herbivorous Animal Biotechnology, Faculty of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China

OPEN ACCESS

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

Reviewed by:
Zeying Wang,
Shenyang Agricultural University, China
Rui Su,
Inner Mongolia Agricultural University, China

****Correspondence:***
Jiqing Wang
wangjq@gsau.edu.cn

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

Received: *02 April 2022*
Accepted: *09 May 2022*
Published: *13 June 2022*

Citation:
Shen J, Hao Z, Luo Y, Zhen H, Liu Y, Wang J, Hu J, Liu X, Li S, Zhao Z, Liu Y, Yang S and Wang L (2022) Deep Small RNA Sequencing Reveals Important miRNAs Related to Muscle Development and Intramuscular Fat Deposition in Longissimus dorsi Muscle From Different Goat Breeds. Front. Vet. Sci. 9:911166. doi: 10.3389/fvets.2022.911166

MicroRNAs (miRNAs) are a class of small non-coding RNAs that have been shown to play important post-transcriptional regulatory roles in the growth and development of skeletal muscle tissues. However, limited research into the effect of miRNAs on muscle development in goats has been reported. In this study, Liaoning cashmere (LC) goats and Ziwuling black (ZB) goats with significant phenotype difference in meat production performance were selected and the difference in *Longissimus dorsi* muscle tissue expression profile of miRNAs between the two goat breeds was then compared using small RNA sequencing. A total of 1,623 miRNAs were identified in *Longissimus dorsi* muscle tissues of the two goat breeds, including 410 known caprine miRNAs, 928 known species-conserved miRNAs and 285 novel miRNAs. Of these, 1,142 were co-expressed in both breeds, while 230 and 251 miRNAs were only expressed in LC and ZB goats, respectively. Compared with ZB goats, 24 up-regulated miRNAs and 135 miRNAs down-regulated were screened in LC goats. A miRNA-mRNA interaction network showed that the differentially expressed miRNAs would target important functional genes associated with muscle development and intramuscular fat deposition. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis revealed that the target genes of differentially expressed miRNAs were significantly enriched in Ras, Rap 1, FoxO, and Hippo signaling pathways. This study suggested that these differentially expressed miRNAs may be responsible for the phenotype differences in meat production performance between the two goat breeds, thereby providing an improved understanding of the roles of miRNAs in muscle tissue of goats.

Keywords: microRNA (miRNA), muscle development, intramuscular fat, small RNA sequencing, goat

INTRODUCTION

MicroRNAs (miRNAs) are a class of non-coding small RNA molecules (∼22 nucleotides), which are evolutionarily conserved in eukaryotes (1). In recent years, miRNAs are well-recognized as negative regulators of gene expression at post-transcriptional level, in that they can either inhibit translation or promote degradation of mRNA by complementary binding to the 3′-untranslated regions

(3′-UTR) of the target genes. The miRNAs are therefore involved in a wide variety of cell biological processes, including proliferation, differentiation, death, and fate specification (1, 2).

In modern animal husbandry, skeletal muscle is considered as the most economically important tissue of producing-meat livestock. Many studies have confirmed that miRNAs played essential roles in the growth and development of skeletal muscle. For example, miR-1 and miR-206 have been reported to facilitate differentiation and inhibit proliferation of bovine skeletal muscle satellite cells by targeting *PAX7* (3). The over-expression of miR-486 induced skeletal muscle hypertrophy of mice *via* activation of protein kinase B (Akt) (4). The miR-27b regulated caprine myogenic proliferation and differentiation of skeletal muscle satellite cells by inhibiting the expression of *PAX3* (5).

Up to now, research into expression profiles of miRNAs in the skeletal muscle tissue of domestic animals have mainly been focused on pigs (6–9), cattle (10–13), and sheep (14–17). It was found from these studies described above that miRNAs were differentially expressed in skeletal muscle at different developmental periods, or between different breeds. These further demonstrated crucial effect of miRNAs on skeletal muscle development.

In goats, the studies of miRNA expression profiles in skeletal muscle tissue have mainly been focused on different development stages. For example, Wang et al. identified 336 differentially expressed miRNAs in skeletal muscle of Huanghuai goats between fetal stage and 6-month-old stage, of which miR-424-5p, miR-29a, miR-129-3p, miR-181b, and miR-181d were involved in multiple important pathways related to muscle development (18). Guo et al. and Ling et al. also found some important differentially expressed miRNAs in skeletal muscle from prenatal stages to neonatal stage in Jianzhou Da′er goats and Anhui white goats (19, 20). However, little is known about the miRNA profiles of muscle tissues in other goat breeds, or between different goat breeds.

Ziwuling black (ZB) goats and Liaoning cashmere (LC) goats are both indigenous goat breeds in China. There are significant differences in meat production performance and muscle nutrients between the two breeds. For example, LC goats had higher carcass weight, muscle mass and intramuscular fat content, but lower muscle fiber density as well as contents of linoleic (C18: 2n-6), 11C, 14C-eicosadienoic acid (C20: 2n-6), moisture, and crude ash in meat when compared to ZB goats ($P < 0.05$) (21). In this context, elucidating the molecular mechanisms regulating these phenotypic differences between LC and ZB goats can provide insight for improving meat production performance of goats and other livestock. In this study, the expression profiles of miRNAs were compared in the *Longissimus dorsi* muscle between LC and ZB goats using small RNA sequencing. The differentially expressed miRNAs were screened between the two caprine breeds and the roles of miRNAs were also uncovered in skeletal muscle development and intramuscular fat deposition in goats.

MATERIALS AND METHODS

Ethics Statement

All animal procedures in this study were approved by Animal Experiment Ethics Committee of Gansu Agricultural University with an approval number of GSAU-ETH-AST-2021-028.

Longissimus dorsi Muscle Sample Collection and RNA Extraction

Ten healthy, 9-month-old male goats (five LC goats and five ZB goats) were selected from Yongfeng Goat Breeding Company in Huan County, Gansu Province, China. All goats were raised under the same environmental conditions and nutrition levels. After being slaughtered, the *Longissimus dorsi* muscle samples from the area between 12th and 13th ribs on the left carcass of each goat were collected and then frozen in liquid nitrogen immediately until further use. The meat production performance, muscle fiber size and intramuscular fat content from these LC and ZB goats have been reported by Wang et al. (21) and were also presented in **Supplementary Material 1**.

Total RNA from *Longissimus dorsi* muscle samples was extracted using a Trizol reagent kit (Invitrogen, Carlsbad, CA, USA). The concentration and purity of the RNA extracted were assessed using a Nanodrop 2000 (Thermo Scientific, MA, USA). Only samples with an RNA concentration >80 ng/uL and a purity of 1.80–2.10 were used for the study. The Agilent 2100 Bioanalyzer (Agilent, CA, USA) was used to assess RNA Integrity Number (RIN) of samples. Only RNA samples with RIN value ≥ 7 were used for small RNA enrichment.

Small RNA Library Construction and Sequencing

Ten small RNA libraries were generated using a TruSeq^{TM} Small RNA Sample Prep Kits (Illumina, San Diego, CA, USA) and then sequenced using an Illumina HiSeq^{TM}4000 sequencer (Illumina, San Diego, CA, United States) at the Gene Denovo Biotechnology Co., Ltd (Guangzhou, China). The clean reads were obtained by removing the reads containing adapters, low quality reads with quality scores < Q20 (the proportion of read bases whose error rate is <1%) or with unknown nucleotides, and the reads shorter than 18nt in length in the raw reads, using fastp v0.18.0. First, the clean reads were mapped to GenBank database v209.0 and Rfam database v11.0 to annotate and remove other non-coding RNAs, including ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), and small cytoplasmic RNA (scRNA). Secondly, the clean reads were mapped to the Caprine Genome Assembly ARS1 (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1) to remove their exons, introns and repeated sequences. Subsequently, the remaining clean reads were searched against miRbase v22.0 to annotate known caprine miRNAs and known miRNAs from other species (named known species-conserved miRNA). Finally, for the reads that were not annotated to miRBase V22.0, but matched the Caprine Genome Assembly ARS1, they were used to predict novel miRNAs using the miReap v.0.2. To ensure the uniquely annotated results for the reads, the following annotation ranking was used: rRNA > caprine miRNA > caprine miRNA edit > species-conserved miRNA > repeat sequence > exon sequence > novel miRNA > intron sequence.

Differentially Expressed miRNAs Analysis and Small RNA Sequencing Results Validation

The expression level of miRNAs was first normalized using transcripts per million (TPM). The TPM value is calculated by actual reads of each miRNA*10^6 by total reads of all miRNAs. The DESeq v2.0 (22) was used to screen differentially expressed miRNAs in *Longissimus dorsi* muscle tissues between LC and ZB goats, using the thresholds of a |fold change| > 2.0 and P-value < 0.05. To validate the accuracy of small RNA sequencing results, 23 differentially expressed miRNAs were selected for reverse transcription-quantitative PCR (RT-qPCR) analysis, including eight up-regulated miRNAs (miR-628-5p, miR-885-3p, novel-m0312-3p, miR-1994-3p, miR-67-3p, miR-278-3p, miR-307-3p, and novel-m0298-5p) and 15 down-regulated miRNAs (miR-381, miR-127-3p, miR-200c, miR-136-3p, miR-487b-3p, miR-200a, miR-410-3p, miR-136-5p, miR-127-5p, miR-141, miR-200b, miR-276-3p, novel-m0213-5p, miR-2796-3p, and miR-429) in *Longissimus dorsi* muscle of LC goats compared to ZB goats. The same RNA samples as those used for the small RNA sequencing were used to generate cDNA using a miRNA 1st Strand cDNA Synthesis Kit (Accurate Biology, Hunan, China). The caprine *U6* and *18SrRNA* were used as internal references to normalize the relative expression level of miRNAs (18, 19). The RT-qPCR was performed in triplicate using 2 × ChamQ SYBR qPCR Master system (Vazyme, Nanjing, China) on an Applied Biosystems QuantStudio 6 Flex (Thermo Lifetech, MA, United States) platform. A 20 μL reaction system was used for the RT-qPCR analysis including 2.0 μL of the cDNA, 0.4 μL of each primer, 10 μL of SYBR qPCR master mix (Vazyme, Nanjing, China) and 7.2 μL of RNase-free water. The thermal profile included an initial denaturation of 30 s at 95°C, followed by 45 cycles of 95°C for 10 s, 60°C for 34 s, and 95°C for 15 s, and finished by 60°C for 60 s. The $2^{-\Delta\Delta Ct}$ method was used to calculate the relative expression level of the miRNAs. The primer information used for RT-qPCR was presented in **Supplementary Material 2**.

Prediction, Validation, and Pathway Enrichment Analysis of the Target Genes of Differentially Expressed miRNAs

To investigate the potential roles of the differentially expressed miRNAs, miReap v0.2 (23), Miranda v3.3a (24), and TargetScan v7.0 (25) were used to predict their target genes and the predicted results from the three kinds of software were overlapped. To further verify the target relationship between the miRNAs and predicted target genes, a RT-qPCR analysis was performed to detect their relative expression levels in *Longissimus dorsi* muscle tissues of LC and ZB goats. The caprine *GAPDH* was used as an internal reference (5), and the primer information was listed in **Supplementary Material 2**. The RNA samples that were the same as those used for the small RNA sequencing analysis, were used to synthesize cDNA using SuperScript II reverse transcriptase (Invitrogen, Carlsbad, CA, United States). The same conditions and thermal profiles described above were used to perform the RT-qPCR analysis. The Pearson's coefficients in expression levels between the miRNAs and the target genes were calculated using SPSS v24.0. For negatively correlative pairs of miRNA-mRNA, the Cytoscape v3.5.1 (26) was used to construct an interaction network. The enrichment analysis of the signaling pathway of the target genes was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (27). The significant pathways ($P < 0.05$) were defined by hypergeometric test, and the P-values were corrected using the calculated False Discovery Rate (FDR) value.

RESULTS

Quality Control of Small RNA Sequencing Data

The concentration of 10 RNA samples collected from *Longissimus dorsi* muscle ranged from 124 to 272 ng/uL, while their purity ranged from 1.92 to 2.05 (**Supplementary Material 3**). On average, a total of 16,061,071 and 15,645,993 raw reads were generated from *Longissimus dorsi* muscle tissues of LC and ZB goats, respectively. The raw reads obtained in the study have been deposited in GenBank with accession numbers SRR16760528-SRR16760537. After removing low quality reads, adaptors and reads shorter than 18nt in length, an average of 15,337,813 and 14,815,683 clean reads were obtained in LC and ZB goats, respectively, of which 79.4 and 79.3% reads were mapped well to the caprine reference genome ARS1. Of these reads obtained, most of small RNA ranged from 18 to 24 nucleotides in length and the reads with 22 nucleotides were the most common, accounting for 45.7 and 45.4% of total reads in LC and ZB goats, respectively (**Figure 1**).

Identification of Known and Novel Caprine miRNAs

A total of 1,623 miRNAs were detected in *Longissimus dorsi* muscle tissues from both LC and ZB goats, including 410 known caprine miRNAs, 928 known species-conserved miRNAs and 285 novel miRNAs (**Supplementary Material 4**). Of these miRNAs, 1,142 were co-expressed in both breeds, while 230 and 251 miRNAs were only expressed in LC and ZB goats, respectively. Among all the small RNAs annotated in this study, the known miRNAs including mature caprine miRNA and species-conserved miRNA were the most abundant, which represented 84.0 and 83.6% of the total number of small RNA reads in LC and ZB goats, respectively (**Figure 2**).

Of the 1,623 miRNAs identified, miR-133a-3p was the most abundant with an TPM value of 125,235 and 136,882 in LC goats and ZB goats, respectively, followed by miR-26a-5p, miR-1, miR-99a-5p, and miR-27b-3p (**Supplementary Material 4**). Most notably, miR-133a-3p and miR-1 are members of myomiRs (namely miRNAs specific to muscle tissues) (28). In addition,

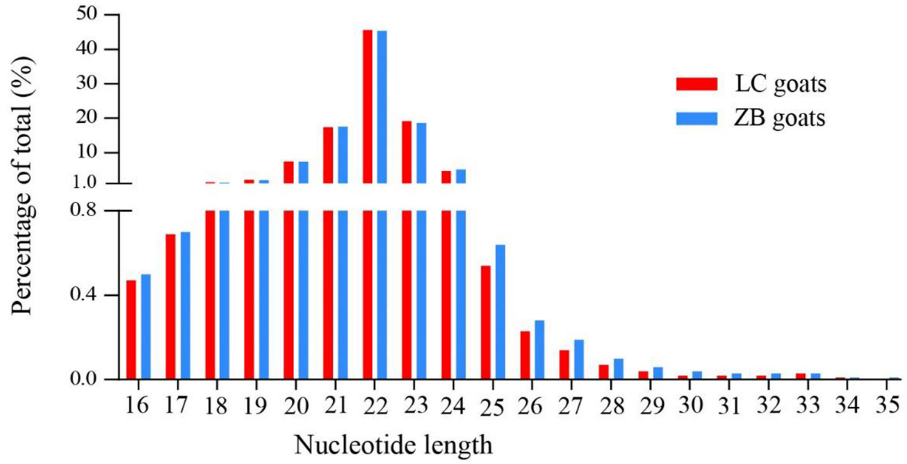


FIGURE 1 | The nucleotide length distribution of small RNA reads obtained from *Longissimus dorsi* muscle tissues of Liaoning cashmere (LC) and Ziwuling black (ZB) goats.

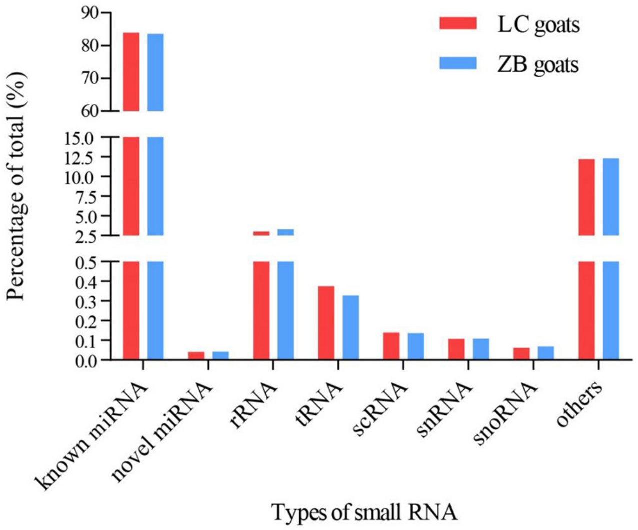


FIGURE 2 | The percentage of small RNA types in *Longissimus dorsi* muscle of Liaoning cashmere (LC) and Ziwuling black (ZB) goats. Known miRNAs included mature caprine miRNAs and species-conserved miRNAs. Others RNAs included the sequences aligned to exon and intron, repeated sequences, caprine miRNA edit, and sequences that were not aligned to any database.

myomiRs also included other highly expressed miRNAs, such as miR-206, miR-133b, and miR-208b (28, 29).

Screening and Validation of Differentially Expressed miRNAs

A total of 159 miRNAs were identified to be differentially expressed in *Longissimus dorsi* muscle tissues when comparing LC goats and ZB goats. Twenty-four miRNAs had higher expression in LC goats compared to ZB goats including one known caprine miRNA, 16 known species-conserved miRNAs and seven novel miRNAs (**Supplementary Material 5**). Among these up-regulated miRNAs in LC goats, miR-1994-3p was the most significant differentially expressed miRNA, followed by miR-67-3p, miR-278-3p, miR-307-3p, and miR-503-5p.

A total of 135 down-regulated miRNAs were identified in LC goats including 40 known caprine miRNAs, 86 known species-conserved miRNAs and nine novel miRNAs (**Supplementary Material 5**). Of these miRNAs, the most prominent down-regulated miRNA was miR-381, followed by miR-276-3p, miR-429, miR-2796-3p, and miR-136-3p.

The results from the RT-qPCR for 23 differentially expressed miRNAs were in consistency with those obtained from the small RNA sequencing analysis (**Figure 3**). Because miR-1994-3p and miR-429 were only expressed in LC and ZB goats, respectively, the $\log_2$ fold-change for LC goats relative to ZB goats was infinity for these two miRNAs. In this context, their relative expression levels are not presented in **Figure 3**. These results demonstrate the repeatability and reliability of small RNA sequencing results.

Predication and KEGG Analysis of the Target Genes of Differentially Expressed miRNAs

The results from interaction analysis of miReap v0.2, Miranda v3.3a, and TargetScan v7.0 revealed a total of 15,029 target genes identified for the 159 differentially expressed miRNAs. To clearly exhibit the interaction between the miRNAs and their target genes, 12 differentially expressed miRNAs were further selected, including the four most up-regulated miRNAs (miR-1994-3p, miR-67-3p, miR-278-3p, and miR-307-3p) and the five most down-regulated miRNAs (miR-381, miR-276-3p, miR-429, miR-2796-3p, and miR-136-3p) in LC goats, as well as two novel up-regulated miRNAs (novel-m0312-3p and novel-m0298-5p) and one novel down-regulated miRNA (novel-m0213-5p) in LC goats. It was notable that the novel-m0312-3p and novel-m0298-5p had the highest expression levels in LC goats among all novel up-regulated miRNAs, while the novel-m0213-5p had the highest expression levels in ZB goats among all novel down-regulated miRNAs. There were 5,407 target genes

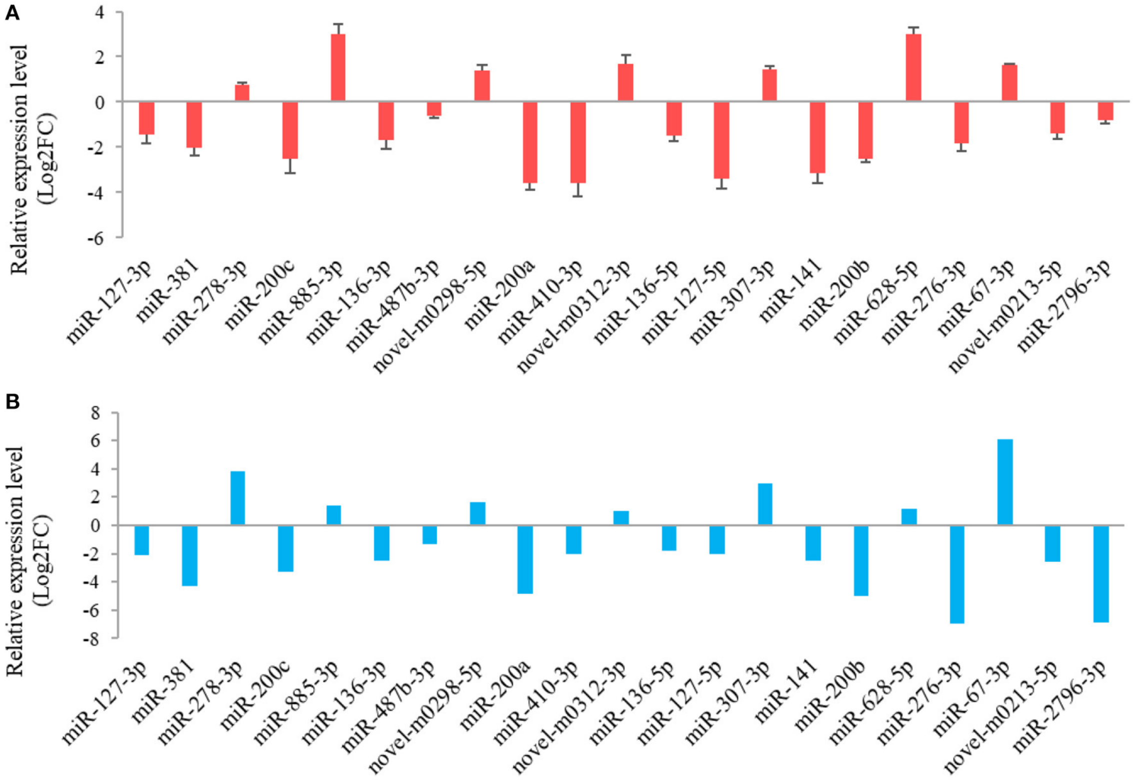


FIGURE 3 | RT-qPCR validation **(A)** of 21 differentially expressed miRNAs in *Longissimus dorsi* muscle tissues between Liaoning cashmere (LC) and Ziwuling black (ZB) goats identified using small RNA sequencing **(B)**. The error bars represent standard deviation value of the means for three independent replicates for each sample.

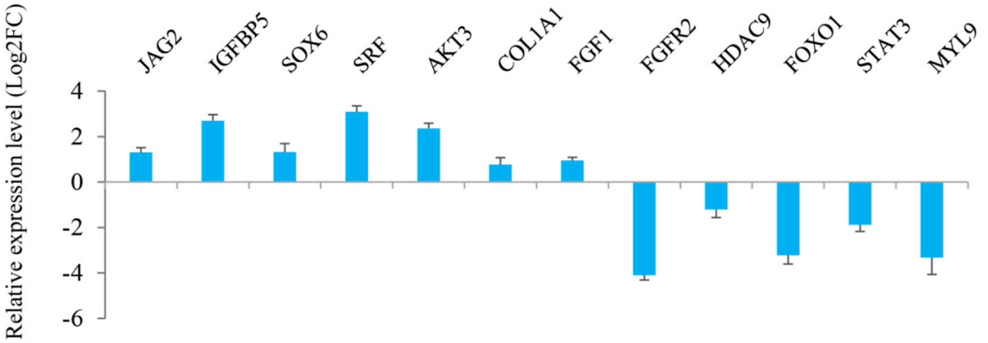


FIGURE 4 | Relative expression level of the target genes of 12 miRNAs selected in *Longissimus dorsi* muscle between Liaoning cashmere (LC) and Ziwuling black (ZB) goats detected using RT-qPCR.

in total for these 12 differentially expressed miRNAs, ranging from 87 target genes for miR-2796-3p to 1,295 targets for miR-429. For the 5,407 target genes, the genes related to muscle development and intramuscular fat deposition were further selected and their target relationships with corresponding miRNAs were verified using RT-qPCR (**Figures 3**, **4**). As shown in **Supplementary Material 6**, there were negative correlations in expression levels between the 12 miRNAs selected and their target genes. These suggest potential target relationships between these miRNAs and their target genes. Finally, a miRNA-mRNA interaction network was constructed (**Figure 5**). Some functional genes that have been previously described to be related with skeletal muscle development and intramuscular fat deposition were identified in this analysis. For example, the target genes *JAG2*, *IGFBP5*, *HDAC9*, and *FOXO1* were closely associated with myogenesis, while *SOX6* and *COL1A1* were reported to regulate adipogenesis (30–35) (**Figure 5**).

To further investigate the possible function of the target genes of differentially expressed miRNAs identified, a KEGG pathway analysis was performed (**Supplementary Material 7**, **Figure 6**).

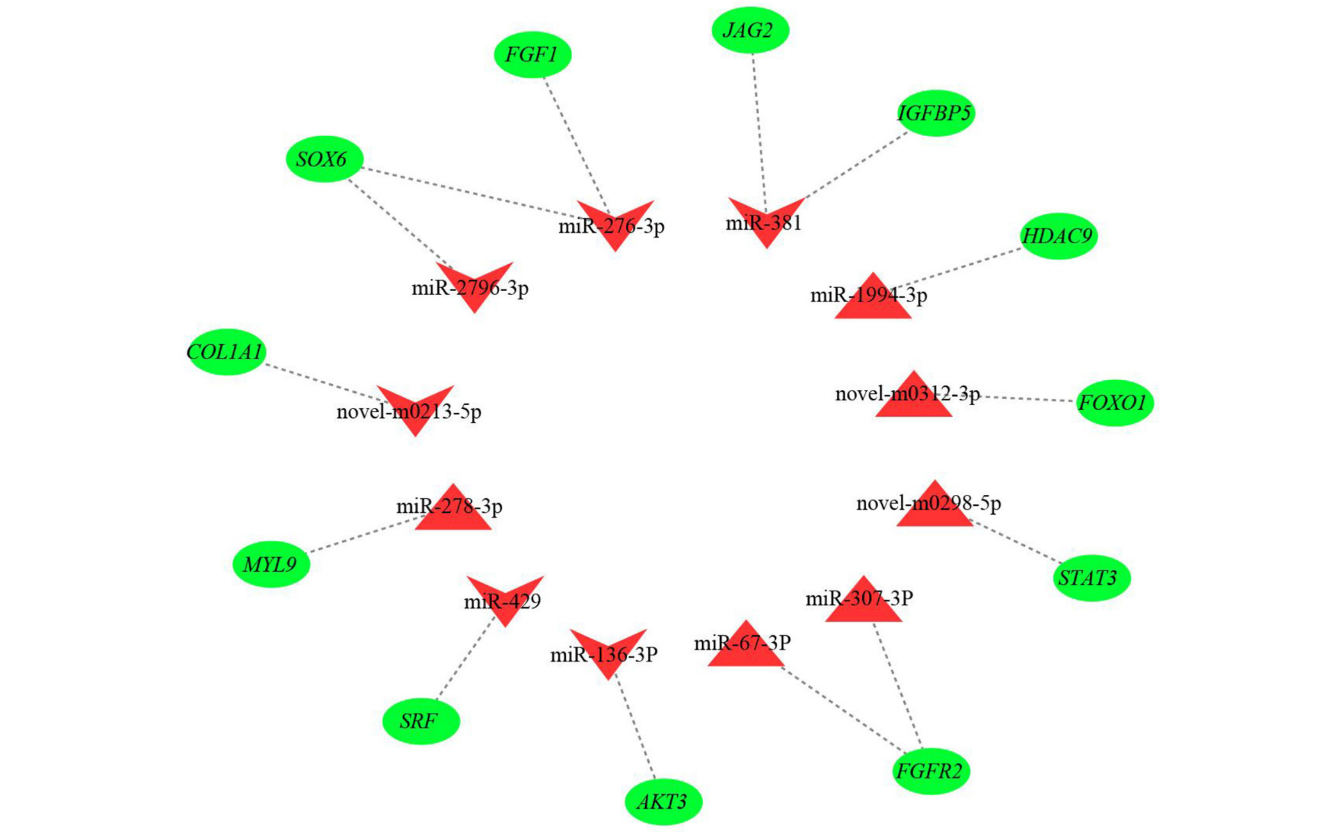


FIGURE 5 | The miRNA-mRNA interaction network of 12 differentially expressed miRNAs and their target genes. The red triangles and inverted triangles represent up-regulated and down-regulated miRNAs in *Longissimus dorsi* muscle of Liaoning cashmere (LC) goats compared to Ziwuling black (ZB) goats, respectively. The green circles represent the target genes of the miRNAs.

For up-regulated miRNAs in *Longissimus dorsi* muscle of LC goats, the most enriched pathway with the lowest *P*-value was axon guidance (*P*-value = 2.55E-09), followed by Ras signaling pathway (*P*-value = 1.82E-07) and Rap1 signaling pathway (*P*-value = 4.57E-07; **Figure 6A**, **Supplementary Material 7**). In addition, FoxO signaling pathway was also found among the top 10 significant pathways (*P*-value = 4.46E-06; **Figure 6A**). For down-regulated miRNAs in LC goats, the target genes were most enriched in pathways in cancer (*P*-value = 3.77E-17), followed by Hippo signaling pathway (*P*-value = 1.17E-16) and metabolic pathways (*P*-value = 2.46E-16; **Figure 6B**, **Supplementary Material 7**).

DISCUSSION

This study compared the *Longissimus dorsi* muscle tissue expression profiles of miRNAs of LC goats, with those of ZB goats that had lower carcass weight, muscle mass and intramuscular fat content. It was observed in the study that most miRNAs were 22 nucleotides in length identified in both caprine breeds and this was in accordance with the typical size range of mature miRNAs from Dicer-derived products (36). The results were also identical to the length distribution of small RNA reads in skeletal muscle tissues of other goat breeds (18, 20, 37), pigs (38), sheep (14, 15), and cattle (11). Additionally, our observation that the vast majority of small RNAs was known miRNAs in the two goat breeds has been also observed in previous studies of *Longissimus dorsi* muscle tissue (18, 20, 37). For example, Wang et al. found 81.5 and 82.5% of known miRNAs in muscle tissue of Huanghuai goats with fetal stage and 6-month-old stage, respectively (18). Similarly, 66.7 and 76.5% known miRNAs were found in skeletal muscle tissues of Anhui white goats (20) and Boer goats (37), respectively.

Of the top five highly expressed miRNAs found in both goat breeds, miR-133a-3p and miR-1 are members of myomiRs, which were specifically expressed in muscle tissues. The myomiRs have been reported to play key roles in regulating hypertrophy and regeneration of muscle fiber, as well as the differentiation and proliferation of muscle satellite cells (3, 39, 40). The miR-133a-3p was found to be the most highly expressed miRNA in *Longissimus dorsi* muscle of Boer goats (37). In Anhui white goats, Boer goats, Huanghuai goats and Jianzhou Da′er goats (18–20, 37), miR-1 was also one of the most abundant miRNAs in skeletal muscle tissues. The top five highly expressed miRNAs in the study also included miR-26a-5p, miR-99a-5p, and miR-27b-3p. The miR-99a-5p

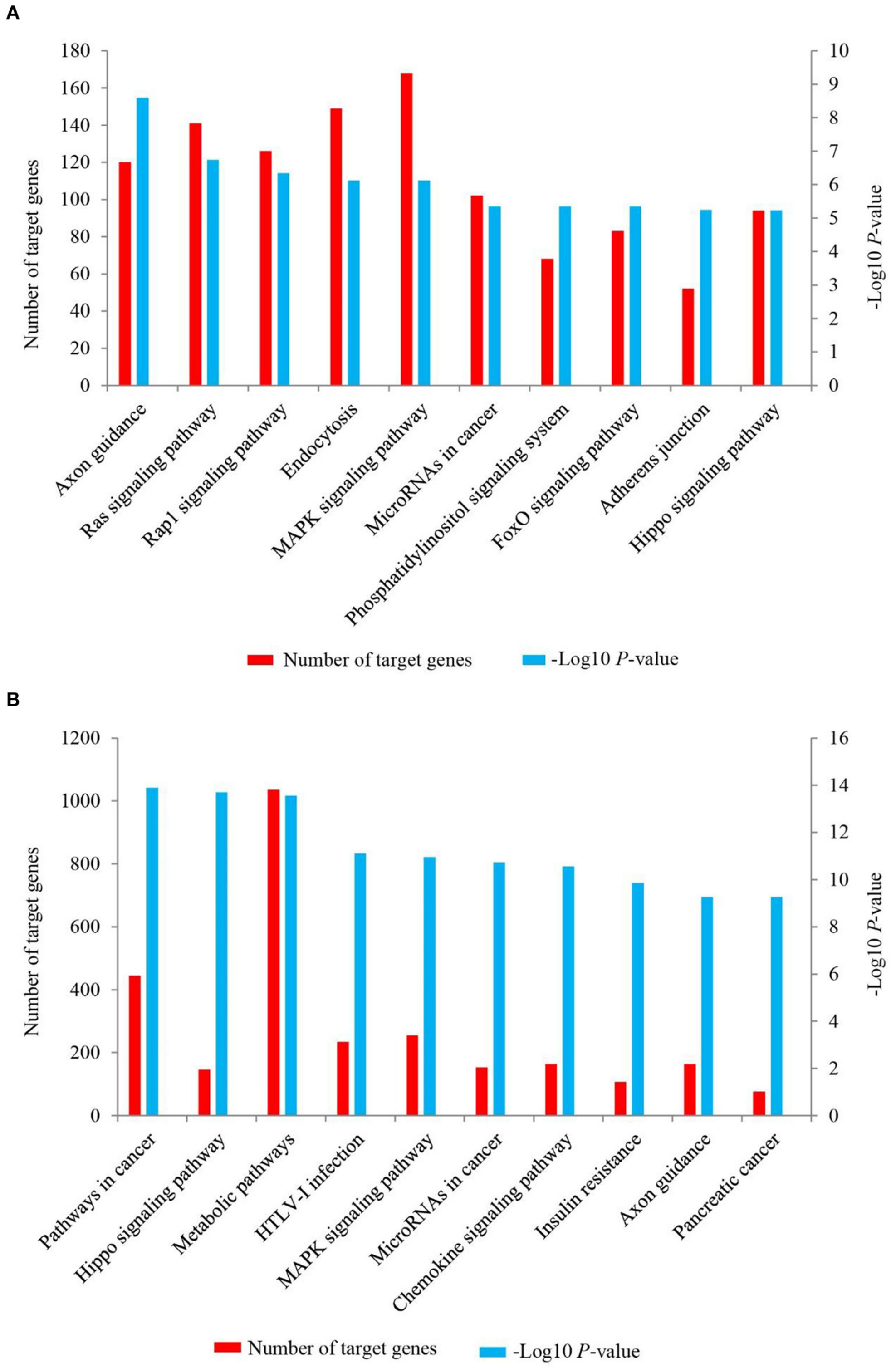


FIGURE 6 | The top 10 KEGG signaling pathways for the target genes of up-regulated **(A)** and down-regulated **(B)** miRNAs in the *Longissimus dorsi* muscle of Liaoning cashmere (LC) goats compared to Ziwuling black (ZB) goats. The left side Y-axis represents the number of the target genes of differentially expressed miRNAs involved in the pathway, while the Y axis on the right side shows the value of -Log10 (*P*-value).

and miR-27b-3p have been reported to regulate proliferation and differentiation of skeletal muscle satellite cells in chicken (41) and goats (5), respectively, while miR-26a-5p promoted myogenesis by targeting *Smad1* and *Smad4* of TGF-β/BMP pathway (42). In short, these highly expressed miRNAs may be necessary for the growth and development of caprine skeletal muscle.

In this study, the expression levels of miR-200 family were very low only with 0-101.3 TPM values and also down-regulated in muscle tissues of LC goats, including miR-429, miR-200b, miR-200a, miR-200c, and miR-141. The miR-200 family acts primarily as a negative regulator of muscle cells and adipocytes development. For example, miR-200b suppressed proliferation of C2C12 myoblast (43) and differentiation of ovine preadipocytes (44). The miR-200c and miR-429 have been reported to inhibit differentiation of C2C12 myoblast (45), porcine preadipocytes, and C2C12 myoblast (46, 47), respectively. These suggest that down-regulated expression of miR-200 family may be responsible for the higher carcass weight, muscle fiber size and intramuscular fat content in LC goats.

The miR-381 was the most down-regulated miRNA in LC goats in the study. The differential expression of miR-381 in muscle tissues between different breeds with divergent meat production performance has also been found in sheep (15) and pigs (38). This may reflect the breed-specific expression pattern of miR-381. Although the molecular mechanism of miR-381 regulating on the growth and development of skeletal muscle is unclear, a miRNA-mRNA network showed that miR-381 would target some important functional genes, such as jagged canonical Notch ligand 2 (*JAG2*), insulin like growth factor binding protein 5 (*IGFBP5*), phosphatase and tensin homolog (*PTEN*), etc. (**Figure 5**). *JAG2* and *IGFBP5* are important components of notch signaling and IGF signaling pathways, respectively. The two signaling pathways promoted muscle growth and development by regulating the activity of muscle satellite cells (30, 31). It was inferred that higher expression level of miR-381 in ZB goats may result in lower carcass weight by more inhibiting the expression level of the target genes *JAG2* and *IGFBP5*.

It was notable that known species-conserved miRNAs identified in this study may play key roles in the growth and development of caprine skeletal muscle, although their sequences have not been deposited in known caprine miRNA database. For example, up-regulated known species-conserved miR-885-3p in LC goats was reported to promote myoblasts proliferation in cattle (48). On the contrary, down-regulated species-conserved miR-370-3p in LC goats played a negative role in skeletal myogenesis in mice (49). The two species-conserved miRNAs may partly explain higher meat production performance in LC goats compared to ZB goats. As one of the most down-regulated miRNAs in LC goats, the known species-conserved miR-276-3p and miR-2796-3p would target SRY-box transcription factor 6 (*SOX6*; **Figure 5**), which promoted adipogenesis in human by activating adipogenic regulators including PPARγ, C/EBPα, and MEST (34). We therefore speculate that the down-regulation of the two species-conserved miRNAs in LC goats may promote adipogenesis by elevating expression of *SOX6*, resulting in increased deposition of intramuscular fat in LC goats. Similarly, the most up-regulated species-conserved miR-1994-3p may contribute to higher carcass weight and muscle mass of LC goats as it would target histone deacetylase 9 (*HDAC9*). The gene was found to play negative roles in muscle cell differentiation (32).

Four down-regulated known caprine miRNAs (miR-127-3p, miR-217-5p, miR-410-3p, and miR-487b-3p) in LC goats attracted our attention. The miR-127-3p has been reported to inhibit proliferation of C2C12 myoblast (50), as well as proliferation and differentiation of porcine skeletal muscle satellite cells (51). Similarly, miR-217-5p and miR-487b-3p inhibited the differentiation of skeletal muscle cells in rats (52) and mice (53), respectively. Additionally, the inhibition of miR-410-3p on the differentiation of adipocyte by targeting *IRS-1* has also been reported in humans (54). It could be therefore inferred that the lower expression levels of the four miRNAs in LC goats may be responsible for its higher meat production performance and intramuscular fat content compared to ZB goats.

It was noteworthy that some novel miRNAs identified in this study were also differentially expressed between the two goat breeds. The roles of these miRNAs in the growth and development of muscle and adipose tissues may be reflected by the function of their target genes. For example, as a positive regulator of muscle atrophy (33), Forkhead box O1 (*FOXO1*) would be targeted by up-regulated novel-m0312-3p in LC goats. The up-regulation of novel-m0312-3p would result in higher muscle mass in LC goats by inhibiting the expression of *FOXO1* and its effect on muscle atrophy. Meanwhile, down-regulated novel-m0213-5p found in LC goats would target SMAD family member 4 (*SMAD4*), which was contributed to the proliferation of porcine intramuscular preadipocyte (55). It was therefore inferred that down-regulation expression of novel-m0213-5p was responsible for higher intramuscular fat content in LC goats by less inhibition of *SMAD4* in expression compared to ZB goats.

As might be expected, the target genes of differentially expressed miRNAs identified in the study were involved in the growth and development of skeletal muscle or adipose tissues. As one of the most enriched pathways for the target genes of up-regulated miRNAs in LC goats, Ras signaling has been reported to negatively regulate skeletal muscle myogenesis (56). Rap 1 signaling pathway negatively regulated adipocyte differentiation and was also associated with myogenic differentiation (57, 58). FoxO signaling pathway accelerated skeletal muscle atrophy by inducing proteolytic and apoptotic (59, 60). These enriched pathways found in the study have also been described previously. For example, Ras signaling pathway was significantly enriched by the target genes of differentially expressed miRNAs identified in skeletal muscle of yak with different ages (61), and Rap1 signaling pathway was enriched by differentially expressed genes in muscle tissues of pigs (62) and goats (63) during different development stages. The three pathways described above may partly explain why LC goats have higher carcass weight and intramuscular fat content compared to ZB goats. Hippo signaling pathway was one of the most enriched pathways for the target genes of down-regulated miRNAs in LC goats with larger muscle fiber size and carcass weight. This is not surprising as the pathway is necessary for the increase of skeletal muscle mass (64, 65). Interestingly, MAPK signaling pathway was significantly enriched by both the target genes of up-regulated miRNAs and down-regulated miRNAs in LC goats. This suggests that the pathway may play dual roles in muscle development. The speculation was subsequently confirmed. MAPK signaling pathway has been mainly recognized as a positive regulator of myogenesis in animals (66). However, an inhibition effect on myogenesis in mice was also reported by Weston et al. (67).

CONCLUSION

This study compared the skeletal muscle tissue expression profiles of miRNAs between different goat breeds. The miR-381, miR-127-3p, miR-200b, miR-200c, miR-429, miR-217-5p, miR-885-3p, miR-370-3p, miR-1994-3p, miR-487b-3p, and novel-m0312-3p were found to be associated with muscle development, while miR-200b, miR-429, miR-276-3p, miR-2796-3p, miR-410-3p, and novel-m0213-5p were related to intramuscular fat deposition in goats.

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 GenBank [accession: SRR16760528-SRR16760537].

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Experiment Ethics Committee of Gansu Agricultural University (Ethic approval file No. GSAU-ETH-AST-2021-028). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

JS and JW did the data analysis and wrote the manuscript. ZH, YuzL, HZ, YaL, JH, and XL performed investigation and collected the samples. SL, ZZ, YuaL, SY, and LW performed the formal analysis, methodology, and software. JW did the project administration and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the fund for Basic Research Creative Groups of Gansu Province (18JR3RA190), the Fuxi Young Talents Fund of Gansu Agricultural University (Gaufx-02Y02), and the Projects of Gansu Agricultural University (GSAU-ZL-2015-033).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.911166/full#supplementary-material

REFERENCES

1. Carrington JC, Ambros V. Role of microRNAs in plant and animal development. *Science.* (2003) 301:336–8. doi: 10.1126/science.1085242
2. Ambros V. The functions of animal microRNAs. *Nature.* (2004) 431:350–5. doi: 10.1038/nature02871
3. Chen JF, Tao Y, Li J, Deng Z, Yan Z, Xiao X, et al. MicroRNA-1 and microRNA-206 regulate skeletal muscle satellite cell proliferation and differentiation by repressing *Pax7*. *J Cell Biol.* (2010) 190:867–79. doi: 10.1083/jcb.200911036
4. Hitachi K, Nakatani M, Tsuchida K. Myostatin signaling regulates Akt activity *via* the regulation of miR-486 expression. *Int J Biochem Cell B.* (2014) 47:93–103. doi: 10.1016/j.biocel.2013.12.003
5. Ling YH, Sui MH, Zheng Q, Wang KY, Wu H, Li WY, et al. MiR-27b regulates myogenic proliferation and differentiation by targeting *Pax3* in goat. *Sci Rep.* (2018) 8:3909. doi: 10.1038/s41598-018-22262-4
6. Lu J, Ye H, Hui W, Miao Y, Li X, Cao J, et al. Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs. *Sci Rep.* (2015) 5:11953. doi: 10.1038/srep11953
7. Sun J, Xie M, Huang Z, Li H, Chen T, Sun R, et al. Integrated analysis of non-coding RNA and mRNA expression profiles of 2 pig breeds differing in muscle traits. *J Anim Sci.* (2017) 95:1092. doi: 10.2527/jas2016.0867
8. Wang Q, Qi R, Wang J, Huang W, Wu Y, Huang X, et al. Differential expression profile of miRNAs in porcine muscle and adipose tissue during development. *Gene.* (2017) 618:49–56. doi: 10.3390/genes8100271
9. Wei W, Li B, Liu K, Jiang A, Dong C, Jia C, et al. Identification of key microRNAs affecting drip loss in porcine longissimus dorsi by RNA-Seq. *Gene.* (2018) 647:276–82. doi: 10.1016/j.gene.2018.01.005
10. Huang YZ, Sun JJ, Zhang LZ, Li CJ, Womack JE, Li ZJ, et al. Genome-wide DNA methylation profiles and their relationships with mRNA and the microRNA transcriptome in bovine muscle tissue (Bos taurine). *Sci Rep.* (2014) 4:6546. doi: 10.1038/srep06546
11. Sun J, Sonstegard TS, Li C, Huang Y, Li Z, Lan X, et al. Altered microRNA expression in bovine skeletal muscle with age. *Anim Genet.* (2015) 46:227–38. doi: 10.1111/age.12272
12. Li N, Zhang Y, Li HP, Han L, Yan XM, Li HB, et al. Differential expression of mRNA-miRNAs related to intramuscular fat content in the *longissimus dorsi* in Xinjiang brown cattle. *PLoS ONE.* (2018) 13:e0206757. doi: 10.1371/journal.pone.0206757
13. Kappeler B, Regitano L, Poleti M, Cesar A, Moreira G, Gasparin G, et al. MiRNAs differentially expressed in skeletal muscle of animals with divergent estimated breeding values for beef tenderness. *BMC Mol Biol.* (2019) 20:1. doi: 10.1186/s12867-018-0118-3
14. Liu Z, Li C, Li X, Yao Y, Ni W, Zhang X, et al. Expression profiles of microRNAs in skeletal muscle of sheep by deep sequencing. *Asian Austral J Anim.* (2019) 32:757–66. doi: 10.5713/ajas.18.0473
15. Sun L, Lu S, Bai M, Xiang L, Li J, Jia C, et al. Integrative microRNA-mRNA analysis of muscle tissues in Qianhua Mutton Merino and Small Tail Han sheep reveals key roles for oar-miR-655-3p and oar-miR-381-5p. *DNA Cell Biol.* (2019) 38:423–35. doi: 10.1089/dna.2018.4408
16. Kaur M, Kumar A, Siddaraju NK, Fairoze MN, Chhabra P, Ahlawat S, et al. Differential expression of miRNAs in skeletal muscles of Indian sheep with diverse carcass and muscle traits. *Sci Rep.* (2020) 10:16332. doi: 10.1038/s41598-020-73071-7

17. Yuan C, Zhang K, Yue Y, Guo T, Liu J, Niu C, et al. Analysis of dynamic and widespread lncRNA and miRNA expression in fetal sheep skeletal muscle. *Peer J.* (2020) 8:e9957. doi: 10.7717/peerj.9957
18. Wang Y, Zhang C, Fang X, Zhao Y, Chen X, Sun J, et al. Identification and profiling of microRNAs and their target genes from developing caprine skeletal muscle. *PLoS ONE.* (2014) 9:e96857. doi: 10.1371/journal.pone.0096857
19. Guo J, Wei Z, Zhan S, Li L, Zhong T, Wang L, et al. Identification and expression profiling of miRNAome in goat longissimus dorsi muscle from prenatal stages to a neonatal stage. *PLoS ONE.* (2016) 11:e0165764. doi: 10.1371/journal.pone.0165764
20. Ling Y, Zheng Q, Jing J, Sui M, Zhu L, Li Y, et al. RNA-Seq reveals miRNA role shifts in seven stages of skeletal muscles in goat fetuses and kids. *Front Genet.* (2020) 11:684. doi: 10.3389/fgene.2020.00684
21. Wang JQ, Shen JY, Liu X, Li SB, Luo YZ, Zhao ML, et al. Comparative analysis of meat production traits, meat quality, and muscle nutrient and fatty acid contents between Ziwuling black goats and Liaoning cashmere goats. *Acta Pratacult Sin.* (2021) 30:166–77. doi: 10.11686/cyxb2020199
22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* (2014) 15:550. doi: 10.1186/s13059-014-0550-8
23. Liu W, Xu L, Wang Y, Shen H, Zhu X, Zhang K, et al. Transcriptome-wide analysis of chromium-stress responsive microRNAs to explore miRNA-mediated regulatory networks in radish (*Raphanus sativus* L.). *Sci Rep.* (2015) 5:14024. doi: 10.1038/srep14024
24. Turner DA. Miranda: a non-strict functional language with polymorphic types. In: *Conference on Functional Programming Languages and Computer Architecture*. Berlin; Heidelberg: Springer (1985). p. 1–16. doi: 10.1007/3-540-15975-4_26
25. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell.* (2005) 120:15–20. doi: 10.1016/j.cell.2004.12.035
26. 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
27. Minoru K, Michihiro A, Susumu G, Masahiro H, Mika H, Masumi I, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* (2008) 36:480–4. doi: 10.1093/nar/gkm882
28. McCarthy JJ. The MyomiR network in skeletal muscle plasticity. *Exerc Sport Sci Rev.* (2011) 39:150–4. doi: 10.1097/JES.0b013e31821c01e1
29. Li D, Xia L, Chen M, Lin C, Wu H, Zhang Y et al. miR-133b, a particular member of myomiRs, coming into playing its unique pathological role in human cancer. *Oncotarget.* (2017) 8:50193–208. doi: 10.18632/oncotarget.16745
30. Shan T, Xu Z, Wu W, Liu J, Wang Y. Roles of Notch1 signaling in regulating satellite cell fates choices and postnatal skeletal myogenesis. *J Cell Physiol.* (2017) 232:2964–7. doi: 10.1002/jcp.25730
31. Zhang L, Wang XH, Wang H, Du J, Mitch WE. Satellite cell dysfunction and impaired IGF-1 signaling cause CKD-induced muscle atrophy. *J Am Soc Nephrol.* (2010) 21:419–27. doi: 10.1681/ASN.2009060571
32. Haberland M, Arnold MA, McAnally J, Phan D, Kim Y, Olson EN. Regulation of HDAC9 gene expression by MEF2 establishes a negative-feedback loop in the transcriptional circuitry of muscle differentiation. *Mol Cell Biol.* (2007) 27:518–25. doi: 10.1128/MCB.01415-06
33. Oyabu M, Takigawa K, Mizutani S, Hatazawa Y, Fujita M, Ohira Y, et al. FOXO1 cooperates with C/EBPδ and ATF4 to regulate skeletal muscle atrophy transcriptional program during fasting. *FASEB J.* (2022) 36:e22152. doi: 10.1096/fj.202101385RR
34. Leow SC, Poschmann J, Too PG, Yin J, Joseph R, McFarlane C, et al. The transcription factor SOX6 contributes to the developmental origins of obesity by promoting adipogenesis. *Development.* (2016) 143:950–61. doi: 10.1242/dev.131573
35. Côté JA, Guénard F, Lessard J, Lapointe M, Biron S, Vohl MC, et al. Temporal changes in gene expression profile during mature adipocyte dedifferentiation. *Int J Genomics.* (2017) 2017:5149362. doi: 10.1155/2017/5149362
36. Zhang B, Stellwag EJ, Pan X. Large-scale genome analysis reveals unique features of microRNAs. *Gene.* (2009) 443:100–9. doi: 10.1016/j.gene.2009.04.027
37. Ling YH, Ding JP, Zhang XD, Wang LJ, Zhang YH, Li YS, et al. Characterization of microRNAs from goat (*Capra hircus*) by Solexa deep-sequencing technology. *Genet Mol Res.* (2013) 12:1951–61. doi: 10.4238/2013.June.13.4
38. Hou X, Yang Y, Zhu S, Hua C, Zhou R, Mu Y, et al. Comparison of skeletal muscle miRNA and mRNA profiles among three pig breeds. *Mol Genet Genomics.* (2016) 291:559–73. doi: 10.1007/s00438-015-1126-3
39. Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, et al. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet.* (2006) 38:228–33. doi: 10.1038/ng1725
40. Nakasa T, Ishikawa M, Shi M, Shibuya H, Adachi N, Ochi M. Acceleration of muscle regeneration by local injection of muscle-specific microRNAs in rat skeletal muscle injury model. *J Cell Mol Med.* (2010) 14:2495–505. doi: 10.1111/j.1582-4934.2009.00898.x
41. Cao X, Tang S, Du F, Li H, Shen X, Li D, et al. MiR-99a-5p regulates the proliferation and differentiation of skeletal muscle satellite cells by targeting MTMR3 in chicken. *Genes.* (2020) 11:369. doi: 10.3390/genes11040369
42. Dey BK, Gagan J, Yan Z, Dutta A. MiR-26a is required for skeletal muscle differentiation and regeneration in mice. *Gene Dev.* (2012) 26:2180–91. doi: 10.1101/gad.198085.112
43. Yao CX, Wei QX, Zhang YY, Wang WP, Xue LX, Yang F, et al. MiR-200b targets GATA-4 during cell growth and differentiation. *RNA Biol.* (2013) 10:465–80. doi: 10.4161/rna.24370
44. Jin X, Wang J, Hu J, Liu X, Li S, Lu Y, et al. MicroRNA-200b regulates the proliferation and differentiation of ovine preadipocytes by targeting p27 and KLF9. *Animals.* (2021) 11:2417. doi: 10.3390/ani11082417
45. D'Agostino M, Torcinaro A, Madaro L, Marchetti L, Sileno S, Beji S, et al. Role of miR-200c in myogenic differentiation impairment *via* p66shc: implication in skeletal muscle regeneration of dystrophic mdx mice. *Oxid Med Cell Longev.* (2018) 2018:4814696. doi: 10.1155/2018/4814696
46. Peng Y, Chen FF, Ge J, Zhu JY, Shi XE, Li X, et al. MiR-429 inhibits differentiation and promotes proliferation in porcine preadipocytes. *Int J Mol Sci.* (2016) 17:2047. doi: 10.3390/ijms17122047
47. Nguyen MT, Min KH, Lee W. Palmitic acid-induced miR-429-3p impairs myoblast differentiation by downregulating CFL2. *Int J Mol Sci.* (2021) 22:10972. doi: 10.3390/ijms222010972
48. Elsaeid Elnour I, Dong D, Wang X, Zhansaya T, Khan R, Jian W, et al. Bta-miR-885 promotes proliferation and inhibits differentiation of myoblasts by targeting MyoD1. *J Cell Physiol.* (2020) 235:6625–36. doi: 10.1002/jcp.29559
49. Zhang P, Du J, Guo X, Wu S, He J, Li X, et al. LncMyoD promotes skeletal myogenesis and regulates skeletal muscle fiber-type composition by sponging miR-370-3p. *Genes.* (2021) 12:589. doi: 10.3390/genes12040589
50. Yuan R, Zhang X, Fang Y, Nie Y, Cai S, et al. MiR-127-3p inhibits the proliferation of myocytes by targeting KMT5a. *Biochem Bioph Res Co.* (2018) 503:970–6. doi: 10.1016/j.bbrc.2018.06.104
51. Chen X, Zhao C, Dou M, Sun Y, Yu T, et al. Deciphering the miRNA transcriptome of Rongchang pig longissimus dorsi at weaning and slaughter time points. *J Anim Physiol An N.* (2020) 104:954–64. doi: 10.1111/jpn.13314
52. Zhu M, Chen G, Yang Y, Yang J, Qin B, Gu L. MiR-217-5p regulates myogenesis in skeletal muscle stem cells by targeting FGFR2. *Mol Med Rep.* (2020) 22:850–8. doi: 10.3892/mmr.2020.11133
53. Wang J, Tan J, Qi Q, Yang L, Wang Y, Zhan C, et al. MiR-487b-3p suppresses the proliferation and differentiation of myoblasts by targeting *IRS1* in skeletal muscle myogenesis. *Int J Biol Sci.* (2018) 14:760–74. doi: 10.7150/ijbs.25052
54. Sun D, Ding Z, Shen L, Yang F, Han J, Wu G. MiR-410-3P inhibits adipocyte differentiation by targeting IRS-1 in cancer-associated cachexia patients. *Lipids Health Dis.* (2021) 20:115. doi: 10.1186/s12944-021-01530-9
55. Zhang Q, Cai R, Tang G, Zhang W, Pang W. MiR-146a-5p targeting SMAD4 and TRAF6 inhibits adipogenensis through TGF-β and AKT/mTORC1 signal pathways in porcine intramuscular preadipocytes. *J Anim Sci Biotechno.* (2021) 12:12. doi: 10.1186/s40104-020-00525-3
56. Mitin N, Kudla AJ, Konieczny SF, Taparowsky EJ. Differential effects of Ras signaling through NFkappaB on skeletal myogenesis. *Oncogene*. (2001) 20:1276–86. doi: 10.1038/sj.onc.1204223

57. Pizon V, Cifuentes-Diaz C, Mège R, Baldacci G, Rieger F. Expression and localization of RAP1 proteins during myogenic differentiation. *Eur J Cell Biol.* (1996) 69:224.
58. Yeung F, Ramírez CM, Mateos-Gomez PA, Pinzaru A, Ceccarini G, Kabir S, et al. Nontelomeric role for Rap1 in regulating metabolism and protecting against obesity. *Cell Rep.* (2013) 3:1847–56. doi: 10.1016/j.celrep.2013.05.032
59. Mcloughlin TJ, Smith SM, Delong AD, Wang H, Unterman TG, Esser KA. FoxO1 induces apoptosis in skeletal myotubes in a DNA-binding-dependent manner. *Am J Physiol Cell Physiol.* (2009) 297:C548–555. doi: 10.1152/ajpcell.00502.2008
60. Egerman MA, Glass DJ. Signaling pathways controlling skeletal muscle mass. *Crit Rev Biochem Mol Biol.* (2014) 49:59–68. doi: 10.3109/10409238.2013.857291
61. Ji H, Wang H, Ji Q, Ji W, Zhong J. Differential expression profile of microRNA in yak skeletal muscle and adipose tissue during development. *Genes Genom.* (2020) 42:1347–59. doi: 10.1007/s13258-020-00988-8
62. Wang Y, Wang J, Hu H, Wang H, Wang C, Lin H, et al. Dynamic transcriptome profiles of postnatal porcine skeletal muscle growth and development. *BMC Genom Data.* (2021) 22:32. doi: 10.1186/s12863-021-00984-1
63. Zhan S, Zhao W, Song T, Dong Y, Zhang H. Dynamic transcriptomic analysis in hircine longissimus dorsi muscle from fetal to neonatal development stages. *Funct Integr Genomic.* (2018) 18:43–54. doi: 10.1007/s10142-017-0573-9
64. Watt KI, Turner BJ, Hagg A, Zhang X, Davey JR, Qian H, et al. The Hippo pathway effector YAP is a critical regulator of skeletal muscle fiber size. *Nat Commun.* (2015) 6:6048. doi: 10.1038/ncomms7048
65. Olouyomi G, Marc F, Louise D. Hippo pathway and skeletal muscle mass regulation in mammals: a controversial relationship. *Front Physiol.* (2017) 8:190. doi: 10.3389/fphys.2017.00190
66. Keren A, Tamir Y, Bengal E. The p38 MAPK signaling pathway: a major regulator of skeletal muscle development. *Mol Cell Endocrinol.* (2006) 252:224–30. doi: 10.1016/j.mce.2006.03.017
67. Weston AD, Sampaio AV, Ridgeway AG, Underhill TM. Inhibition of p38 MAPK signaling promotes late stages of myogenesis. *J Cell Sci.* (2003) 116:2885–93. doi: 10.1242/jcs.00525

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

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

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fvets.2022.911685

Key miRNAs and Genes in the High-Altitude Adaptation of Tibetan Chickens

Binlong Chen[1], *Diyan Li*[2]*, *Bo Ran*[3], *Pu Zhang*[3] *and Tao Wang*[2]*

[1] College of Animal Science, Xichang University, Xichang, China, [2] School of Pharmacy, Chengdu University, Chengdu, China, [3] College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, China

Tibetan chickens living at high altitudes show specific physiological adaptations to the extreme environmental conditions. However, the regulated base of how chickens adapt to high-altitude habitats remains largely unknown. In this study, we sequenced 96 transcriptomes (including 48 miRNA and 48 mRNA transcriptomes of heart, liver, lung, and brain) and resequenced 12 whole genomes of Tibetan chickens and Peng'xian yellow chickens. We found that several miRNAs show the locally optimal plastic changes that occurred in miRNAs of chickens, such as miR-10c-5p, miR-144-3p, miR-3536, and miR-499-5p. These miRNAs could have effects on early adaption to the high-altitude environment of chickens. In addition, the genes under selection between Tibetan chickens and Peng'xian yellow chickens were mainly related to oxygen transport and oxidative stress. The I-kappa B kinase/NF-kappa B signaling pathway is widely found for high-altitude adaptation in Tibetan chickens. The candidate differentially expressed miRNAs and selected genes identified in this study may be useful in current breeding efforts to develop improved breeds for the highlands.

Keywords: chicken, miRNAs, genes, high-altitude, adaptation

OPEN ACCESS

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

Reviewed by:
Chao Tong,
University of Pennsylvania,
United States
Zhang Gen Xi,
Yangzhou University, China

****Correspondence:***
Diyan Li
lidiyan860714@163.com
Tao Wang
wanttao@hotmail.com

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

Received: *03 April 2022*
Accepted: *09 June 2022*
Published: *14 July 2022*

Citation:
Chen B, Li D, Ran B, Zhang P and
Wang T (2022) Key miRNAs and
Genes in the High-Altitude Adaptation
of Tibetan Chickens.
Front. Vet. Sci. 9:911685.
doi: 10.3389/fvets.2022.911685

INTRODUCTION

Because of many malconditions, such as more intensive ultraviolet radiation, lower atmospheric pressure, and lower oxygen partial pressure, the high-altitude plateau is unfavorable to animal survival; however, it is so astounding that many kinds of animals are living there. Some morphological studies indicate that high-altitude species with medium or large builds, such as yak, Tibetan mastiff, and Tibetan pig, have different tissue structures and blood biochemical indexes from their close relative species (1–3). These studies showed that plateau animals usually have more capillaries in the lung, thinner blood-gas barrier, and exceptional abundant red cells and hemoglobin. There can be no doubt that these animal species have adapted to high-altitude environments. Phenotypic plasticity is the capacity that a single genotype can produce different phenotypes in response to environmental change (4). The role of phenotypic plasticity in adaptive evolution is a controversial issue that whether plasticity constrains or facilitates adaptive evolution (5–9). Our previous research showed that phenotypic plasticity could help chickens readapt to their ancestral environments (10). Anyway, plasticity changes are widespread when organisms are faced with new or altered environments.

To investigate whether some genes of these animals are changed in the progression of plateau environmental adaptation, many genome studies have been conducted. The study on domestic yaks (*Bos grunniens*) and their closely related low-altitude cattle (*Bos taurus*) showed that *Adam17*, *Arg2*, and *Mmp3* are low, especially in strong positive selection in yaks. All three genes are related to hypoxia-inducible factor-1α (11). Another study indicated that some genes that are involved in DNA repair and the production of ATPase are low in positive selection in Tibetan antelope (*Pantholops hodgsonii*) (12). Some genomic studies on pigs (13), chickens (14), and hot-spring snakes (15) also demonstrate similar findings. The results of genomic studies reveal the inherent differences between high-altitude animals and their low-altitude close relative species. Only a small number of studies focus on the regulatory factors of genes, such as microRNA (miRNA). miRNA is one kind of small noncoding RNA with 18–25 nt in length. Many studies have shown that miRNAs widely participate in various biological processes, such as angiogenesis (16, 17), DNA damage repair (18, 19), and erythropoiesis (20, 21). Few studies reveal the miRNA expression profile of high-altitude native animals and their low-altitude close relatives (22, 23). It is still unclear whether miRNA plays extensive roles in high-altitude adaptation.

The Tibetan chicken is one of the chicken breeds found on the Tibetan plateau dating back to the seventh century A.C., which are widely distributed in farming areas of Tibet, including Shigatse, Lhasa, Lhoka, and Nyingchi (24). A previous study showed that incubation at high altitude of fertilized eggs laid by sea-level hens markedly restricted fetal growth compared with those laid by high-altitude hens; in contrast, incubation at sea level of fertilized eggs laid by high-altitude hens not only restored but enhanced fetal growth (25), which suggested that these hens have a high-altitude adaption. The basis of genetic adaptations to the extreme environmental conditions of the Tibetan plateau has recently been partly investigated in Tibetan chickens (26–28). At transcriptome levels, chorioallantoic membrane samples of Tibetan chickens and Chahua chickens were analyzed to explore hypoxic adaptation in Tibetan chickens (29). Comparative transcriptomic and proteomic analyses indicated that differentially expressed genes and proteins were mainly enriched in angiogenesis pathways that might be helpful for hypoxic adaptation in the embryos of Tibetan chickens (30). miR-15a was significantly increased in embryonic lung tissue (31) and GgmiRNA-454 is a time-dependent and tissue-differential expression miRNA of Tibetan chickens (32). Furthermore, several key proteins and pathways were identified and considered important candidates for high-altitude adaptation in Tibetan chickens (33). However, transcriptome analysis, especially of these regulatory RNAs such as miRNA across multiple tissues, environments, and breeds, was rarely conducted to explore the mechanism of high-altitude adaption of the Tibetan chicken populations.

In this study, we performed the reciprocal transplant test (highland and lowland chicken breeds raised in both highland and lowland environments) of Tibetan chickens and a lowland chicken breed (Peng'xian yellow chicken) to explore the miRNAs involved in the process of plateau environment adaptation of Tibetan chickens. In addition, whole-genome resequencing of Tibetan chickens and Peng'xian yellow chickens was also conducted. Based on miRNA and mRNA data from heart, liver, lung, and brain tissues integrated with whole-genome resequencing data, we explored how chickens adapt to high-altitude environments. Our work uncovers several miRNAs have a plastic change as the living environment (altitude) changes. These miRNAs could be involved in the response of hypoxia, inflammation, or other stress under a high-altitude environment and help chickens adapt, indicating that miRNAs could play key roles in the adaptation to a high-altitude environment.

MATERIALS AND METHODS

Chicken Breeds and Sample Collection

Tibetan chickens (the high-altitude chicken breed) were hatched in A'ba (altitude, 3,300 m) and Ya'an (altitude, 670 m), and the same operation was performed on Peng'xian yellow chickens (the low-altitude chicken breed). All the chickens were fed normally. When the experimental chickens were 120 days old, for each group, we collected blood samples and four tissue samples (heart, liver, lung, and brain) from three healthy males with similar body weights. The blood samples and tissue samples were stored at −20°C and −80°C until DNA and RNA extraction. These 12 individuals were genome resequenced (**Supplementary Table 1**) and 48 tissue samples from them were transcriptome sequenced.

RNA Isolation and Sequencing

The standard TRIzol method was used to isolate the total RNA from tissue samples (heart, liver, lung, and brain). The concentration and purity of RNA were determined using a Nanodrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA), and the integrity of RNA was confirmed *via* a 2% agarose gel. Using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA, USA), the RNA integrity number (RIN) value was obtained. Using the QiaQuick PCR Purification Kit, the mRNA library was constructed. From the total RNA, 1 μg of RNA was obtained. By using the Truseq^{TM} Small RNA Sample Preparation Kit, adapter ligation and reverse transcription-polymerase chain reaction (PCR) were performed to obtain the cDNA. Sequencing was performed based on the HiSeq platform (Illumina; San Diego, CA, USA).

DNA Extraction and Sequencing

Genomic DNA was extracted from blood samples using the traditional phenol-chloroform protocol. DNA purity and quality were assessed using a Nanodrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA), 2% gel electrophoresis, and an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA, USA). The sequencing library was constructed using the TruSeq DNA Sample Preparation Kit (Illumina Inc., San Diego, CA, USA) following the manufacturer's protocols. The whole-genome resequencing was performed based on the HiSeq platform (Illumina; San Diego, CA, USA).

Analysis of Chicken Transcriptome

To explore the function of miRNA in the process of plateau environment adaptation of lowland chickens, we performed miRNA sequencing of four tissues (heart, liver, lung, and brain) from Peng'xian yellow chickens, which were raised in Ya'an ("LC" hereinafter) and A'ba ("HLC" hereinafter), and Tibetan chickens,

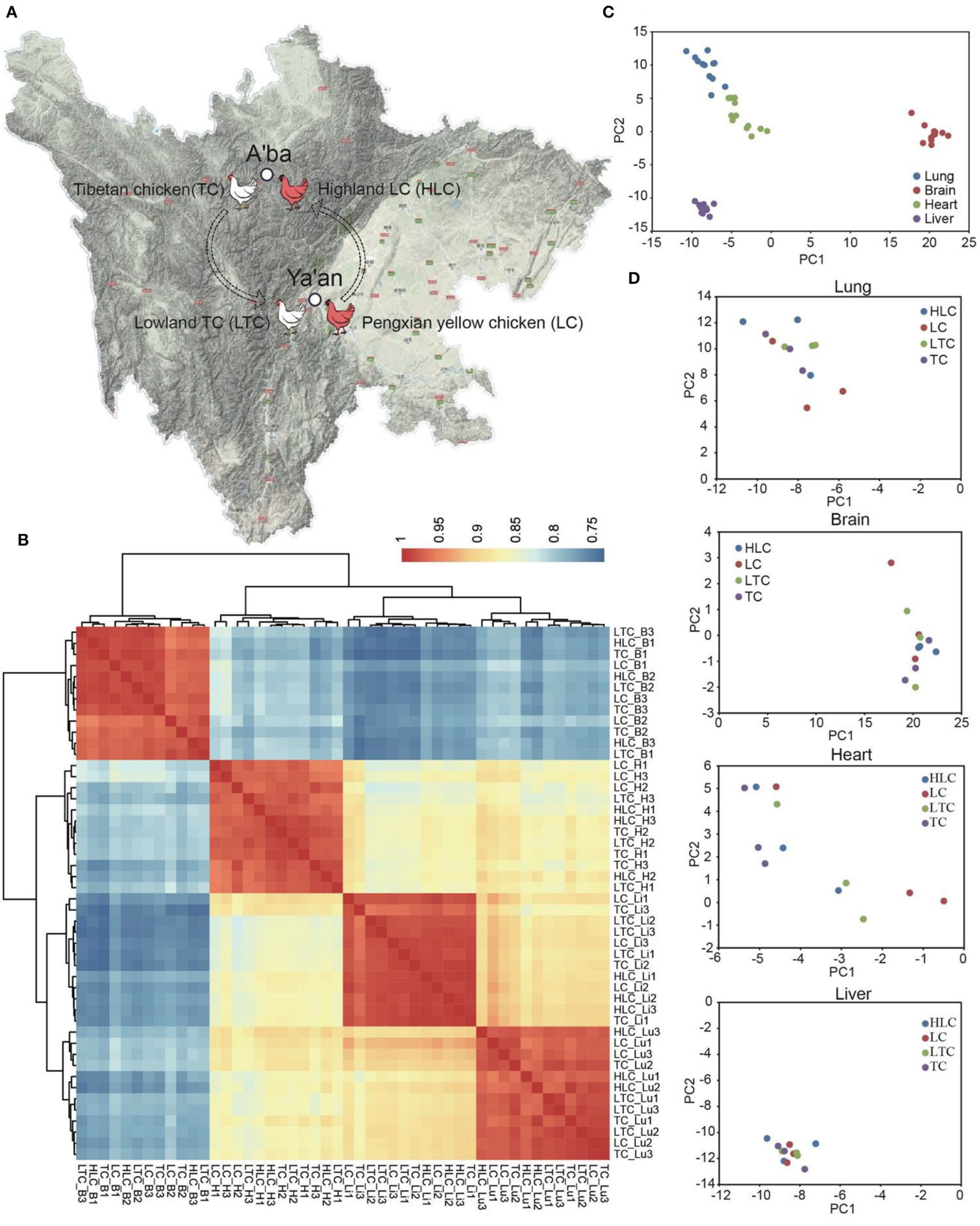


FIGURE 1 | The locations and animal groups of this study. **(A)** Sample collection sites. As the gray deepens, the altitude gets higher. The color of the chicken in this picture does not represent the real color of Tibetan chicken and Peng'xian yellow chicken. **(B)** Heatmap shows the correlation of all samples by Pearson correlation analysis based on CPM (B, H, Li, and Lu represent brain, heart, liver, and lung, respectively). **(C,D)** PCA plots show the difference in all samples and the samples of the same tissue.

which were raised in Ya'an ("LTC" hereinafter) and A'ba ("TC" hereinafter) (**Figure 1A**).

For miRNA data, Cutadapt version 1.12 (34) was used to remove the adaptor sequences. Trimmed reads were compared to known miRNAs of chicken from the miRbase database (version 22) by Bowtie (35). Next, after filtering known miRNA sequences, the remaining sequences were BLAST searched against the *Gallus gallus* genome. The sequences matching the chicken genome were used to predict the novel miRNA by mirDeep2 (36) using default parameters.

For transcriptome data, we employed TopHat2 (37) to do reads mapping against the chicken reference genome (Ensemble release 92). The miRNA expression level was normalized by edgeR (38). Then, we did correlation analysis (Pearson correlation) and principal component analysis (PCA) by SAS based on count-per-million (CPM), which was generated from edgeR. To identify differentially expressed miRNAs (DEMs), we first did a pairwise comparison by edgeR and obtained six compared groups, namely, "HLC-LC," "HLC-TC," "HLC-LTC," "LC-TC," "LC-LTC," and "TC-LTC," respectively. DEMs were identified with a log2 fold threshold of 1.5, log2 CPM > 2, and FDR ≤ 0.05.

We speculated that due to long-term living in highlands or lowlands, some miRNAs that are related to environmental adaptation have their unique expression patterns in Tibetan chickens and Peng'xian yellow chickens and help chickens adapt to their living environments. In short, that is what we called ENMs. To screen out ENMs, we first identified differentially expressed miRNAs (DEMs) between highland and lowland experimental chickens. In the samples of the same tissue, known miRNAs were compared in pairs to identify DEMs using edgeR. We focused on the miRNAs that are caused by the EN factor. So, we merged the DEMs that are in every comparative group of the same tissue to reduce the influence of other factors. The concrete implementation method is as follows.

In the comparative groups of the same tissue, the EN factor works in four groups ("TC-LC, "TC-LTC," "HLC-LC," and "HLC-LTC," respectively) that we named effective groups. Because of the same growth environment, the EN factor does not exist in the other groups ("TC-HLC" and "LTC-LC") that we named ineffective groups. In consideration of the condition that ENMs should be stable in effective groups, we focused on the miRNAs that are at the intersection of the effective groups and not in the ineffective groups.

We then screened miRNAs that changed with the environment to evaluate whether the optimal plasticity changes of miRNAs occurred in the chickens of the "HLC" group. In this study, the differentially expressed miRNAs can be divided into three parts, namely, caused by breed (BR), environment (EN), and experiment error (EE), respectively. For example,

DEMs TC-LC = FBR + FEN + FEE. In this group, there are different chicken breeds and different areas where these chickens grew up. So, the differentially expressed miRNAs were caused by all three factors in the "TC-LC" group.

DEMs HLC-TC = FBR + FEE. In this group, both Peng'xian yellow chickens and Tibetan chickens grew up in the same environment, A'ba. There is no environmental factor.

DEMs HTC-LTC = FEN + FEE. In this group, Tibetan chickens were raised in different areas, A'ba and Ya'an. BR factor was not working here.

To investigate the role of ENMs, we employed miRDB to predict their target genes. To ensure the reliability of the prediction results, we only keep the part the target score >80 in miRDB. Functional enrichment analysis of GO terms, KEGG pathway analysis, and pathway enrichment were performed using Metascape. Because there is no chicken database in Metascape, we map the genes to human homologous genes for further analysis.

Chicken Genome Resequencing Analysis

Trimmomatic (39) was used to remove adapter and low-quality data. Trimmed reads were compared to the chicken reference genome (GRCg6a) by Burrows-Wheeler Aligner (BWA). SAMtools (40) and GATK (41) were used to detect single nucleotide polymorphisms (SNPs), and the software Annovar (42) was used to annotate the SNPs with genomic elements. Subsequently, we did a selective sweep analysis. A 40-kb sliding window and a 20-kb step size were used to detect the selective sweep region of the genome. PopGenome (43) was used to calculate the *Fst* and π ratio of the SNPs in each window. Significantly, GO terms and KEGG pathways of the gene with *Fst* ≥ 0.3 were identified using Metascape (44) with $P \leq 0.01$.

RESULTS

Summary of the miRNA and mRNA Data

In this study, miRNA data were analyzed using Illumina deep sequencing technology. A total of 508.8 million clean reads were obtained. After annotation, a total of 1,079 known miRNAs were identified in all the samples. In total, 269.15 Gb of raw data from 48 mRNA libraries were obtained, with an average of 5.51Gb per sample. After quality filtering, 264.36 Gb of clean data with an average of 94.85% of Q20 was retained for further analysis.

To evaluate the correlation of samples in this study, we calculated the correlation coefficient between every two samples using the Pearson correlation method. The figure showed that the samples of the same tissue have higher correlations than the samples of different tissues for both miRNA and mRNA (**Figure 1B**). There is a very high correlation between every two samples of the same tissue, which is not subject to breed or growth environments. It seems like altitude has a small effect on the miRNA transcriptome of the chicken. Furthermore, we used the principal component analysis (PCA) method to reduce the dimensionality of these data and further assess the difference in samples. The result of PCA showed similar characteristics to correlation analysis. The difference between the different tissue samples is obvious (**Figure 1C**). The first three eigenvalues can explain the sample variability of 36.17% (PC1), 15.28% (PC2), and 11.26% (PC3), respectively. But the arrangement of the same tissue samples is chaotic (**Figure 1D**).

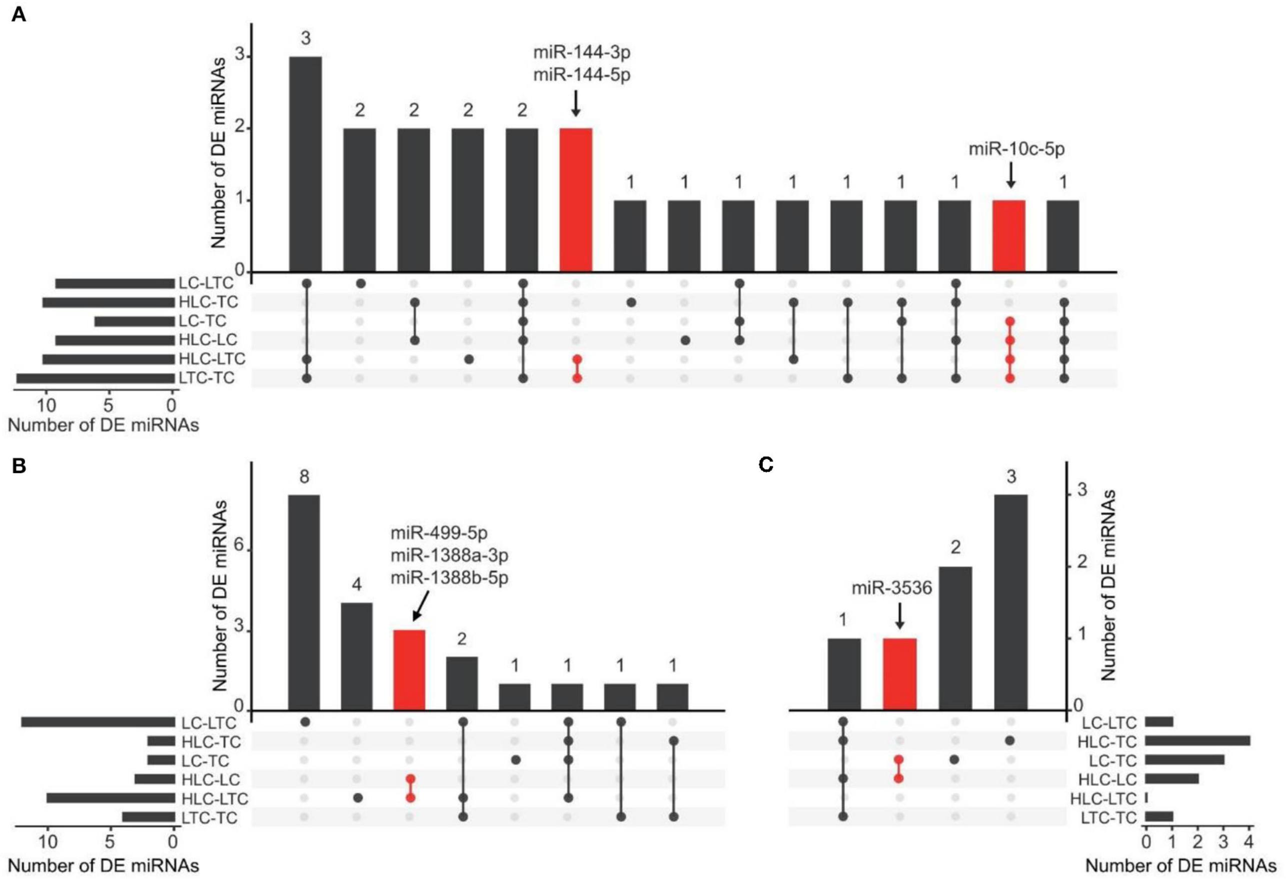


FIGURE 2 | UpSet plot showing the screening of candidate ENMs in the liver **(A)**, lung **(B)**, and heart **(C)**. The bars on the left or right of each panel were the numbers of DE miRNAs in each comparison of two groups of chickens. The dots represent relative DE miRNAs detected in the corresponding comparison. For example, in **(B)**, the red dots represent miRNAs detected to be differentially expressed in both two comparisons (HLC vs. LC, and HLC vs. LTC). Three and two candidate ENMs were obtained from the LC vs. HLC and LTC vs. TC comparisons, respectively.

miRNAs Showed the Locally Optimal Plasticity Changes Alongside Altitude Change

In this study, a total of 1,076 known miRNAs were detected. In addition, 354 novel miRNAs were detected (**Supplementary Table 2**). For these known miRNAs, 45 DEMs were detected. A total of 22, 21, 7, and 1 DEMs were detected in the liver, lung, heart, and brain, respectively.

In this study, the influence of the environmental (EN) factor is mainly reflected in the change in altitude. We focused on the miRNAs that are caused by the EN factor. The expression of miRNAs in this part is characterized by the change in higher elevation. It showed the plasticity of miRNAs. These miRNAs may play a key regulatory role in the process of plateau environmental adaptation of chickens. We merged these groups that include the EN factor to screen out the miRNAs that are caused by the EN factor. We focused on the miRNAs that always have a high expression level in the experimental chickens at a certain altitude, regardless of the chicken breeds. We named them EN miRNAs (ENMs).

In the liver, nine miRNAs have obvious expression changes in experimental chickens at high and low altitude environments. They are miR-1692, miR-10c-5p, miR-10b-5p, miR-144-5p, miR-144-3p, miR-2184-5p, miR-375, miR-1736-3p, and miR-205a, respectively. Six of these nine miRNAs (miR-1692, miR-10b-5p, miR-2184-5p, miR-375, miR-1736-3p, and miR-205a) appeared in the ineffective groups ("HLC-TC" and "LTC-LC"). Among the remaining three miRNAs, miR-10c-5p was always highly expressed in the liver of chickens that grew up in low-altitude areas ("LC" and "LTC"). Conversely, miR-144-5p and miR-144-3p were highly expressed in high-altitude experimental chickens ("TC" and "HLC"). These 3 miRNAs are found at the intersection of four effective groups and were marked (**Figure 2**) as ENMs.

In the same way, we also checked the intersection of the other tissues. At heart, we only found four DEMs in the effective groups. Two DEMs were found at the intersection of effective groups. In them, miR-375 did not show an obvious altitude preference. miR-3536 was highly expressed in high-altitude experimental chickens. In addition, miR-3536

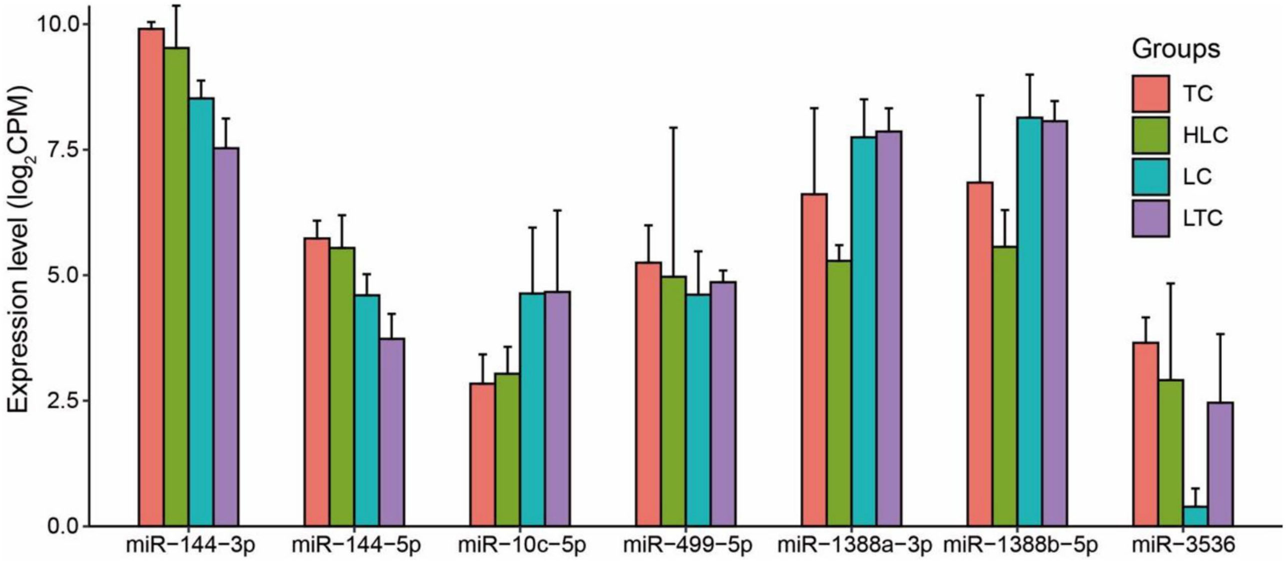


FIGURE 3 | The expression level of candidate miRNAs.

was not the DEMs of ineffective groups. In lung, six DEMs (miR-122-5p, miR-215-5p, miR-449d-5p, miR-449-5p, miR-1388b-5p, and miR-1388a-3p) were found at the intersection of effective groups. In them, miR-499-5p, miR-1388b-5p, and miR-1388a-3p were highly expressed in a certain living environment. And, these miRNAs did not appear in the DEMs of ineffective groups. We were surprised by the analysis result of the comparative groups of the brain. There was only one DEM (miR-194) in the effective groups and it was not at the intersection.

In brief, seven candidate ENMs were obtained from four tissues of experimental chickens. All these screened miRNAs showed plasticity changes in the environment. The expression pattern of these miRNAs was similar to native chickens' when Tibetan chickens and Peng'xian yellow chickens were brought into each other's living environment (**Figure 3**).

We further used a two-way analysis of variance to further analyze the influence of BR and EN factors on the expression levels of candidate miRNAs (**Supplementary Table 3**). The results showed that the expression level of all candidate miRNAs was not significantly affected by genetic factors, and the interaction effect of BR and EN factors was significantly affected by the expression of some miRNAs, such as miR-144-3p, miR-144-5p, miR-1388a-3p, and miR-3536. The expression levels of miR-10c-5p and miR-1388b-5p were only significantly affected by EN factors. Their expression levels in low-altitude chicken groups (LC and LTC) were significantly higher than those of high-altitude chicken groups (TC and HLC). miR-10c-5p and miR-1388b-5p showed the plasticity changes when altitude changed and were defined as ENMs. Then we explored the function of ENMs in the altitude adaption of chickens.

Function Prediction of Environment-Related miRNAs

We combined miRDB and Metascape to explore the functions of environment-related miRNAs (ENMs). miRDB was used to predict the target gene of ENMs. For miR-10c-5p and miR-1388b-5p, we found a total of 82 target genes, 67 and 15, respectively. We then performed enrichment analysis of these target genes using Metascape. The target genes of miR-10c-5p were significantly enriched in a total of 73 GO terms and 3 KEGG pathways ($P \leq$ The target genes of miR-10c-5p were significantly enriched in a total of 73 GO terms The target genes of miR-10c-5p were involved in the GO terms related to high-altitude stress, such as "regulation of I-kappaB kinase/NF-kappaB signaling" (GO:0043122), "ephrin receptor signaling pathway" (GO:0048013), and "stress-activated MAPK cascade" (GO:0051403) (**Figure 4**). In addition, the target genes of miR-1388b-5p were significantly enriched in 7 GO terms and 0 KEGG pathways, such as "protein serine/threonine kinase activity" (GO:0004674) and "maintenance of location" (GO:0051235) (**Supplementary Figure 1**).

Studies have shown that increased exposure to hypoxia could cause insufficient oxygen supply for tissue and further induce inflammation (45, 46). As mentioned above, we found that some target genes of miR-10c-5p are involved in the "regulation of I-kappaB kinase/NF-kappaB signaling" (GO:0043122) and "ephrin receptor signaling pathway" (GO:0048013). It showed that miR-10c-5p may help chickens resist the hypoxic environment.

The Target Genes of miR-10c-5p Showed Locally Optimal Plasticity Changes When the Environment Changed

To evaluate differentially expressed mRNAs (DEGs) between different tissues, breeds, and environments, we compared

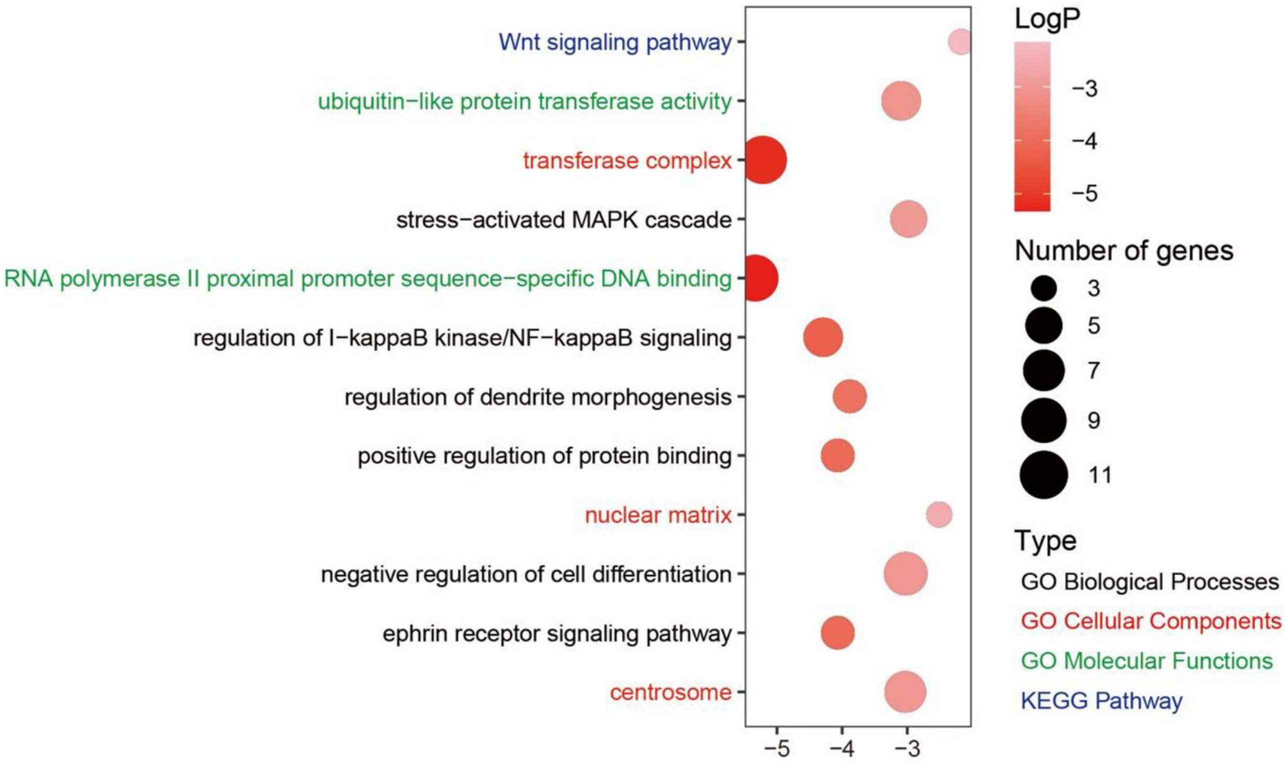


FIGURE 4 | Enrichment analysis of miR-10c-5p.

the expression levels with the threshold of FC (fold change) ≥ 2 (or ≤ 0.5) plus a Bonferroni-adjusted *P* adjusted a Boa t-test. Overall, we compared 28 pairs and identified 18–1,634 mRNAs as significantly differentially expressed (**Supplementary Table 4**). These DEGs were involved in 0–97 pathways (**Supplementary Table 4**). The genes highly expressed in Tibetan chickens were found to be enriched in immune-related GO terms (such as "innate immune response," "defense response to virus," and "cytokine activity") (**Supplementary Figure 2**).

miR-10c-5p shows a strong altitude preference. Its target genes may also show functional adaptations when the living environment is changed. Then we explored the expression pattern of the target genes of miR-10c-5p. In fact, through joint analysis of miRNA data and transcriptome data, we found that expression pattern change occurred in several genes that are involved in these GO terms between high and low-altitude experimental chickens (**Figure 5**), such as zinc finger MYND-type containing 11 (*ZMYND11*), OTU deubiquitinase 7A (*OTUD7A*), tripartite motif-containing 13 (*TRIM13*), and T-cell lymphoma invasion and metastasis 1 (*TIAM1*). *ZMYND11* has a lower expression level in high-altitude experimental chickens ("HLC" and "TC" groups) than in low-altitude chickens ("LTC" and "LC" groups). *OTUD7A*, *TRIM13*, and *TIAM1* have similar expression patterns. As shown in **Figure 5**, both the PCA plot and histogram showed that there is no significant difference between lowland chicken groups ("LC" and "LTC") in these genes. Analogously, the expression pattern of these genes in chickens of the "HLC" group was similar to those of the "TC" group. It indicated that some target genes of miR-10c-5p also showed optimal plasticity changes when the environment changes.

The Genes That Are Under Selection Were Related to Oxygen Transport and Oxidative Stress

We obtained a total of 235.65 Gb of clean reads from genome resequencing. After mapping to the genome, the reads with unique read alignments were used to analyze common genetic polymorphisms, whereas the remaining reads were discarded. In all, 10,819,586 SNPs were identified from all chickens, of which 150,897 were from the exonic region (**Supplementary Table 5**).

To detect the regions with extreme divergence in allele frequency (*Fst*) on autosomes, we scanned the genome regions in 40-kb sliding windows. In total, we identified 82 candidate regions (with Fst ≥ 0.3) that included 50 genes. We further investigated the function of these genes. Results showed that the top-ranked gene ontology (GO) terms were related to vasculature development, including regulation of blood vessel size, regulation of blood vessel diameter, regulation of tube size, and vascular process in the circulatory system. Comparisons against the Kyoto Encyclopedia of Genes and Genomes (KEGG) indicated that

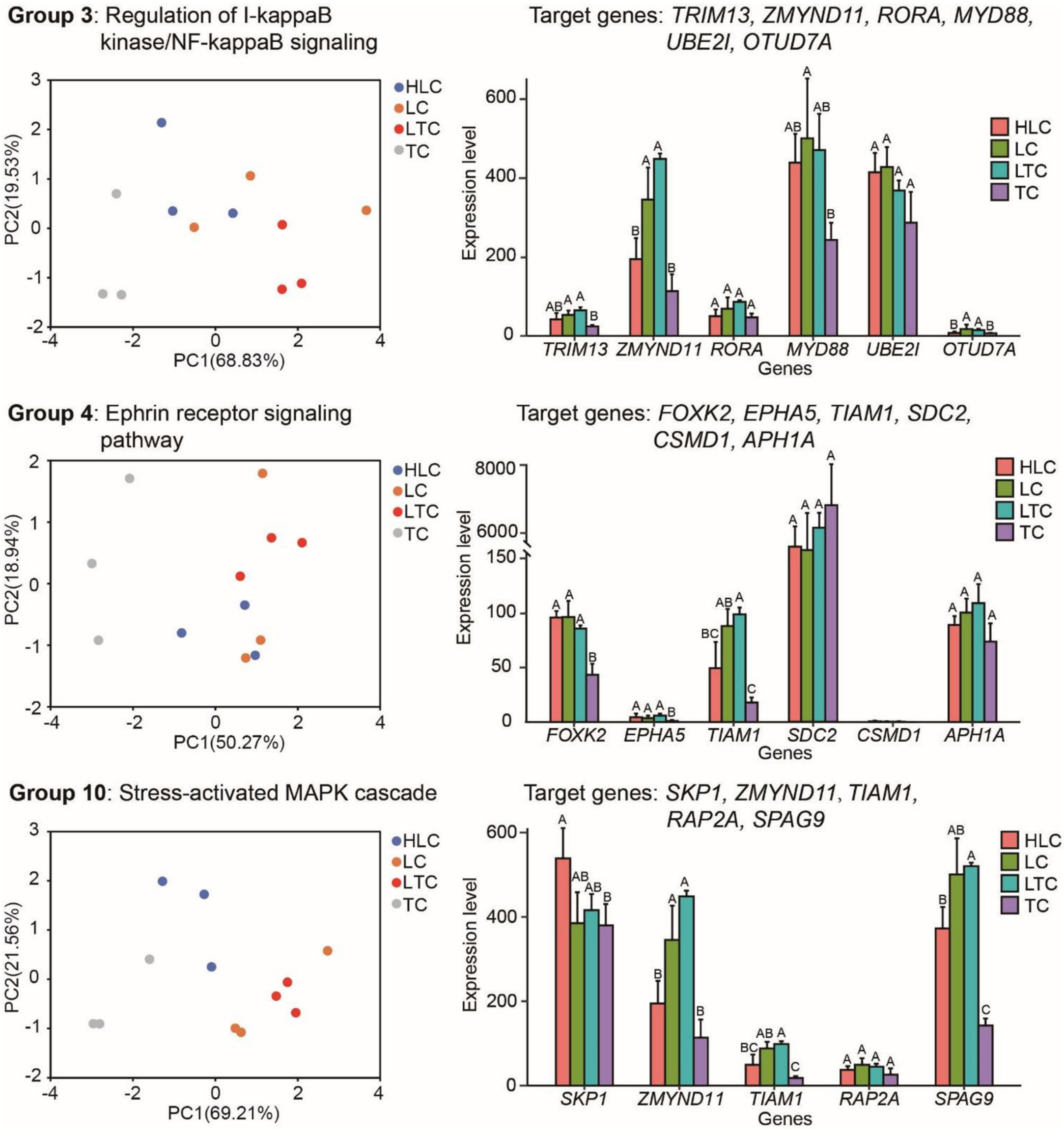


FIGURE 5 | Target genes of miR-10c-5p have different expression patterns between high and low-altitude chickens. The left, middle, and right panels represent the target genes of miR-10c-5p and the pathways these genes are involved in (left panels), PCA plots based on these target genes, and the expression of target genes of miR-10c-5p, respectively. Different letters mean that there is a statistically significant difference between groups ($P \leq 0.05$).

several genes were clustered into cytokine–cytokine receptor interaction and Rap1 signaling pathway (**Figure 6**).

Combined with transcriptome results, we further analyzed the expression of selected genes in multiple tissues (**Figure 7**). Among these genes, *ATP2B1*, *EGFR*, *KCNMA1*, *MTNR1B*, and *PREX1* were associated with the regulation of anatomical structure size, such as blood vessel diameter.

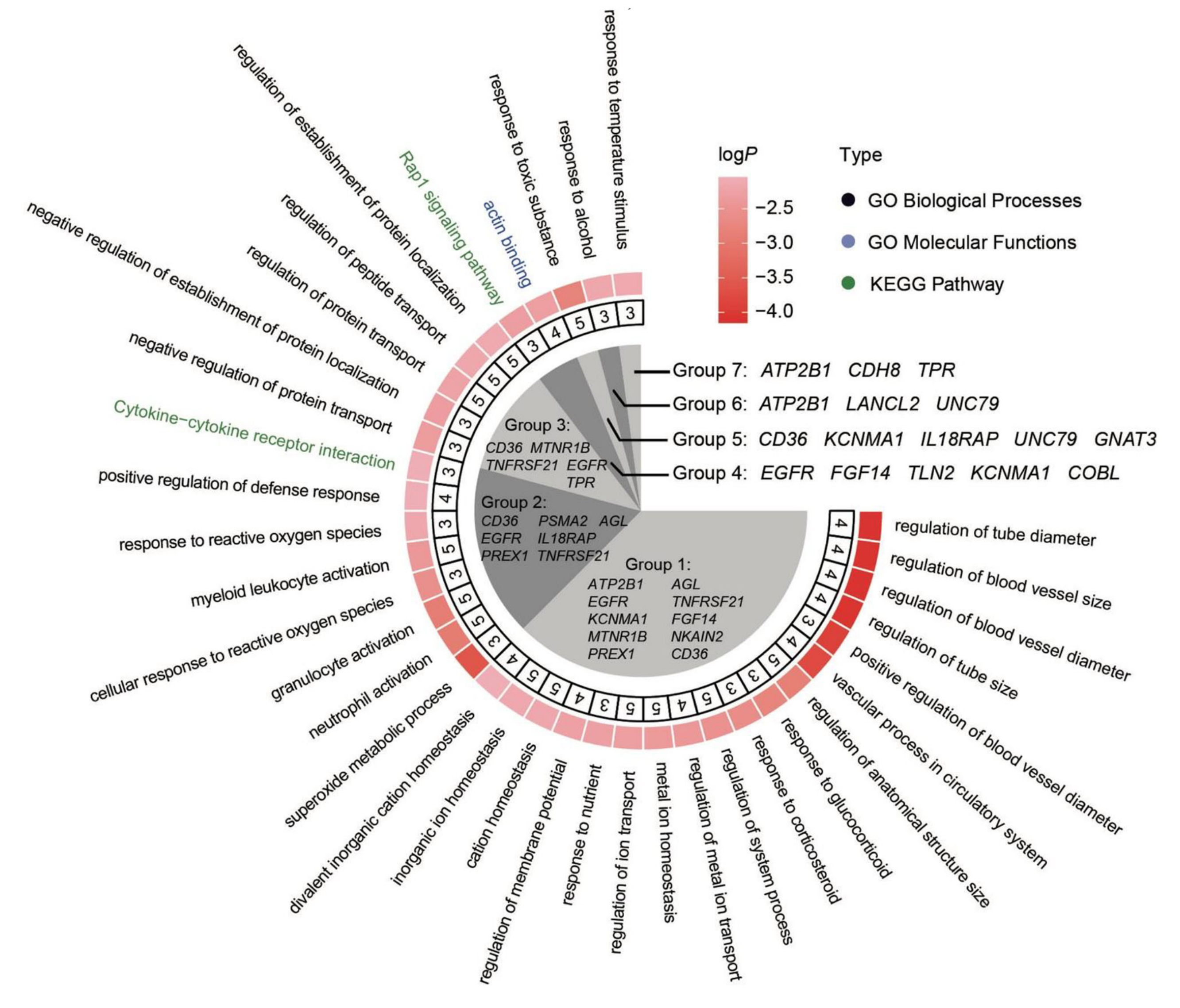


FIGURE 6 | Enrichment analysis of the genes under selection. All enrichment terms can be divided into seven groups. The genes involved in these groups were marked in the inner fan. The second annulus represents the number of genes that are involved in the specific terms.

DISCUSSION

The Tibetan chicken is a unique breed that has adapted to the high-altitude hypoxic conditions of the Tibetan plateau (30). A number of positively selected genes have been reported in these chickens; however, the mechanisms of gene expression and regulation for hypoxia adaptation are not fully understood. In this study, we conducted reciprocal transplant experiments of Tibetan chicken and Peng'xian yellow chicken to investigate whether some miRNAs have a plastic response to a different environment and play roles in the progress of plateau environment adaptation of chickens. Furthermore, through genome analysis, we found that the genes that are related to the regulation of blood vessel diameter and cellular response to reactive oxygen genes are under selection between Tibetan chicken and Peng'xian yellow chicken. Some selected genes showed altitude-specific expression, and it indicated that these genes are under regulation.

At the miRNA transcriptome level, we found more differentially expressed miRNAs (DEMs) in liver and lung samples than in the heart and brain, which indicated that the functions of the liver and lung could play an important role in the progress of chickens' plateau environmental adaptation. This is similar to other mammals living on the Tibetan plateau. The low-altitude cattle (*Bos taurus*) are closely related to domestic yak (*Bos grunniens*). The divergence time for them is five million years ago (11). The research on yaks and cattle found a total of 85 differentially expressed mature miRNAs in the lung (70 DEMs) and heart (29 DEMs) with a threshold fold change > 2 and FDR < 0.05 (47). The research on Tibetan

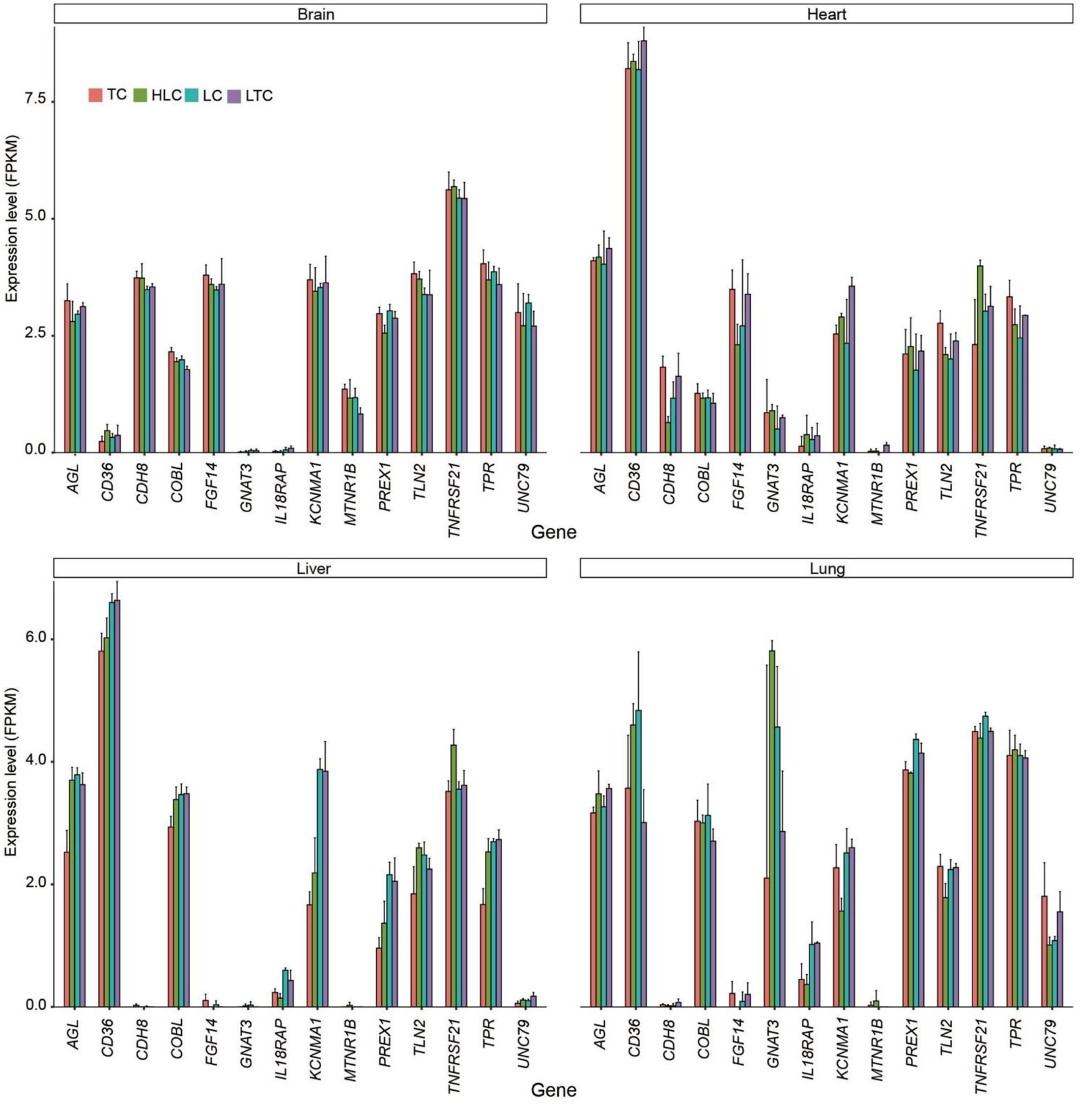


FIGURE 7 | The expression level ($\log_2$FPKM) of selected genes in four tissues.

pigs and Yorkshire pigs found 51 DEMs in the liver with the threshold of Fisher's exact test $P < 0.01$ and fold change > 2 (48). These results suggest that neural activity is very sensitive to the change in oxygen concentration. When the body is in a low-oxygen environment, the oxygen may be preferentially used for brain supply, resulting in a sufficient oxygen supply for the brain.

Correlation analysis and PCA showed a high correlation and low difference between every two samples in the same tissue. In fact, we only found a few DEMs between every two groups. Through the selection of DEMs of all comparative groups, we observed that some miRNAs have a significant environmental preference. Further analysis revealed that, although Tibetan chickens adapt to the highland in many ways, it shows similar characteristics as lowland chickens in miRNA and transcriptome when Tibetan chickens are brought back to the lowland (10). Some miRNAs and genes of Peng'xian yellow chickens also show plastic changes when they grow up in the highland.

These miRNAs and genes show a similarity expression pattern with highland chickens. In short, our findings indicated that several miRNAs could play a key role in the process of plateau environmental adaptation of chickens, such as miR-10c-5p and miR-1388b-5p.

It is known that hypoxia can induce the expression of inflammatory cytokines and chemokines (49, 50). The liver, as an immune organ, plays an important role in the hepatic inflammatory response (51). The vast majority of hepatic T cells (Th1/Tc1) secrete inflammatory cytokines, including interferon-γ, TNF-α, and interleukin-2 (52). In our study, miR-10c-5p may play an important role in regulating inflammatory responses in the liver. It always has a high expression level in the livers of lowland experimental chickens, whether Tibetan chickens or Peng'xian yellow chickens. Functional enrichment analysis and pathways analysis showed that some target genes of miR-10c-5p were involved in the "regulation of I-kappaB kinase/NF-kappaB signaling," "ephrin receptor signaling pathway," and "stress-activated MAPK cascade." A recent study also indicates that genes related to Tibetan chicken environmental adaptations were involved in the "regulation of I-kappaB kinase/NF-kappaB signaling" (24) and MAPK (29) pathways. Differentially expressed miRNAs analysis also found the targeted genes were involved in the I-kappa B kinase/NF-kappa B signaling pathway (53). These previous and current studies suggest the key role of the I-kappa B kinase/NF-kappa B signaling pathway in the hypoxia adaptation of Tibetan chickens.

At the mRNA transcriptome level, further analysis of miR-10c-5p's target genes that were involved in the enrichment groups as described above found that four genes, namely, *ZMYND11*, *OTUD7A*, *TRIM13*, and *TIAM1*, were expressed differentially between lowland and highland experimental chickens. These genes have significantly lower expression in the TC group than in the lowland experimental groups ("LTC" and "LC"). While in the HLC group, the expression level of these genes is biased to the TC group. It indicates that these genes could play an important role in the progress of high-altitude adaptation of chickens, such as *ZMYND11*. *ZMYND11*, and zinc finger MYND-type containing 11, is known as *BS69* and works as a transcriptional repressor (54, 55). Research shows that *ZMYND11* could negatively regulate the NF-kappaB signaling pathway in chickens (56). In this study, we found that *ZMYND11* has a lower expression level in high-altitude experimental chickens. It indicates that miR-10c-5p may activate the NF-κB signaling pathway by regulating *ZMYND11* in high-altitude chickens' livers, further promoting pro-inflammatory cytokine expression to defend against the tissue inflammation that is caused by hypoxia at highland. The specific roles of these genes in the form of high-altitude adaptation need more studies to determine.

At the genome level, a total of 50 candidate genes were identified using a sliding window analysis. Hypoxia typically causes a release of intracellular Ca^{2+}, mediated by the ryanodine receptors, which then leads to increased cell contraction (57). Consistent with previous studies, which indicated that several candidate genes in the calcium-signaling pathway are possibly involved in adaptation to the hypoxia experienced by Tibetan chickens (26). In this study, *ATP2B1* (ATPase plasma membrane Ca2+ transporting 1) plays a key role in intracellular calcium homeostasis and is further involved in vascular smooth muscle contraction and regulation of blood pressure (58, 59), and was also identified to be potentially positively selected. *EGFR* (epidermal growth factor receptor) is the receptor for EGF, which is involved in the mitogenic, stimulating proliferation of many cell types, including human microvascular endothelial cells (60) and lymphatic endothelial cells (61). *PREX1* (phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1) is a member of the Rac exchange factor family of guanine nucleotide exchange factors. The *PREX* proteins can activate Rho GTPases, which regulate cell motility, proliferation, glucose uptake, and reactive oxygen species generation (62). Some selected genes showed environment-specific expressions, such as *KCNMA1* and *PREX1*. In the lung, the expression levels of these genes were similar in low-altitude chickens ("LC" and "LTC") and similar in high-altitude chickens ("TC" and "HLC"). It showed that the expression of these genes is under the action of regulatory factors.

CONCLUSION

In summary, our study uncovers that several miRNAs have a plastic change alongside living environment (altitude) changes. These miRNAs could be involved in the response to hypoxia, inflammation, or other stresses in high-altitude environments and help chickens adapt. The target genes of these miRNAs also showed differential expression under different environmental conditions. Further genome analyses revealed that several candidate genes in the calcium-signaling and immunity pathways are possibly involved in adaptation to the hypoxia experienced by these Tibetan chickens. The I-kappa B kinase/NF-kappa B signaling pathway was widely found in the hypoxia adaptation of Tibetan chickens. The candidate differentially expressed miRNAs, genes, and selected genes identified in this study may be useful targets for breeding efforts to develop improved chicken breeds for the Tibet plateau.

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 Institutional Animal Care and Use Committee at the Chengdu University.

AUTHOR CONTRIBUTIONS

DL and TW conceived the study and the analytical strategy. BC wrote the manuscript. BR and PZ performed the statistical analyses. TW revised the article. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by Sichuan Science and Technology Program Grants 2019JDTD0009, 2020YFH0138, and 2020YFSY0040.

ACKNOWLEDGMENTS

We thank Xueying Ma for helping with the sample collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2022.911685/full#supplementary-material

REFERENCES

1. Wen J, Ning Y. Analysis on the parameters of blood in healthy Tibetan mastiffes. *J Qinghai Med College.* 20:7–9.
2. Dolt KS, Mishra MK, Karar J, Baig MA, Ahmed Z, Pasha MA. cDNA cloning, gene organization and variant specific expression of HIF-1 alpha in high altitude yak (Bos grunniens). *Gene.* (2007) 386:73–80. doi: 10.1016/j.gene.2006.08.004
3. Zong BY, Peng S, Feng LJ, Jiu WD, Zhuo DB, Hong W, et al. Preliminary study on pulmonary tissue and hypoxia adaptation to plateau for tibetan pigs. *Arch Anim Breed.* (2012) 40:541–55. doi: 10.1007/s10745-012-9487-2
4. Gollin E. *Developmental Plasticity and Evolution*. Oxford: Oxford University Press. (2003).
5. Price TD, Qvarnstrom A, Irwin DE. The role of phenotypic plasticity in driving genetic evolution. *Proc Biol Sci.* (2003) 270:1433–40. doi: 10.1098/rspb.2003.2372
6. Ghalambor C, McKay J, Carroll S, Reznick D. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Funct Ecol.* (2007) 21:394–407. doi: 10.1111/j.1365-2435.2007.01283.x
7. Ghalambor CK, Hoke KL, Ruell EW, Fischer EK, Reznick DN, Hughes KA. Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature.* (2015) 525:372–5. doi: 10.1038/nature15256
8. Ho WC, Zhang J. Evolutionary adaptations to new environments generally reverse plastic phenotypic changes. *Nat Commun.* (2018) 9:350. doi: 10.1038/s41467-017-02724-5
9. Ho WC, Zhang J. Genetic gene expression changes during environmental adaptations tend to reverse plastic changes even after the correction for statistical nonindependence. *Mol Biol Evol.* (2019) 36:1847–8. doi: 10.1093/molbev/msz073
10. Ho W-C, Li D, Zhu Q, Zhang J. Phenotypic plasticity as a long-term memory easing readaptations to ancestral environments. *Sci Adv.* (2020) 6:eaba3388. doi: 10.1126/sciadv.aba3388
11. Qiang Q, Guojie Z, Tao M, Wubin Q, Junyi W, Zhiqiang Y, et al. The yak genome and adaptation to life at high altitude. *Nat Genet.* (2012) 44:946–9. doi: 10.1038/ng.2343
12. Ge RL, Cai Q, Shen YY, San A, Lan M, Yong Z, et al. Draft genome sequence of the Tibetan antelope. *Nat Commun.* (2013) 4:1858. doi: 10.1038/ncomms2860
13. Mingzhou L, Shilin T, Long J, Guangyu Z, Ying L, Yuan Z, et al. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet.* (2013) 45:1431–U1180. doi: 10.1038/ng.2811
14. Kharrati-Koopaee H, Ebrahimie E, Dadpasand M, Niazi A, Esmailizadeh A. Genomic analysis reveals variant association with high altitude adaptation in native chickens. *Sci Rep.* (2019) 9:9224. doi: 10.1038/s41598-019-45661-7
15. Li J-T, Gao Y-D, Xie L, Deng C, Shi P, Guan M-L, et al. Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. *Proc Natl Acad Sci U S A.* (2018) 115:8406–11. doi: 10.1073/pnas.1805348115
16. Sun CY, She XM, Qin Y, Chu ZB, Chen L, Ai LS, et al. miR-15a and miR-16 affect the angiogenesis of multiple myeloma by targeting VEGF. *Carcinogenesis.* (2013) 34:426–35. doi: 10.1093/carcin/bgs333
17. Liu J, Sun F, Wang X, Bi Q. miR-27b promotes angiogenesis and skin repair in scalded rats through regulating VEGF-C expression. *Lasers Med Sci.* (2020) 35:1577–88. doi: 10.1007/s10103-020-02991-7
18. He M, Zhou W, Li C, Guo M. MicroRNAs, DNA damage response, and cancer treatment. *Int J Mol Sci.* (2016) 17. doi: 10.3390/ijms17122087
19. Zeng H, Hu M, Lu Y, Zhang Z, Xu Y, Wang S, et al. microRNA 34a promotes ionizing radiation–induced DNA damage repair in murine hematopoietic stem cells. *FASEB J.* (2019) 33:8138–47. doi: 10.1096/fj.201802639R
20. Chan JL, Hu XX, Wang CC, Xu QH. miRNA-152 targets GATA1 to regulate erythropoiesis in Chionodraco hamatus. *Biochem Biophys Res Commun.* (2018) 501:711–7. doi: 10.1016/j.bbrc.2018.05.053
21. Xu P, Palmer LE, Lechauve C, Zhao GW, Yao Y, Luan J, et al. Regulation of gene expression by miR-144/451 during mouse erythropoiesis. *Blood.* (2019) 133:2518–28. doi: 10.1182/blood.2018854604
22. Sun L, Fan F, Li R, Niu B, Zhu L, Yu S, et al. Different erythrocyte MicroRNA profiles in low- and high-altitude individuals. *Front Physiol.* (2018) 9:1099. doi: 10.3389/fphys.2018.01099
23. Long K, Feng S, Ma J, Zhang J, Jin L, Tang Q, et al. Small non-coding RNA transcriptome of four high-altitude vertebrates and their low-altitude relatives. *Sci Data.* (2019) 6, 192. doi: 10.1038/s41597-019-0204-5
24. Yuan J, Li S, Sheng Z, Zhang M, Liu X, Yuan Z, et al. Genome-wide run of homozygosity analysis reveals candidate genomic regions associated with environmental adaptations of Tibetan native chickens. *BMC Genom.* (2022) 23:91. doi: 10.1186/s12864-021-08280-z
25. Giussani DA, Salinas CE, Villena M, Blanco CE. The role of oxygen in prenatal growth: studies in the chick embryo. *J Physiol.* (2007) 585:911–7. doi: 10.1113/jphysiol.2007.141572
26. Wang MS, Li Y, Peng MS, Zhong L, Wang ZJ, Li QY, et al. Genomic analyses reveal potential independent adaptation to high altitude in tibetan chickens. *Mol Biol Evol.* (2015) 32:1880–9. doi: 10.1093/molbev/msv071
27. Zhang Q, Gou W, Wang X, Zhang Y, Ma J, Zhang H, et al. Genome resequencing identifies unique adaptations of tibetan chickens to hypoxia and high-dose ultraviolet radiation in high-altitude environments. *Genome Biol Evol.* (2016) 8:765–76. doi: 10.1093/gbe/evw032
28. Zhang Z, Qiu M, Du H, Li Q, Yu C, Gan W, et al. Whole genome re-sequencing identifies unique adaption of single nucleotide polymorphism, insertion/deletion and structure variation related to hypoxia in Tibetan chickens. *Gene Expr Patterns.* (2021) 40:119181. doi: 10.1016/j.gep.2021.119181
29. Zhang Y, Su W, Zhang B, Ling Y, Kim WK, Zhang H. Comprehensive analysis of coding and non-coding RNA transcriptomes related to hypoxic adaptation in Tibetan chickens. *J Anim Sci Biotechnol.* (2021) 12:60. doi: 10.1186/s40104-021-00582-2
30. Zhang Y, Zheng X, Zhang Y, Zhang H, Zhang X, Zhang H. Comparative transcriptomic and proteomic analyses provide insights into functional genes for hypoxic adaptation in embryos of Tibetan chickens. *Sci Rep.* (2020) 10:11213. doi: 10.1038/s41598-020-68178-w
31. Hao R, Hu X, Wu C, Li N. Hypoxia-induced miR-15a promotes mesenchymal ablation and adaptation to hypoxia during lung development in chicken. *PLoS ONE.* (2014) 9:e98868. doi: 10.1371/journal.pone.0098868
32. Chen M, Zhang S, Xu Z, Gao J, Mishra SK, Zhu Q, et al. MiRNA profiling in pectoral muscle throughout pre- to post-natal stages of chicken development. *Front Genet.* (2020) 11:570–570. doi: 10.3389/fgene.2020.00570
33. Zhang Y, Gou W, Zhang Y, Zhang H, Wu C. Insights into hypoxic adaptation in Tibetan chicken embryos from comparative proteomics. *Comp Biochem Physiol Part D Genomics Proteomics.* (2019) 31:100602. doi: 10.1016/j.cbd.2019.100602

34. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* (2011) 17:13. doi: 10.14806/ej.17.1.200

35. Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics.* (2010) 11:7. doi: 10.1002/0471250953.bi1107s32

36. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* (2012) 40:37–52. doi: 10.1093/nar/gkr688

37. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 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

38. Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139. doi: 10.1093/bioinformatics/btp616

39. Bolger AM, Marc L, Bjoern U. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170

40. Kai W, Mingyao L, Hakon H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010) 38:e164. doi: 10.1093/nar/gkq603

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

42. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read massively parallel ligation sequencing using two-base encoding. *Genome Res.* (2009) 19:1527–41. doi: 10.1101/gr.091868.109

43. Pfeifer B, Wittelsbürger U, Onsins S, Lercher MJ. PopGenome: an efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol.* (2014) 31:1929–36. doi: 10.1093/molbev/msu136

44. Zhou Y, Zhou B, Pache L, Chang M, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* (2019) 10:1523. doi: 10.1038/s41467-019-09234-6

45. Eltzschig HK, Carmeliet P. Hypoxia and inflammation. *N Engl J Med.* (2011) 364:656. doi: 10.1056/NEJMra0910283

46. Chen T, Yang C, Li M, Tan X. Alveolar hypoxia-induced pulmonary inflammation: from local initiation to secondary promotion by activated systemic inflammation. *J Vasc Res.* (2016) 53:317–29. doi: 10.1159/000452800

47. Guan J, Long K, Ma J, Zhang J, Luo X. Comparative analysis of the microRNA transcriptome between yak and cattle provides insight into high-altitude adaptation. *PeerJ.* (2017) 5:e3959. doi: 10.7717/peerj.3959

48. Li Y, Li X, Sun WK, Cheng C, Chen YH, Zeng K, et al. Comparison of liver microRNA transcriptomes of tibetan and yorkshire pigs by deep sequencing. *Gene.* (2016) 577:244–50. doi: 10.1016/j.gene.2015.12.003

49. Ghezzi P, Dinarello CA, Bianchi M, Rosandich ME, Repine JE, White CW. Hypoxia increases production of interleukin-1 and tumor necrosis factor by human mononuclear cells. *Cytokine.* (1991) 3:189–94. doi: 10.1016/1043-4666(91)90015-6

50. Hartmann G, Tschop M, Fischer R, Bidlingmaier C, Riepl R, Tschop K, et al. High altitude increases circulating interleukin-6, interleukin-1 receptor antagonist and C-reactive protein. *Cytokine.* (2000) 12:246–52. doi: 10.1006/cyto.1999.0533

51. Gao B, Jeong WI, Tian Z. Liver: an organ with predominant innate immunity. *Hepatology.* (2008) 47:729–36. doi: 10.1002/hep.22034

52. Dancygier, H. (2010). *The Liver as an Immune Organ in Clinical Hepatology: Principles and Practice of Hepatobiliary Diseases*. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 141–52.

53. Zhang Z, Qiu M, Du H, Li Q, Yu C, Gan W, et al. Small RNA sequencing reveals miRNAs important for hypoxic adaptation in the Tibetan chicken. *Br Poult Sci.* (2020) 61:632–9. doi: 10.1080/00071668.2020.1792835

54. Hateboer G, Gennissen A, Ramos YF, Kerkhoven RM, Sonntag-Buck V, Stunnenberg HG, et al. BS69 a novel adenovirus E1A-associated protein that inhibits E1A transactivation. *EMBO J.* (1995) 14:3159–69. doi: 10.1002/j.1460-2075.1995.tb07318.x

55. Wei G, Schaffner AE, Baker K, Mansky K, Ostrowski MC. Ets-2 interacts with Co-repressor BS69 to repress target gene expression. *Anticancer Res.* (2003) 23:2173. doi: 10.1245/ASO.2003.08.017

56. Hu Q, Zhao Y, Wang Z, Hou Y, Bi D, Sun J, et al. Chicken gga-miR-19a Targets ZMYND11 and plays an important role in host defense against Mycoplasma gallisepticum (HS Strain) infection. *Front Cell Infect Microbiol.* (2016) 6:102–102. doi: 10.3389/fcimb.2016.00102

57. Wang YX, Zheng YM. ROS-dependent signaling mechanisms for hypoxic Ca(2+) responses in pulmonary artery myocytes. *Antioxid Redox Signal.* (2010) 12:611–23. doi: 10.1089/ars.2009.2877

58. Kobayashi Y, Hirawa N, Tabara Y, Muraoka H, Umemura S. Mice lacking hypertension candidate gene ATP2B1 in vascular smooth muscle cells show significant blood pressure elevation. *Hypertension.* (2012) 59:854–60. doi: 10.1161/HYPERTENSIONAHA.110.165068

59. Shin YB, Lim JE, Ji SM, Lee HJ, Park SY, Hong KW, et al. Silencing of Atp2b1 increases blood pressure through vasoconstriction. *J Hypertens.* (2013) 31:1575–83. doi: 10.1097/HJH.0b013e32836189e9

60. Nezu E, Ohashi Y, Kinoshita S, Manabe R. Recombinant human epidermal growth factor and corneal neovascularization. *Jpn J Ophthalmol.* (1992) 36:401–6.

61. Liu NF, He QL. The regulatory effects of cytokines on lymphatic angiogenesis. *Lymphology.* (1997) 30:3–12.

62. Barrows D, He JZ, Parsons R. PREX1 protein function is negatively regulated downstream of receptor tyrosine kinase activation by p21-activated Kinases (PAKs). *J Biol Chem.* (2016) 291:20042–54. doi: 10.1074/jbc.M116.723882

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

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

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 18 August 2022
DOI 10.3389/fgene.2022.934757

Genetic analysis of morphological traits in spring wheat from the Northeast of China by a genome-wide association study

Wenlin Liu[1], Yuyao Li[2], Yan Sun[1], Jingquan Tang[1], Jingyu Che[3], Shuping Yang[1], Xiangyu Wang[1], Rui Zhang[4] and Hongji Zhang[1]*

[1]Crop Resources Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, China, [2]Heilongjiang Academy of Agricultural Sciences, Harbin, China, [3]KeShan Branch of Heilongjiang Academy of Agricultural Sciences, Qiqihaer, China, [4]Institute of Forage and Grassland Sciences, Heilongjiang Academy of Agricultural Sciences, Harbin, China

OPEN ACCESS

EDITED BY
Kefei Chen,
Curtin University, Australia

REVIEWED BY
Chengsong Zhu,
University of Texas Southwestern Medical Center, United States
Rajiv Sharma,
Scotland's Rural College,
United Kingdom

*CORRESPONDENCE
Hongji Zhang,
zhanghongji1969@163.com

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

RECEIVED 03 May 2022
ACCEPTED 15 July 2022
PUBLISHED 18 August 2022

CITATION
Liu W, Li Y, Sun Y, Tang J, Che J, Yang S, Wang X, Zhang R and Zhang H (2022), Genetic analysis of morphological traits in spring wheat from the Northeast of China by a genome-wide association study.
Front. Genet. 13:934757.
doi: 10.3389/fgene.2022.934757

Identification of the gene for agronomic traits is important for the wheat marker-assisted selection (MAS) breeding. To identify the new and stable loci for agronomic traits, including flag leaf length (FLL), flag leaf width (FLW), uppermost internode length (UIL), and plant morphology (PM, including prostrate, semi-prostrate, and erect). A total of 251 spring wheat accessions collected from the Northeast of China were used to conduct genome-wide association study (GWAS) by 55K SNP arrays. A total of 30 loci for morphological traits were detected, and each explained 4.8–17.9% of the phenotypic variations. Of these, 13 loci have been reported by previous studies, and the other 17 are novel. We have identified seven genes involved in the signal transduction, cell-cycle progression, and plant development pathway as candidate genes. This study provides new insights into the genetic basis of morphological traits. The associated SNPs and accessions with more of favorable alleles identified in this study could be used to promote the wheat breeding progresses.

KEYWORDS

GWAS, marker-assisted selection, *Triticum aestivum* L, spring wheat, morphological traits

Abbreviations: BLUE, Best linear unbiased estimation; FLL, Flag leaf length; FLW, Flag leaf width; PM, Plant morphology; GWAS, Genome-wide association study; hb2, Broad-sense heritability; KASP, Kompetitive allele-specific PCR; LD, Linkage disequilibrium; MAS, Marker-assisted selection; QTL, Quantitative trait loci; R2, Phenotypic variance explained; RIL, Recombinant inbred line; SNP, Single nucleotide polymorphism; UIL, Uppermost internode length.

Background

Common wheat (*Triticum aestivum* L.) is an essential cereal crop and provides nearly 20% of the total caloric input to the global population (He et al., 2010; Wang et al., 2019). Due to the complicated genetic architecture (Gao et al., 2017; Li et al., 2021), the progresses for the improvement of morphological traits are difficult. Heilongjiang and Jilin provinces located at the Northeast of China are the major spring wheat-producing regions (Li et al., 2021). Although the wheat production has been improved largely in the past decades, it is still not enough to meet the needs of the people (He et al., 2010; Xu et al., 2021).

Morphological traits include various traits, such as flag leaf length (FLL), flag leaf with (FLW), uppermost internode length (UIL), and plant morphology (PM, including prostrate, semi-prostrate, and erect) (Li et al., 2018; Li et al., 2021). Both the genetic and environmental factors influenced morphological traits. Marker-assisted selection (MAS) based on molecular marker is an available and effective approach for the further improvement of morphological traits of wheat (He et al., 2010; Rasheed et al., 2016; Wang et al., 2020). However, the effective and reliability of MAS depend on the number and quality of available genes/QTLs and associated markers for target traits. Until now, over 30 genes associated with morphological traits were cloned by homologous cloning or map-based cloning, and over 50 functional markers or kompetitive allele-specific PCR (KASP) markers were developed (Cui et al., 2014; Rasheed et al., 2016; Nadolska-Orczyk et al., 2017; Lozada et al., 2019). Besides, over 120 loci for morphological traits were detected by linkage mapping and association analysis (Gao et al., 2017; Würschum et al., 2017; Rahimi et al., 2019; Pang et al., 2020; Li et al., 2021). However, the loci/gene is still not enough for the improvement of wheat morphological traits. Identifying the novel genes or loci for morphological traits is urgent.

Single nucleotide polymorphism (SNP), insertion and deletion (InDel), simple sequence repeats (SSRs), and diversity array technology (DArT) markers are the mostly used molecular marker for genotyping (Wang et al., 2014; Rasheed et al., 2016). Compared with SSR and DArT markers, SNPs have more abundant and higher coverage. With the development of the gene chip and NGS technology, getting SNP quickly becomes feasible and provides an effective way to identifying genes/QTLs for complex traits (Zhu et al., 2008; Liu et al., 2017; Wang et al., 2020). Now, the 55, 90, and 660K wheat SNP arrays are gradually replacing SSR and DArT markers in genetic analysis and have been widely used in the genetic analysis for yield (Gao et al., 2015; Sun et al., 2017; Beyer et al., 2019; Li et al., 2019), disease resistance (stripe rust, leaf rust, or powdery mildew), end-use quality, procession quality, and biotic or abiotic stress tolerance (drought or flood)-related traits (Liu et al., 2016a; Liu et al., 2017; Valluru et al., 2017; Liu et al., 2019; Li et al., 2021; Quan et al., 2021).

Linkage analysis and association mapping are the two main ways to uncover the genetic mechanism of complex traits (Zhu et al., 2008; Liu et al., 2016b; Liu et al., 2017). Compared with the bi-parental linkage analysis, association mapping is based on natural population (including wild types, landraces, released cultivars, and improved accessions) and offers an effective and reliable approach to uncover the genetic architecture of complex traits (Liu et al., 2017; Wang et al., 2020). Besides, linkage analysis focused on specific traits, whereas the GWAS could be used to analyze all traits based on the same set of genotype data (Zhu et al., 2008). Nowadays, association mapping has been widely used in the genetic analysis of complex traits in wheat, including grain yield-related traits, disease-related traits, and biotic and abiotic stresses (Cui et al., 2014; Zuo et al., 2020; Quan et al., 2021). The rapid, efficient, and accurate genotyping is the basis to conduct GWAS. Thus, the SNP chip provides an effective and feasible way for association mapping.

In this study, 251 spring wheat accessions mainly collected from the Northeast of China (Heilongjiang and Jilin province) were used to 1) identify loci for morphological traits in spring wheat, 2) get new insights into the genetic architecture of target traits, and 3) search for candidate genes for further study.

Materials and methods

Plant materials and field trials

A total of 251 spring wheat accessions from the Northeast of China (Heilongjiang or Jilin province) were collected for the GWAS of morphological traits (Supplementary Table S1). The diverse panel was grown at the Harbin and Keshan experimental station of the Heilongjiang academy of agricultural science in Heilongjiang province during the 2018–2019 and 2019–2020 cropping seasons. A complete randomized block design with three replicates was employed in field trials. For both Harbin and Keshan experimental stations, each plot comprised four 2.0 m rows spaced 20 cm apart, with 40 seeds in each row. Agronomic management was performed according to local practices.

Phenotyping and statistical analysis for flag leaf length, flag leaf width, uppermost internode length, and plant morphology

Four traits related to morphological were conducted in all four environments, including FLL, FLW, UIL, and PM. For FLL and FLW, 10 random flag leaves in each plot at the mid-grain-fill stage were used to measure FLL and FLW, represented by the distance between the base and the tip, and width at the widest point, respectively. UIL is the mean distance between the stem base and the top of spikes excluding awns and the mean length of the uppermost internode. Ten single plants in each plot were randomly selected at physiological maturity for measuring UIL. BLUE for four traits among four environments was calculated by a one-stage approach using the R package sommer (https://cran.r-project.org/web/packages/sommer/index.html) as $\boldsymbol{y} = 1_{\boldsymbol{n}}\mu + \boldsymbol{Xg} + \boldsymbol{Ze} + \boldsymbol{v} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the n-dimensional multi-environment phenotypic records, n is the number of multi-environment phenotypic records, $1_{\boldsymbol{n}}$ is an n-dimensional vector of ones, μ is the intercept, $\boldsymbol{g}$ and $\boldsymbol{e}$ are the vector of genetic and environment effects, respectively, $\boldsymbol{v}$ is the vector of genotype-by-environment interaction effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices of $\boldsymbol{g}$ and $\boldsymbol{e}$, respectively, and$\boldsymbol{\varepsilon}$ is the random residuals. $\boldsymbol{g}$, $\boldsymbol{e}$, $\boldsymbol{v}$, and $\boldsymbol{\varepsilon}$ were all assumed as random effects following normal distributions.

Genotyping, population structure, and linkage disequilibrium

A total of 3000 polymorphic and evenly distributed markers on 21 chromosomes were used to conduct population structure analysis by Structure v2.3.4 (Pritchard et al., 2000) (http://pritchardlab.stanford.edu/structure.html). Besides, principal component analysis (PCA) and neighbor-jointing (NJ) trees were also estimated using the Tassel v5.0 to validate the results of population structure analysis (Breseghello and Sorrells, 2006). The LD decay analysis was calculated for the whole genomes using the full matrix and sliding window options in Tassel v5.0 (Breseghello and Sorrells, 2006). The results of LD decay, population analysis, PCA, and NJ-tree analysis for the 251 spring wheat accessions have been reported by Li et al. (2021). The multi-environment phenotypic data were analyzed by a one-stage approach as $\boldsymbol{y} = 1_{\boldsymbol{n}}\mu + \boldsymbol{Xg} + \boldsymbol{Ze} + \boldsymbol{v} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the n-dimensional multi-environment phenotypic records, n is the number of multi-environment phenotypic records, $1_{\boldsymbol{n}}$ is an n-dimensional vector of ones, μ is the intercept, $\boldsymbol{g}$ and $\boldsymbol{e}$ are the vector of genetic and environment effects, respectively, $\boldsymbol{v}$ is the vector of genotype-by-environment interaction effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices of $\boldsymbol{g}$ and $\boldsymbol{e}$, respectively, and $\boldsymbol{\varepsilon}$ is the random residuals. $\boldsymbol{g}$, $\boldsymbol{e}$, $\boldsymbol{v}$, and $\boldsymbol{\varepsilon}$ were all assumed as random effects following normal distributions. The heritability was estimated using the entry-mean basis formula $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_v^2}{n_e} + \frac{\sigma_\varepsilon^2}{n_e \times n_r}}$, where σ_g^2, σ_v^2, and σ_ε^2 are the variance components of genetic effect, environment effect, and residual, n_e and n_e are the number of environments and number of replicates per accession in each environment, respectively. The one-stage phenotypic analysis was realized using the R package sommer (https://cran.r-project.org/web/packages/sommer/index.html) in statistical software R. We have added the statistical method in the M&M section. The SNP-based heritability was estimated by the GREML-LDMS method based on GCTA (Yang et al., 2011) (https://yanglab.westlake.edu.cn/software/gcta/) software.

Genome-wide association study and the identification of candidate genes

To eliminate the spurious marker-trait associations (MTAs) caused by environment variation, the mixed linear model (MLM, PCA + K) in Tassel v5.0 (Breseghello and Sorrells, 2006) was used as follows: $y = \mu + x\beta + u + e$. Of these, y is the vector of phenotype; μ is the mean; x represents the genotype; β is the effect of the SNP; and u is the random effects. The kinship matrix was treated as a random-effect factor, whereas the PCA was considered as a fixed-effect factor. Both the kinship and PCA matrix were calculated by the software Tassel v5.0. In this study, markers with a -$\log_{10}$ (p-value) $\geq$ 3.0 were regarded as MTAs. Manhattan plots and Q–Q plots were drawn by the CMplot package (https://cran.r-project.org/web/packages/CMplot/index.html) based on R 3.6.5.

The flanking sequences corresponding to the SNP markers (the SNPs in the LD decay interval) significantly associated with morphological traits were used in BLASTn and BLASTx searches against NCBI (http://www.ncbi.nlm.nih.gov/) databases. Besides, the annotation for IWGSC v2.1 was also used to identify candidate genes. Potential candidate genes were then selected with significant MTAs corresponding to non-synonymous SNPs in the coding region of the genes.

Quantitative real-time PCR (qRT-PCR) was conducted to test expression differences of the candidate genes in the accessions with extreme traits (Supplementary Table S2). The flag leaf used for test FLL and FLW was sampled for RNA extraction during the flag leaf fully drawn out; the stem used for test PM was sampled for RNA extraction during the erecting period; the uppermost internode used for test UIL was sampled for RNA extraction during the heading stage. Total RNA was extracted according to the Trizol method, whereas cDNA was synthesized with the HiScript II 1st Strand cDNA Synthesis Kit (Vazyme, Nanjing, China). The primers were designed with Primer Premier 5.0 software9 (Supplementary Table S3). PCR procedure was conducted in a volume of 20 μl, containing 2 μl cDNA, 0.4 μl of each primer, and a 10 μl ChamQ Universal SYBR qPCR Master Mix. The reaction was conducted in the ABI StepOnePlus Real-Time PCR System with Tower. The gene expression level was analyzed with $2^{-\Delta\Delta CT}$ method. *Actin1* was used as an internal control to normalize the expression levels of different samples. All assays were performed in two independent experiments with three repetitions.

Results

Phenotypic evaluation

Significant and continuous variations of adaptive traits were exhibited in the diverse panel. The BLUE values of FLL, FLW, UIL, and PM (0 means prostrate, one means semi-prostrate, two

TABLE 1 ANOVA analysis for the morphological traits in 251 spring wheat accessions from the Northeast of China.

Source of variation	*SS*				
	df	FLL	FLW	UIL	PM
Genotypes	250	7370.3**	6788.0**	65468.23**	303.5**
Environments	3	2274.3**	6726.8**	1187.8**	19.8**
Replicates (nested in environments)	8	144.2**	5717.06**	1133.7**	5.9**
Genotypes*Environments	750	5584.15**	8761.1**	17678.2**	193.4**
Error	1806				

*and ** indicate significance at 0.05 and 0.01 levels.

means erect) were 25.0 cm (19.4–29.0 cm), 14.7 mm (10.8–19.1 mm), 22.3 cm (14.1–33.3 cm), and 2.70 (1.3–3.0). The standard deviation and coefficient of variation of FLL, FLW, UIL, and PM were 7.25 cm (0.29), 2.94 mm (0.20), 5.58 cm (0.25), and 0.59 (0.22). The UIL was negatively correlated with FLW (−0.557, p < 0.001). FLL with FLW, PM and UIL, FLW with PM, and UIL with PM show no significant correlation (Supplementary Table S4). ANOVA showed highly significant effects (p < 0.01) of genotypes, environments, and genotype × environment interactions on all traits (Table 1). The h^2 estimates by a one-stage approach for FLL, FLW, UIL, and PM were 0.78, 0.50, 0.93, and 0.79, whereas the SNP-based heritabilities for FLL, FLW, UIL, and PM were 0.77, 0.50, 0.91, and 0.78, respectively, indicating that morphological traits were determined by genetic factors and affected by environment.

Genotyping, population structure, and linkage disequilibrium decay analysis

Totally, 52,503 polymorphic SNPs after filtering (MAF <0.05, missing rate >0.1) were conducted for further GWAS (Li et al., 2021). Among the SNP markers, 18323, 18691, and 15489 were from A, B, and D genomes, respectively, with an average marker density of 0.273 Mb per marker. All the 251 spring wheat accessions could be divided into three subgroups, subgroups I, II, and III. Of these, subgroup I contained 126 varieties mainly from Heilongjiang ranging from 1950s to 1980s; subgroup II had 75 varieties mainly from Heilongjiang ranging from 1990s to 2010s; whereas subgroup III comprised 50 varieties mainly from Jilin and foreign counties, including United States, Canada, and Japan. Furthermore, the average LD for the whole genome was 8 Mb according to the LOESS curve (Li et al., 2021).

Genome-wide association study

Thirty loci were detected associated with morphological traits (Table 2; Supplementary Table S5; Figure 1; Supplementary Figure S1). Among these, nine loci for FLL were identified on chromosomes 1, 2, 2A, 2B, 3B, 3D, 5D, 6A, and 7D, and each explained 5.7–9.5% of the phenotypic variances, respectively; seven for FLW were identified on chromosomes 1A (2), 5A, 5A, 5B, 5D, and 7A, and explained 5.1–17.9% of the phenotypic variances, respectively; seven for FLL were identified on chromosomes 2B, 5A, 6A, 6D, 7A, 7B, and 7D and explained 4.8–8.8% of the phenotypic variances, respectively; seven for PM were identified on chromosomes 2B, 2B, 3D, 4A, 5A, 5B, and 6A, and each explained 5.3–12.5% of the phenotypic variances, respectively.

Of these, the locus on chromosome 1A (556.0–577.7 Mb) is an identical locus, which showed effects on FLW and FLL; a locus on chromosome 2B (652.7–664.2 Mb) controlled both the FLL and ST; a locus on chromosome 5D (551.2–562.7 Mb) is the same loci for FLL and FLW. Besides, FLL and UIL have the same locus on chromosome 7D (5.4–8.3 Mb).

Candidate genes

Totally, seven candidate genes for morphological traits were identified (Table 3). Two candidate genes encoded for the E3 ubiquitin-protein ligase-like protein (*TraesCS1A01G164400* and *TraesCS2B01G466600*) were identified in the LD decay of the loci on 1A (296.9–297.7 Mb) and 2B (652.7–656.0 Mb) chromosomes. Another gene encoding a cytokinin riboside 5′-monophosphate phosphoribohydrolase (*TraesCS1A01G362500*) was identified in the LD decay of the loci on chromosome 1A (556.0–587.4 Mb). For the loci on chromosome 5A (383.4–385.0 Mb and 595.4–597.7 Mb), candidate genes *TraesCS5A01G182500* and *TraesCS5A01G406800* were identified, which encode the F-box family protein. Besides, the genes *TraesCS2B01G123800* and *TraesCS4A01G055000* encoding the C2H2 zinc finger and leucine-rich repeat receptor protein kinase were identified as the candidate gene for the loci 2B (91.0–92.6 Mb) and 4A (41.4–46.2 Mb). The qRT-PCR of seven candidate genes showed that *TraesCS2B01G466600* and *TraesCS5A01G406800* showed no significant differences

TABLE 2 Loci for morphological traits in 251 spring wheat accessions from the Northeast of China by association analysis.

Trait	Chr	Start (Mb)	End (Mb)	R^2		*p*-value		Environment	Favorable allele	References
				Min	Max	Min	Max			
FLL	1A	572.9	577.7	5.9%	9.0%	2.6E-05	8.6E-04	E1, E3, E4, BLUE	C	Li et al. (2021)
FLL	2A	66.3	68.2	6.0%	9.5%	1.7E-05	8.7E-04	E4, BLUE	G	
FLL	2A	554.6	560.2	5.8%	7.3%	2.2E-04	9.3E-04	E4, BLUE	C	Li et al. (2021)
FLL	2B	652.7	656.0	6.1%	7.9%	1.4E-04	9.2E-04	E1, E2, E4, BLUE	C	Liu et al. (2018)
FLL	3B	121.4	129.6	5.8%	9.3%	1.8E-05	9.8E-04	E1, E2, E4, BLUE	G	Wu et al. (2016)
FLL	3D	82.8	94.5	5.9%	7.8%	1.2E-04	7.7E-04	E1, E2, E4, BLUE	G	
FLL	5D	551.2	556.2	5.9%	8.3%	5.5E-05	9.2E-04	E2, E4, BLUE	G	
FLL	6A	5.1	17.9	5.8%	8.3%	5.4E-05	9.2E-04	E1, E2, E4, BLUE	T	
FLL	7D	5.4	10.5	5.7%	7.0%	2.3E-04	9.8E-04	E1, E4, BLUE	A	Wu et al. (2016)
FLW	1A	296.9	297.7	5.8%	6.6%	3.3E-04	8.4E-04	E1, E3, BLUE	C	
FLW	1A	556.0	587.4	5.3%	8.1%	1.8E-05	7.0E-04	E1, E2, BLUE	A	Li et al. (2021)
FLW	5A	290.1	299.5	6.0%	7.7%	1.5E-04	8.9E-04	E1, E3, E4, BLUE	A	
FLW	5A	595.4	597.7	6.2%	7.7%	1.1E-04	5.3E-04	E2	A	Wu et al. (2016)
FLW	5B	615.7	617.9	5.1%	10.7%	9.0E-05	9.4E-04	E1, E3, E1, BLUE	A	Li et al. (2021)
FLW	5D	562.0	562.7	5.7%	9.3%	1.8E-05	9.2E-04	E1, E3	A	
FLW	7A	18.8	25.9	5.8%	17.9%	1.4E-09	9.5E-04	E2, E1, E3	G	
UIL	2B	409.9	439.5	5.1%	8.2%	1.8E-05	9.1E-04	E1, E2, E3, E4, BLUE	G	
UIL	5A	383.4	385.0	5.6%	8.3%	4.6E-05	9.3E-04	E2	G	Li et al. (2020)
UIL	6A	606.6	611.7	5.2%	5.9%	9.3E-04	9.5E-04	E1, E2, E3, E4, BLUE	G	
UIL	6D	388.2	407.5	4.8%	6.8%	2.5E-04	9.0E-04	E2, E4, BLUE	G	Li et al. (2020)
UIL	7A	496.3	512.9	5.0%	6.8%	2.8E-04	8.1E-04	E1, E2, E3, E4, BLUE	C	
UIL	7B	701.1	701.1	6.6%	8.8%	6.5E-06	1.3E-04	E1, E3, E4, BLUE	A	Li et al. (2020)
UIL	7D	8.3	8.3	4.9%	6.0%	3.1E-04	9.6E-04	E1, E3, E4, BLUE	A	
PM	2B	91	92.6	6.1%	7.7%	1.0E-04	7.8 E-04	E2	A	
PM	2B	663.1	664.2	5.9%	6.6%	3.9E-04	9.2E-04	E2, E4	G	
PM	3D	602.8	606.7	5.9%	8.4%	5.1E-05	8.3E-04	E1, E3, E4	G	
PM	4A	41.4	46.2	5.3%	12.5%		3.7E-04	E2, E4, BLUE	C	Liu et al. (2016)

(Continued on following page)

TABLE 2 (*Continued*) Loci for morphological traits in 251 spring wheat accessions from the Northeast of China by association analysis.

Trait	Chr	Start (Mb)	End (Mb)	R^2		*p*-value		Environment	Favorable allele	References
				Min	Max	Min	Max			
						2.7E-06				
PM	5A	688.9	688.9	7.0%	9.9%	1.5E-06	4.4E-05	E1, E4, BLUE	C	
PM	5B	696.7	697	6.6%	8.9%	2.9E-05	5.0E-04	E1, E2, E3, BLUE	C	
PM	6A	23.8	24.4	6.9%	10.1%	7.5E-06	2.6E-04	E2, BLUE	C	Li et al. (2020)

FLL, flag leaf length; FLW, flag leaf width; UIL, uppermost internode length; PM, plant morphology.
The E1, E2, E3, E4, and E5 indicated the Haerbin 2018, Haerbin 2019, Keshan 2018, Keshan 2019, and the best linear unbiased estimation (BLUE), respectively.

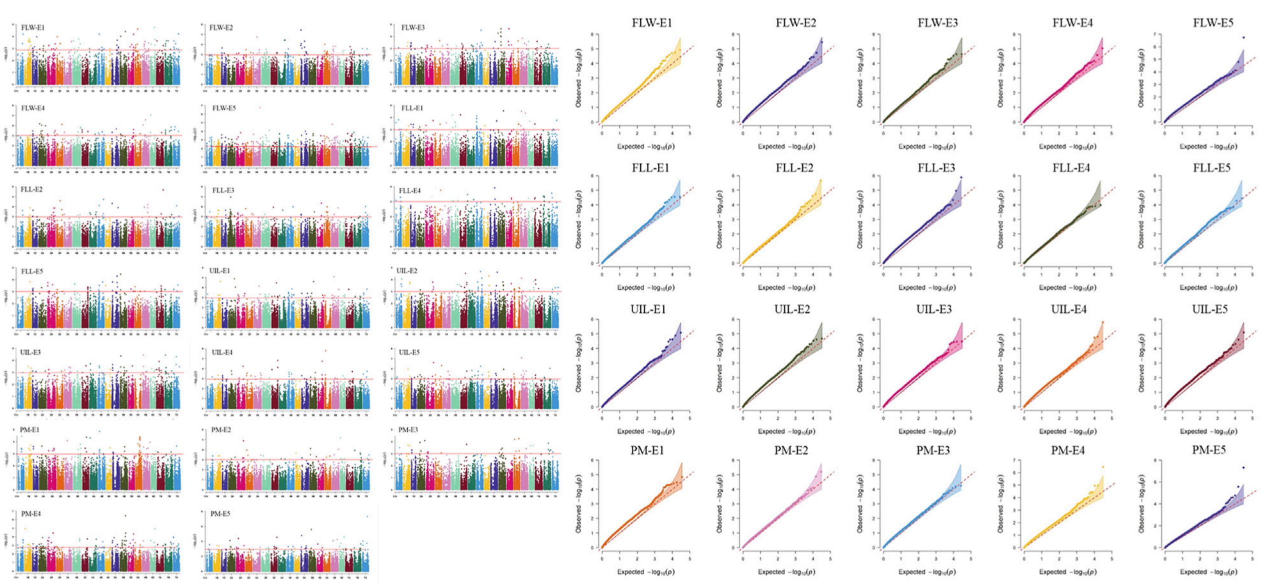


FIGURE 1
Manhattan and Q–Q plot for morphological traits in 251 spring wheat accessions from the Northeast of China analyzed by the mixed linear model (MLM) in Tassel v5.0. FLL, Flag leaf length; FLW, Flag leaf width; UIL, Uppermost internode length; PM, Plant morphology. The 1, 2, 3, 4, and 5 indicated the Harbin 2018, Harbin 2019, Keshan 2018, Keshan 2019, and the best linear unbiased estimation (BLUE), respectively.

TABLE 3 The details for the candidate genes of morphological traits.

Candidate gene	Chromosome	Start	End	Annotation
TraesCS1A01G164400	1A	296212311	296215426	E3 ubiquitin-protein ligase
TraesCS1A01G362500	1A	543067408	543069241	Cytokinin riboside 5′-monophosphate phosphoribohydrolase
TraesCS2B01G123800	2B	91304955	91305926	Zinc finger, C2H2
TraesCS2B01G466600	2B	662376739	662379204	E3 ubiquitin-protein ligase
TraesCS4A01G055000	4A	46141864	46145610	Leucine-rich repeat receptor protein kinase
TraesCS5A01G182500	5A	382291561	382292892	F-box protein-like protein
TraesCS5A01G406800	5A	597793675	597796893	F-box family protein

between the extreme accessions; *TraesCS1A01G164400*, *TraesCS4A01G055000,* and *TraesCS1A01G362500* showed 1.6–3.2-fold higher expressions among the extreme accessions, whereas *TraesCS5A01G182500* and *TraesCS2B01G123800* showed more than 1.9–7.6-fold lower expressions among extreme accessions (Supplementary Figure S1).

Discussion

All the 251 spring accessions could be divided into three subgroups (Li et al., 2021), and the characterization of the subgroups was largely consistent with geographic origins, released years, and pedigrees. Most of the cultivars from Heilongjiang ranging from 1950s to 1980s belonged to subgroup-1; the accessions from Heilongjiang ranging from 1990s to 2010s belonged to subgroup-2, whereas subgroup-3 mainly including the accessions from the Jilin and foreign counties (United States, Canada). Population analysis (PCA, NJ-tree, and Kinship) indicated that a significant population structure existed in the 251 spring wheat accessions. The lack of appropriate correction for population structure can lead to spurious MTAs (Zhu et al., 2008). Thus, to eliminate spurious MTAs, an MLM model with subpopulation data (Q) (fixed-effect factors) and kinship matrix (random-effect factor) was conducted. LD decay was influenced by population structure, allele frequency, recombination rate, and selection progresses and significantly affected the precision of association mapping. The LD decay for the whole genome was about 8 Mb, consistent with previous studies (Liu et al., 2017). The number of markers is enough for the subsequent association analysis.

Comparison with the QTL or gene in previous studies

The genes or loci associated with morphological traits were extensively reported previously (Wu et al., 2016; Li et al., 2018; Li et al., 2021). In this study, the association of morphological traits was performed and 30 significant loci were detected.

Several studies for flag leaf-related traits in common wheat have reported. We have identified nine loci for FLL in eight different chromosomes. Li et al. (2020) have identified eight loci for FLL on chromosomes 1A, 2A (2), 2B, 3A, 5A, 6B, and 6D, and each explained 6.9–19.6% of the phenotypic variances. Of these, the loci identified in the 1A and 2A were overlapped with the regions 1A (572.9–577.7 Mb) and 2A (554.6–560.2 Mb) identified in this study. Wu et al. (2016) mapped two FLL QTL on chromosomes 3B and 7D, which is overlapped with the FLL locus (3B: 121.4–129.6 Mb and 7D: 5.4–10.5 Mb) in the present study. Another FLL QTL was previously identified on chromosome 2B linked with the SSR marker *barc318* (Liu et al., 2018); it is coincided with the present FLL locus (2B: 652.7–656.0 Mb) based on the consensus linkage map (Maccaferri et al., 2015). These loci on chromosomes 2A (66.3–68.2 Mb), 3D (82.8–94.5 Mb), 5D (551.2–556.2 Mb), and 6A (5.1–17.9 Mb) were potential novel loci. Li et al. (2020) have identified five loci for FLW on chromosomes 1A, 3B, 5B (2), and 6B, accounting for 6.9–11.4% of the phenotypic variances. The locus on chromosome 5B (*AX_109519234*) was coincided with the locus (5B: 615.7–617.9 Mb) detected in this study. Besides, the locus on chromosome 1A (296.9–297.7 Mb) was also overlapped with the locus on 1A (*AX_111540798*) identified by Li et al. (2018). Wu et al. (2016) reported a locus for both FLW and flag leaf angle at the position (*IWB4576*) on chromosome 5A, showing a pleiotropic effect, which coincided with the loci identified in this study (5A: 290.1–299.5 Mb). In conclusion, the loci identified in 1A (556.0–587.4 Mb), 5A (595.4–597.7 Mb), 5D (562.0–562.7 Mb), and 7A (18.8–25.9 Mb) were potentially new.

UIL is important for the construction of plant architecture and influences the yield. Several studies have focused on the genetic basis of UIL. Wu et al. (2016) have identified seven loci for UIL. Li et al. (2020) have identified 12 loci for UIL on chromosomes 1A (2), 1B (2), 3A, 5A, 6B (3), 6D (2), and 7B, with a single locus explaining 6.7–16.4% of the phenotypic variances. Of these, the loci 1A (*AX_109449226*), 6D (*AX_109331000*), and 7B (*AX_186165710*) coincided with 5A (383.4–385.0 Mb), 6D (388.2–407.5 Mb), and 7B (701.1 Mb) identified in this study. Besides, the UIL locus (6D: 388.2–407.5 Mb) is about one LD from the QTL associated with both PH and the third internode length (Cui et al., 2011); they are likely to be the same. Except for the three loci talked above, the remaining four loci (2B: 409.9–439.5 Mb; 6A: 606.6–611.7 Mb; 7A: 496.3–512.9 Mb, and 7D: 8.3 Mb) are all likely to be novel.

Until today, a few studies have been conducted on plant morphology for common wheat. Thus, it is difficult to compare with the present results. Li et al. (2002) centered to the Gli-A2 gliadin locus and associated with a QTL affecting prostrate growth trait on chromosome 6A and nearly with the locus identified in this study in 6A (23.8–24.4 Mb) according to Maccaferri et al., 2015. Liu et al. (2016) have identified a region associated with plant morphology on the 4A chromosome, which maybe coincided with the locus identified in our study (4A: 41.4–46.2 Mb). Although some studies have been focused on the plant habit growth in bread wheat, no marker information was available in order to confirm our results. Thus, we think the loci 2B (91–92.6 Mb, 663.1–664.2 Mb), 3D (602.8–606.7 Mb), 5A (688.9 Mb), and 5B (696.7–697 Mb) were potential be novel.

Among the 30 loci for morphological traits, 13 loci talked above (1A: 572.9–577.7 Mb, 2A: 554.6–560.2 Mb, 2B: 652.7–656.0 Mb, 3B: 121.4–129.6 Mb, 7D: 5.4–10.5 Mb, 1A:

296.9–297.7 Mb, 5A: 290.1–299.5 Mb, 5B: 615.7–617.9 Mb, 5A: 383.4–385.0 Mb, 6D: 388.2–407.5 Mb, 7B: 701.1–701.1 Mb, 4A: 41.4–46.2 Mb, and 6A: 23.8–24.4 Mb) should be the same as the QTL reported in previous studies, whereas the remaining 17 are likely to be new. The stable loci validated by our studies and previous studies indicated that they are widespread and maybe stable in various varieties.

Candidate gene analysis

To identify candidate genes for morphological traits, the flanking sequences of SNP markers (in the LD decay interval and corresponding non-synonymous SNPs in the coding region of the genes) significantly associated with morphological traits were used as queries to BLAST against the NCBI. Totally, seven candidate genes were identified. For loci 1A (296.9–297.7 Mb) and 2B (652.7–656.0 Mb), the candidate genes for the E3 ubiquitin-protein ligase-like protein (*TraesCS1A01G164400* and *TraesCS2B01G466600*) were identified in the LD decay distance. E3 ubiquitin-protein ligase-like protein plays a crucial role in plant growth and development (Karki et al., 2021; Zhang et al., 2022). For loci 4A (41.4–46.2 Mb), candidate gene *TraesCS4A01G055000* encoded leucine-rich repeat receptor-like protein kinase family protein, which may trigger multiple physiological pathways (Coleman et al., 2021). For the loci on chromosome 5A (383.4–385.0 Mb and 595.4–597.7 Mb), candidate genes (*TraesCS5A01G182500* and *TraesCS5A01G406800*) for F-box proteins were identified. F-box proteins play crucial roles in cell-cycle progression, transcriptional regulation, flower formation, signal transduction, and many other cellular processes (El-Sharkawy et al., 2021; Guérin et al., 2021). Of these, a cytokinin riboside 5′-monophosphate phosphoribohydrolase (*TraesCS1A01G362500*) was identified in the LD decay of the loci on chromosome 1A (556.0–587.4 Mb). The cytokinin is a positive regulator of shoot growth (Chen et al., 2021) and response to biotic and abiotic stressors (Ellis et al., 2005; Chen et al., 2022). Besides, the gene *TraesCS2B01G123800* was identified as the candidate gene for the loci 2B (91.0–92.6 Mb) and encodes the C2H2 zinc finger, which is important for determining the prostrate/erect plant morphology (Li et al., 2021). The expressions of seven candidate genes showed that *TraesCS2B01G466600* and *TraesCS5A01G406800* showed no significant differences between the extreme accessions; *TraesCS1A01G164400*, *TraesCS4A01G055000,* and *TraesCS1A01G362500* showed 1.6–3.2-fold higher expressions among the extreme accessions; whereas *TraesCS5A01G182500* and *TraesCS2B01G123800* showed more than 1.9- to 7.6-fold lower expressions among extreme accessions. Thus, *TraesCS1A01G164400*, *TraesCS4A01G055000*, *TraesCS1A01G362500*, *TraesCS5A01G182500,* and *TraesCS2B01G123800* are the candidate genes in our study.

Potential implications in wheat breeding

Although conventional breeding has led to improved morphological traits on wheat, selection is time-consuming and not very efficient (He et al., 2010; Rasheed et al., 2016). The SNPs associated with morphological traits identified in this study should facilitate the progresses of MAS and pyramiding favorable alleles will improve morphological traits (Li et al., 2018). Accessions with superior morphological traits and high numbers of favorable alleles (such as Hechun 5, Jichun 158, LongFu 18-171, LongFu 17-5277-12-24 and Longchun 1 with longer FLL; Xinshuguang 4, Dongnong 156597, Kechun 151350, Kenda 163672 and Gang 09-558 with wider FLW; Longmai 17, Kehua, Xiaomaixiaobing 4, Xiaomaixiaobing seven and Yongjie with longer UIL; Kehan 1, Shen 68-71, Mailiduo, Jichun101 and Jichun1 with erect, whereas Beimai6, Lianfeng, Kechun 140865, Beimai one and LongFumai8 with prostrate) could be used as parental lines for the improvement of morphological traits in wheat breeding. The loci with pleiotropic and consistent effects across each environment in this study should be amenable to MAS.

Conclusion

We have identified 30 loci for morphological traits in spring wheat accessions by GWAS. Of these, 17 loci were likely to be new. Besides, five candidate genes were identified for morphological traits. The associated markers and varieties with favorable alleles could be used to accelerate the progresses of wheat MAS breeding.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

WL and YL designed the research. YL and YS analyzed the physiology data. WL and YL drafted the article. WL, JT, JC, SY, XW, and RZ performed the experiment. HZ and YL revised the article. All authors have read, edited, and approved the current version of the article.

Funding

This work was funded by the “Crop Varietal Improvement and Insect Pests Control by Nuclear Radiation.”

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.2022.934757/full#supplementary-material

SUPPLEMENTARY FIGURE S1
The qRT-PCR results of the seven candidate genes for the extreme accessions of the 25 spring wheat accessions from the Northeast of China. The extreme accessions of the 25 spring wheat accessions from the Northeast of China were according to Supplementary Table S2.

References

Beyer, S., Daba, S., Tyagi, P., Bockelman, H., Brown-Guedira, G., and Mohammadi, M. (2019). Loci and candidate genes controlling root traits in wheat seedlings-a wheat root GWAS. *Funct. Integr. Genomics* 19, 91–107. doi:10.1007/s10142-018-0630-z

Breseghello, F., and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172, 1165–1177. doi:10.1534/genetics.105.044586

Chen, L., Jameson, G. B., Guo, Y., Song, J., and Jameson, P. E. (2022). The LONELY GUY gene family: From mosses to wheat, the key to the formation of active cytokinins in plants. *Plant Biotechnol. J.* 20, 625–645. doi:10.1111/pbi.13783

Chen, L., Zhao, J., Song, J., and Jameson, P. E. (2021). Cytokinin glucosyl transferases, key regulators of cytokinin homeostasis, have potential value for wheat improvement. *Plant Biotechnol. J.* 19, 878–896. doi:10.1111/pbi.13595

Coleman, A. D., Maroschek, J., Raasch, L., Takken, F. L., Ranf, S., Hückelhoven, R., et al. (2021). The Arabidopsis leucine-rich repeat receptor-like kinase MIK2 is a crucial component of early immune responses to a fungal-derived elicitor. *New Phytol.* 229, 3453–3466. doi:10.1111/nph.17122

Cui, F., Li, J., Ding, A., Zhao, C., Wang, L., Wang, X., et al. (2011). Conditional QTL mapping for plant height with respect to the length of the spike and internode in two mapping populations of wheat. *Theor. Appl. Genet.* 122, 1517–1536. doi:10.1007/s00122-011-1551-6

Cui, F., Zhao, C., Ding, A., Li, J., Wang, L., Li, X., et al. (2014). Construction of an integrative linkage map and QTL mapping of grain yield-related traits using three related wheat RIL populations. *Theor. Appl. Genet.* 127, 659–675. doi:10.1007/s00122-013-2249-8

El-Sharkawy, I., Ismail, A., Darwish, A., El Kayal, W., Subramanian, J., Sherif, S. M., et al. (2021). Functional characterization of a gibberellin F-box protein, PslSLY1, during plum fruit development. *J. Exp. Bot.* 72, 371–384. doi:10.1093/jxb/eraa438

Ellis, M. H., Rebetzke, G. J., Azanza, F., Richards, R. A., and Spielmeyer, W. (2005). Molecular mapping of gibberellin-responsive dwarfing genes in bread wheat. *Theor. Appl. Genet.* 111, 423–430. doi:10.1007/s00122-005-2008-6

Gao, F., Ma, D., Yin, G., Rasheed, A., Dong, Y., Xiao, Y., et al. (2017). Genetic progress in grain yield and physiological traits in Chinese wheat cultivars of Southern Yellow and Huai Valley since 1950. *Crop Sci.* 57, 760–773. doi:10.2135/cropsci2016.05.0362

Gao, F., Wen, W., Liu, J., Rasheed, A., Yin, G., Xia, X., et al. (2015). Genome-wide linkage mapping of QTL for yield components, plant height and yield-related physiological traits in the Chinese wheat cross Zhou 8425B/Chinese Spring. *Front. Plant Sci.* 6, 1099. doi:10.3389/fpls.2015.01099

Guérin, C., Mouzeyar, S., and Roche, J. (2021). The landscape of the genomic distribution and the expression of the F-Box genes unveil genome plasticity in hexaploid wheat during grain development and in response to heat and drought stress. *Int. J. Mol. Sci.* 22, 3111. doi:10.3390/ijms22063111

He, Z. H., Xia, X. C., and Bonjean, A. P. A. (2010). *Wheat improvement in China*. Mexico: CIMMYT, 51–68.

Karki, S. J., Reilly, A., Zhou, B., Mascarello, M., Burke, J., Doohan, F., et al. (2021). A small secreted protein from *Zymoseptoria tritici* interacts with a wheat E3 ubiquitin ligase to promote disease. *J. Exp. Bot.* 72, 733–746. doi:10.1093/jxb/eraa489

Li, F., Wen, W., He, Z., Liu, J., Jin, H., Cao, S., et al. (2018). Genome-wide linkage mapping of yield related traits in three Chinese bread wheat populations using high-density SNP markers. *Theor. Appl. Genet.* 131, 1903–1924. doi:10.1007/s00122-018-3122-6

Li, F., Wen, W., Liu, J., Zhang, Y., Cao, S., He, Z., et al. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *BMC Plant Biol.* 19, 168. doi:10.1186/s12870-019-1781-3

Li, W. L., Nelson, J. C., Chu, C. Y., Shi, L. H., Huang, S. H., Liu, D. J., et al. (2002). Chromosomal locations and genetic relationships of tiller and spike characters in wheat. *Euphytica* 125, 357–366. doi:10.1023/a:1016069809977

Li, Y., Sun, A., Wu, Q., Zou, X., Chen, F., Cai, R., et al. (2021). Comprehensive genomic survey, structural classification and expression analysis of C2H2-type zinc finger factor in wheat (*Triticum aestivum* L.) *BMC Plant Biol.* 21, 380. doi:10.1186/s12870-021-03016-3

Liu, J., He, Z., Wu, L., Bai, B., Wen, W., Xie, C., et al. (2016a). Genome-wide linkage mapping of QTL for black point reaction in bread wheat (*Triticum aestivum* L.) *Theor. Appl. Genet.* 129, 2179–2190. doi:10.1007/s00122-016-2766-3

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

Liu, J., He, Z., Rasheed, A., Wen, W., Yan, J., Zhang, P., et al. (2017). Genome-Wide Association Mapping of Black Point Reaction in Common Wheat (*Triticum aestivum* L.. *BMC Plant Bio.* 17, 1–12.

Liu, J., Rasheed, A., He, Z., Imtiaz, M., Arif, A., Mahmood, T., et al. (2019). Genome-Wide Variation Patterns Between Landraces and Cultivars Uncover Divergent Selection During Modern Wheat Breeding. *Theor. Appl. Genet.* 132, 2509–2523.

Lozada, D. N., Mason, R. E., Sarinelli, J. M., and Brown-Guedira, G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genet.* 20, 82. doi:10.1186/s12863-019-0785-1

Nadolska-Orczyk, A., Rajchel, I. K., Orczyk, W., and Gasparis, S. (2017). Major Genes Determining Yield-Related Traits in Wheat and Barley. *Theor. Appl. Genet.* 130, 1081–1098.

Pang, Y., Liu, C., Wang, D., Amand, P. S., Bernardo, A., Li, W., et al. (2020). High-resolution genome-wide association study identifies genomic regions and candidate genes for important agronomic traits in wheat. *Mol. Plant* 13, 1311–1327. doi:10.1016/j.molp.2020.07.008

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181. doi:10.1086/302959

Quan, X., Dong, L. J., Zhang, N., Xie, C., Li, H., Xia, X., et al. (2021). Genome-wide association study uncover the genetic architecture of salt tolerance-related traits in common wheat (*Triticum aestivum* L.) *Front. Genet.* 12, 663941. doi:10.3389/fgene.2021.663941

Rahimi, Y., Bihamta, M. R., Taleei, A., Alipour, H., and Ingvarsson, P. K. (2019). Genome-wide association study of agronomic traits in bread wheat reveals novel putative alleles for future breeding programs. *BMC Plant Biol.* 19 (1), 541. doi:10.1186/s12870-019-2165-4

Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., et al. (2016). Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. *Theor. Appl. Genet.* 129, 1843–1860. doi:10.1007/s00122-016-2743-x

Sun, C. W., Zhang, F. Y., Yan, X. F., Zhang, X. F., Dong, Z. D., Cui, D. Q., et al. (2017). Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. *Plant Biotechnol. J.* 15, 953–969. doi:10.1111/pbi.12690

Valluru, R., Reynolds, M. P., Davies, W. J., and Sukumaran, S. (2017). Phenotypic and genome-wide association analysis of spike ethylene in diverse wheat genotypes under heat stress. *New Phytol.* 214, 271–283. doi:10.1111/nph.14367

Wang, J., Wang, R., Mao, X., Zhang, J., Liu, Y., Xie, Q., et al. (2020). RING finger ubiquitin E3 ligase gene *TaSDIR1-4A* contributes to determination of grain size in common wheat. *J. Exp. Bot.* 71, 5377–5388. doi:10.1093/jxb/eraa271

Wang, S. C., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi:10.1111/pbi.12183

Wang, Y., Hou, J., Liu, H., Li, T., Wang, K., Hao, C., et al. (2019). *TaBT1*, affecting starch synthesis and thousand kernel weight, underwent strong selection during wheat improvement. *J. Exp. Bot.* 70, 1497–1511. doi:10.1093/jxb/erz032

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., and Longin, C. F. H. (2018). Phenotypic and Genetic Analysis of Spike and Kernel Characteristics in Wheat Reveals Long-Term Genetic Trends of Grain Yield Components. *Theor. Appl. Genet.* 131, 2071–2084.

Xu, X. P., Ping, H., Chuan, L. M., Liu, X. Y., Liu, Y. X., Zhang, J. J., et al. (2021). Regional distribution of wheat yield and chemical fertilizer requirements in China. *J. Integr. Agric.* 20, 2772–2780. doi:10.1016/s2095-3119(20)63338-x

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

Zhang, D., Zhang, X., Xu, W., Hu, T., Ma, J., Zhang, Y., et al. (2022). *TaGW2L*, a GW2-like RING finger E3 ligase, positively regulates heading date in common wheat (*Triticum aestivum* L.) *Crop J.* doi.org/doi:10.1016/j.cj.2021.12.002

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1, plantgenome2008.02.0089. doi:10.3835/plantgenome2008.02.0089

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fgene.2022.972557

G × EBLUP: A novel method for exploring genotype by environment interactions and genomic prediction

Hailiang Song[1†], Xue Wang[2†], Yi Guo[2] and Xiangdong Ding[2*]

[1]Beijing Key Laboratory of Fisheries Biotechnology, Fisheries Science Institute, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [2]Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture and Rural Affairs, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China

OPEN ACCESS

EDITED BY
Li Ma,
University of Maryland, United States

REVIEWED BY
Rostam Abdollahi,
Aviagen, United Kingdom
Hailan Liu,
Maize Research Institute of Sichuan Agricultural University, China

*CORRESPONDENCE
Xiangdong Ding,
xding@cau.edu.cn,
orcid.org/0000000226842551

†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 18 June 2022
ACCEPTED 01 August 2022
PUBLISHED 12 September 2022

CITATION
Song H, Wang X, Guo Y and Ding X (2022), G × EBLUP: A novel method for exploring genotype by environment interactions and genomic prediction. *Front. Genet.* 13:972557.
doi: 10.3389/fgene.2022.972557

Genotype by environment (G × E) interaction is fundamental in the biology of complex traits and diseases. However, most of the existing methods for genomic prediction tend to ignore G × E interaction (GEI). In this study, we proposed the genomic prediction method G × EBLUP by considering GEI. Meanwhile, G × EBLUP can also detect the genome-wide single nucleotide polymorphisms (SNPs) subject to GEI. Using comprehensive simulations and analysis of real data from pigs and maize, we showed that G × EBLUP achieved higher efficiency in mapping GEI SNPs and higher prediction accuracy than the existing methods, and its superiority was more obvious when the GEI variance was large. For pig and maize real data, compared with GBLUP, G × EBLUP showed improvement by 3% in the prediction accuracy for backfat thickness, while our findings indicated that the trait of days to 100 kg of pig was not affected by GEI and G × EBLUP did not improve the accuracy of genomic prediction for the trait. A significant advantage was observed for G × EBLUP in maize; the prediction accuracy was improved by ~5.0 and 7.7% for grain weight and water content, respectively. Furthermore, G × EBLUP was not influenced by the number of environment levels. It could determine a favourable environment using SNP Bayes factors for each environment, implying that it is a robust and useful method for market-specific animal and plant breeding. We proposed G × EBLUP, a novel method for the estimation of genomic breeding value by considering GEI. This method identified the genome-wide SNPs that were susceptible to GEI and yielded higher genomic prediction accuracies and lower mean squared error compared with the GBLUP method.

KEYWORDS

G × E interaction, snps, bayes factors, traits, genomic prediction

Introduction

Genomic selection (GS) (Meuwissen et al., 2001), which relies on linkage disequilibrium between single nucleotide polymorphisms (SNPs) and causative variants, has become a useful tool in animal (VanRaden et al., 2009) and plant (Zhong et al., 2009) breeding. However, GS analytical modelling usually assumes no G × E interaction (GEI) and opposes the true genetic architecture of complex traits. In fact, interaction is fundamental in biology, and there is growing interest in estimating breeding value by considering GEI and using genome-wide SNPs.

The current state-of-the-art methods for the estimation of genomic breeding value without considering GEI include GBLUP (VanRaden, 2008) and Bayes-Alphabet (e.g., Bayes A, Bayes B and Bayes C) (Habier et al., 2011). Multi-trait (Richard, 1996) and reaction norm models (Rebecka et al., 2002) are the two prevalent GEI-handling methods that are used for genomic evaluations. However, the multi-trait model could only capture GEI in a limited number of environments, and the computational demands of multi-trait models would increase rapidly with an increase in the number of environment levels (Song et al., 2020a). The reaction norm model captures only part of the GEI because it needs to accommodate a continuous range of environmental values and cannot select excellent individuals using the unique estimated breeding value in actual breeding (Jarquin et al., 2014; Song et al., 2020b).

To explore GEI, certain G × E interaction-affected methods have been proposed for detecting SNPs. Wang et al. (2019) proposed several methods (Bartlett, F-Killeen, L-mean and L-median) that can be used to infer GEI from variance quantitative trait locus (vQTL) analysis without requiring environmental factor measurements. Moore et al. (2019) proposed StructLMM, which is useful for studying interactions with hundreds of environment variables. Moreover, Kerin and Marchini. (2020) proposed LEMMA, which infers GEI using a Bayesian whole-genome regression model. However, in Wang's method, GEI-affected SNP detection is possible because of selection, epistasis and phantom vQTL, instead of only GEI (Wang et al., 2019). StructLMM and LEMMA do not currently enable accounting for relatedness, and these methods cannot be efficiently applied to livestock and plant breeding due to the close genetic relationships that widely exists between individuals (Kerin et al., 2020; Moore et al., 2019). Therefore, it is essential to develop new methods for the estimation of genomic breeding value by considering GEI.

In this study, we proposed a novel approach for the estimation of genomic breeding value by considering GEI, which can handle environment variables in different dimensions. The basic principle of the new approach was to first detect the genome-wide markers affected by GEI (G × EWAS), in which a score-test statistics was implemented to identify the significant GEI-associated SNPs. Next, all markers were classified into SNPs with/without GEI to construct the genomic relationship matrices separately and predict the genomic breeding value using the mixed model (G × EBLUP). For its general application, the efficiency of the proposed method was evaluated through simulation study and real data from pigs and maize.

Methods

Ethics statement

The animal study was reviewed and approved by Animal handling and sample collection were conducted according to protocols approved by the Institutional Animal Care and Use Committee (IACUC) at China Agricultural University. All authors strictly complied with the Regulations on the Administration of Laboratory Animals (Order No. 2 of the State Science and Technology Commission of the People's Republic of China, 1988). There was no use of human participants, data or tissues.

G × EWAS

G × EWAS extends the conventional linear mixed model by including an additional per-individual effect term accounting for G × E, which can be represented as N × 1 vector, β_{GxE}. The model was defined as follows:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{x}\boldsymbol{\beta}_{\boldsymbol{G}} + \boldsymbol{x} \odot \boldsymbol{\beta}_{\boldsymbol{GxE}} + \boldsymbol{u} + \boldsymbol{e} \qquad (1)$$

where $\mathbf{y}$ is the vector of observed phenotypic values; b is the vector of fixed effects; $\beta_{\mathbf{G}}$ is the average effect of gene substitution of a particular SNP; and $\mathbf{x}$ is the vector of the genotype indicator variable of the variant coded as 0, 1 or 2. $\mathrm{x} \odot \beta_{\mathrm{G\times E}} = \mathrm{diag}(\mathrm{x})\beta_{\mathrm{G\times E}}$, where $\odot$ denotes the element-wise (Hadamard) product and $\mathrm{diag}(\mathrm{x})$ denotes the $\mathrm{N} \times \mathrm{N}$diagonal matrix whose diagonal is $\boldsymbol{x}$.The per-individual effect size vector $\beta_{\mathrm{G\times E}}$ is defined as a random effect, following the multivariate normal distribution $\beta_{\mathrm{G\times E}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{G\times E}} \sum)$, where $\sigma^2_{\mathrm{G\times E}}$, is the variance and covariance matrix of the G × E effect, $\sum \in \mathbb{R}^{\mathrm{N\times N}}$ parameterises how per-individual effects covary across individuals and is calculated as a function of observed environment variables. $\sum \equiv \sum(\mathrm{E}) = \mathrm{EE}'$, where E is the N × L matrix of L observed environments. The linear covariance function (EE′) was primarily used because of two appealing properties. First, as the number of samples typically exceeds the number of environments in larger populations (L << N), a low-rank linear covariance is noted, which enables parameter inference with a computational complexity that scales linearly with the increasing population size. Second, a linear covariance is directly interpretable as there is one-to-one correspondence between G × EWAS and linear regression using L covariates to account for

GEI. Notably, for the special case of $\sigma^2_{G\times E} = 0$, the model G × EWAS reduces to a standard linear mixed model for genome-wide association study; thus, G × EWAS is a single-SNP regression model; u is the vector of random polygenic effects with a normal distribution $u \sim N(0, G\sigma^2_u)$, in which σ^2_u is the polygenic variance and G is the genomic relationship matrix. It was constructed using the markers according to VanRaden (2008); X is the incidence matrix linking **b** to **y**; **e** is the vector of random errors with normal distribution of N (0, **I** σ^2_e), where σ^2_e is the residual variance and I is the identity matrix. The analysis of G × EWAS was based only on the reference data to avoid the double counting of the SNP effect in genomic prediction.

For the parameter inference, we considered the marginalised form of the model in Eq. 1, which was obtained by integrating over the G × E effects $\beta_{G\times E}$ and the random effect component u **:**

$$\mathbf{y} \sim \mathbf{N}\left(\mathbf{Xb} + \mathbf{x}\boldsymbol{\beta}_{\mathbf{G}},\ \sigma^2_{\mathrm{GxE}}\mathrm{diag}(\mathbf{x})\mathbf{EE}^{\mathrm{T}}\mathrm{diag}(\mathbf{x}) + \sigma^2_{\mathrm{g}}\mathbf{G} + \sigma^2_{\mathrm{e}}\mathbf{I}\right) \qquad (2)$$

Using the marginalised model in Eq. 2, a G × E interaction test corresponds to the alternative hypothesis $\sigma^2_{G\times E} > 0$. We defined an efficient score-based test that enabled the *p*-value calculation with a complexity that scaled linearly with the number of individuals, provided that there is low-rank environment covariance $\sum$. The null model of the interaction test reduced to a standard linear mixed model with a low-rank covariance matrix for additive genetic effects, and the existing efficient inference strategies for the standard linear mixed model can be reused. The score-test statistics can be computed in an analogous manner according to the procedure described by Wu et al. (2011):

$$\begin{aligned}\mathbf{Q} &= \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{PK}_1\mathbf{Py} = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{P}\left(\mathrm{diag}(\mathbf{x})\mathbf{EE}^{\mathrm{T}}\mathrm{diag}(\mathbf{x})\right)\mathbf{Py}\\ &= \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{P}\left(\mathrm{diag}(\mathbf{x})\mathbf{E}\right)\left(\mathrm{diag}(\mathbf{x})\mathbf{E}\right)'\mathbf{Py} = \frac{1}{2}\left\|\mathbf{W}^{\mathrm{T}}\mathbf{Py}\right\|^2\end{aligned} \qquad (3)$$

Where

$$\mathbf{K}_1 = \mathrm{diag}(\mathbf{x})\sum\mathrm{diag}(\mathbf{x})$$

$$\mathrm{W} = \mathrm{diag}(\mathbf{x})\mathbf{E}$$

$$\mathbf{P} = \mathbf{H}_0^{-1} - \mathbf{H}_0^{-1}[\mathbf{X},\mathbf{x}]\left([\mathbf{X},\mathbf{x}]^{\mathrm{T}}\mathbf{H}_0^{-1}[\mathbf{X},\mathbf{x}]\right)^{-1}[\mathbf{X},\mathbf{x}]^{\mathrm{T}}\mathbf{H}_0^{-1}$$

The matrix H_0 denotes the total covariance matrix estimated under the null model $H_0 = \sigma^2_g G + \sigma^2_e I$ **Q** follows a mixture of χ^2 distributions (Wu et al., 2011; Lippert et al., 2014): $Q \sim \sum_k a_k\chi^2_1$, where the vector of the coefficients $a = [a_k]_k$ can be computed as the eigenvalues of $P^{\frac{T}{2}}K_1P^{\frac{1}{2}}$. According to the procedure in SKAT (Wu et al., 2011), as the distribution of the score-test statistics was a mixture of χ^2, the *p*-values were computed using the Davies method (Davies, 1980). Alternatively, the Liu method (Huan et al., 2008) was employed when the Davies method failed to converge.

The evidences for individual environment variables or environment sets for driving the observed G × E effects can be assessed by comparing the model log marginal likelihoods between models with and without including these environments. The Bayes factors (BF) obtained from such comparisons is directly calibrated as the parameter number fitted using maximum likelihood and is independent of the environment variable numbers.

Given a variant and set of L environment $L = (e_1, e_2, e_3, \ldots, e_L)$,

$$(Log(BF) = LML(L) - LML(L_i)) \qquad (4)$$

where LML(L) and LML(L_i) represent the marginal log-likelihood of the model described in Eq. 2, either considering the full or reduced environment sets to define the G × E environment covariance, respectively. log(BF) < 0 indicates the lack of contribution of the environmental impact on G × E interaction, whereas log(BF) > 3 indicates strong G × E environment interaction (Kass and Raftery., 1995).

G × EBLUP

The G × EBLUP model includes additive genetic and GEI effects. The model is as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu}_{\mathbf{G\times E}} + \mathbf{Zu} + \mathbf{e} \qquad (5)$$

where **y**, X, b and e denote the same parameters as in the G × EWAS model, $u_{G\times E}$ is the vector of genomic values captured by genetic markers associated with GEI, following a normal distribution of $N(0, G_{G\times E}\sigma^2_{G\times E})$; **u** is the vector of genomic values captured by the remaining genetic marker sets (SNPs that are not significantly associated with GEI), following a normal distribution $N(0,\ G_{\boldsymbol{u}}\sigma^2_u)$ and Z is an incidence matrix that links $u_{G\times E}$ and u to **y**. Matrices $G_{G\times E}$ and G_u were constructed similarly as G; the former was constructed using only the genetic marker set defined by GEI, as described below, and the latter was constructed using the remaining markers. $\sigma^2_{G\times E}$ and σ^2_u are the variance components explained by the variants with and without GEI, respectively. When an SNP was significant the GEI with phenotypes based on the prespecified significance cutoff level (E01-E05), showing the SNP was considered to impact the GEI.

Data simulation

So far, only few genomic data simulating softwares considering GEI have been available. In this study, we proposed a reaction norm model accounting for heterogeneous residual variances to simulate phenotypic and environmental values.

$$\mathbf{y} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 * \boldsymbol{c} + \boldsymbol{e}_0 + \boldsymbol{e}_1 * \boldsymbol{c}$$

where $\mathbf{y}$ is the vector of phenotypic value, $\mathbf{c}$ is the vector of environmental value; $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are the random additive genetic effects for the intercept and slope, respectively; and e_0 and $\mathbf{e}_1$ are the random residual effects for the intercept and slope, respectively.

The environmental value $\boldsymbol{c}$ is further divided into two components:

$$\mathbf{c} = \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where β is the vector of the random genetic effect and $\boldsymbol{\varepsilon}$ is the vector of the random residual effect.

We assumed that $\boldsymbol{\alpha}_0$, β and $\boldsymbol{\alpha}_1$ are affected by all QTLs simultaneously, and these three effects of each QTL are drawn from a multivariate normal distribution with the vector of means 0 and the variance–covariance structure $\begin{bmatrix} \sigma^2_{\alpha_0} & \sigma_{\alpha_0\beta} & \sigma_{\alpha_0\alpha_1} \\ \sigma_{\alpha_0\beta} & \sigma^2_{\beta} & \sigma_{\beta\alpha_1} \\ \sigma_{\alpha_0\alpha_1} & \sigma_{\beta\alpha_1} & \sigma^2_{\alpha_1} \end{bmatrix}$. The genetic variance of each QTL is computed using $2\ p_i\,(1 - p_i)m_i$, where $p_i$is the frequency of one allele of ith QTL, m_i is the effect of the ith QTL for $\boldsymbol{\alpha}_0$, β or $\boldsymbol{\alpha}_1$. Then, the substitution effects are rescaled to ensure the total variances $\sigma^2_{\alpha_0}$, σ^2_{β} and $\sigma^2_{\alpha_1}$ for $\boldsymbol{\alpha}_0$, β and $\boldsymbol{\alpha}_1$, respectively. The $\sigma_{\alpha_0\beta}$, $\sigma_{\alpha_0\alpha_1}$ and $\sigma_{\beta\alpha_1}$ are re-calculated using the scaled substitution effects of QTL. The $\boldsymbol{e}_0$, $\boldsymbol{e}_1$ and $\boldsymbol{\varepsilon}$ values of each individual are sampled from a multivariate normal distribution with the vector of means 0 and the variance–covariance structure $\begin{bmatrix} \sigma^2_{e_0} & \sigma_{e_0e_1} & \sigma_{e_0\varepsilon} \\ \sigma_{e_0e_1} & \sigma^2_{e_1} & \sigma_{e_1\varepsilon} \\ \sigma_{e_0\varepsilon} & \sigma_{e_1\varepsilon} & \sigma^2_{\varepsilon} \end{bmatrix}$.

For the G × E interaction simulations, the parameter $\sigma^2_{\alpha_1}$ was set to control the extent of the G × E interaction, whereas other parameters ($\sigma^2_{\alpha_0}$, σ^2_{β}, $\sigma_{\alpha_0\beta}$, $\sigma_{\alpha_0\alpha_1}$, $\sigma^2_{e_0}$, $\sigma_{\beta\alpha_1}$, $\sigma^2_{e_1}$, σ^2_{ϵ} , $\sigma_{e_0e_1}$, $\sigma_{e_0\epsilon}$ and $\sigma_{e_1\epsilon}$) were fixed. The pseudo true breeding values (TBVs) of an individual for $\boldsymbol{\alpha}_0$, β and $\boldsymbol{\alpha}_1$ are its QTL effects multiplied by genotypes, followed by the scaling of the means of the pseudo TBVs to 0. Finally, the environmental valuec of each individual is obtained by adding the cumulative effect across all QTLs for β with the residual $\boldsymbol{\epsilon}$, followed by the generation of the phenotype $\mathbf{y}$ of each individual through the model $y = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 {}^* \boldsymbol{c} + \boldsymbol{e}_0$, without accounting for heterogeneous residual variance. For the simulated data, $\boldsymbol{c}$ was used as the environment variable $\mathbf{E}$, as described in formula (1). The real genotypes of 7,334 individuals determined using the Illumina BovineSNP50 BeadChip from the Chinese Holstein population were referred for phenotype and environment simulation, and 45,323 SNPs remained after imputation of missing genotypes and removal of SNPs with a minor allele frequency (MAF) of <0.01. Additional File 3 Supplementary Figure S1 presents the heat map of the genomic relationship matrix of 7,334 Chinese Holsteins. Three simulated datasets with GEI effect variances of 0.25, 1 and 2 were obtained, and the corresponding phenotypic variances were 2.25, 3 and 4, respectively. Additionally, when the GEI effect variance was 0.25, the datasets 2 and 3 covariate environments were simulated and compared. For each dataset, 306 SNPs that affected the trait of interest were simulated and referred to as simulated QTLs in this study. For each scenario, the simulation was repeated 20 times. We used the DMU software (Madsen et al., 2006) to estimate the variances of the additive effect, GEI effect and residual using the reaction norm model for each replicate. As shown in Additional File 2 Supplementary Table S1, these estimated values were close to the assigned values. Moreover, a dataset was randomly selected from the 20 repeated datasets, and 6 SNPs with GEI were randomly selected from this dataset, showing that the phenotypic variation was largely affected by GEI and that it was relatively small in the scenarios with no GEI (Additional File 3: Supplementary Figures S2–S6), thus implying that the simulation fitted well. All analyses with G × EBLUP and GBLUP models and simulation were conducted using in-house scripts written in Python3.8 by the first author.

Real data

Pig data

Yorkshire pigs were sampled from a breeding company with five breeding farms distributed across China (Additional File 3: Supplementary Figure S7). Different farms displayed distinct climates, housing systems, nutritional regimes, disease pressures and stocking densities, potentially leading to GEI. Table 1 presents the phenotype data. We examined two growth traits ‘days to 100 kg (AGE)’ and ‘backfat thickness adjusted to 100 kg (BFT)’. Genotyping was performed using the PorcineSNP80 BeadChip (Illumina, CA, USA), which included 68,528 SNPs across the entire pig genome. A total of 1,778 animals born between 2011 and 2016 were genotyped (Table 1).

Pig data

Maize is one of the most important crops worldwide. It provides food for humans and animals; it is a raw material for industrial processes and a model plant for understanding evolution, domestication and heterosis (Romay et al., 2013). Thus, maize data from 11 regions across China (Additional File 3: Supplementary Figure S7) were obtained to verify G × EBLUP, with the regions used as environment variables. Because of varying conditions of light, temperature, air, water and soil in different regions, under which GEI could show its effect, region effect was considered as an environmental covariate in the present study. A total of 681 maize lines were collected and each line had phenotype records of two traits, grain weight (GW) and water content (WC), in 1–8 environments. Overall, 2,676 observations were collected for the two traits. Table 1 lists the detailed information on GW and WC. Meanwhile, all lines were genotyped using the customised SNP panel of 61,224 markers across the maize genome.

For real pig and maize data, Beagle 4.1 (Browning and Browning, 2009) was used for the imputation of the missing

TABLE 1 Descriptive statistics for pig and maize population traits.

Population	Trait[a]	N-obs[b]	Genotyped individuals	N-env[c]	Mean	SD	Min	Max
Pig	AGE (day)	28,827	1778	5	170.8	13.9	124.0	211.0
	BFT (mm)	28,827	1778	5	11.8	2.4	5.0	30.7
Maize	GW (kg)	2676	681	11	6.75	1.39	0.407	11.24
	WC (%)	2676	681	11	26.89	4.58	14.80	47.80

[a] AGE: days to 100 kg; BFT: backfat thickness adjusted to 100 kg; GW: grain weight; WC: water content.

[b] N-obs: number of observations.

[c] N-env: number of environments.

SNP genotypes, and only loci on autosomes were used for further analysis. PLINK software (v1.90) (Chang et al., 2015) was implemented for quality control. We excluded SNPs with a MAF of <0.05, call rate of <0.90, or those severely deviating from the Hardy–Weinberg equilibrium (HWE) ($p < 10^{-7}$). Similarly, we excluded the pig individuals or maize samples with a call rate of <0.90. Finally, 56,463 and 59,401 SNPs were present in the pig and maize data, respectively, and all genotyped pigs and maize were retained.

Method application

Application to simulated data

Simulated data analysis was performed using G × EWAS proposed in this study to identify markers associated with GEI. We used Bonferroni correction at a significance level of 0.05 to identify significant SNPs. We implemented the four methods (L-median, L-mean, Bartlett and F-Killeen) proposed by Wang et al. (2019) in addition to StructLMM proposed by Moore et al. (2019) to identify SNPs affected by GEI. Based on the G × EWAS results, we performed genomic prediction on each simulated dataset using G × EBLUP. To investigate more SNPs with GEI, *p*-value gradients of 1-E01–1-E05 were chosen as threshold standards to select the SNPs associated with GEI. Five 10-fold cross-validation (10-CV) repetitions were used to assess the genomic prediction using G × EBLUP. In each cross-validation, the reference and validation populations comprised 6,601 and 733 individuals, respectively. The accuracy of genomic prediction was calculated as the Pearson's correlation between original phenotypes Phe and the genomic estimated breeding values (GEBVs) of all validation individuals *r*(Phe, GEBV). Moreover, the mean squared error (MSE) of the prediction ability matrix was used to evaluate the performance of the models; MSE was computed as the average square of the difference between Phe and GEBV centred on zero. In each scenario, we performed the comparisons between G × EBLUP and GBLUP (VanRaden, 2008) at different GEI variances (0.25, 1 and 2) and different number of environment variables (1, 2 and 3).

TABLE 2 Significant G × E interaction single nucleotide polymorphisms (SNPs) detected on simulated data using the proposed G × EWAS method and the five approaches, StructLMM, Bartlett, F-Killeen, L-mean and L-median under different variance of G × E interactions. The SNP numbers overlapping with simulated genotype–environment interaction quantitative trait locus (306) are in parentheses.

Variance of G × E interactions	0.25	1	2
G × EWAS	2435 (43)	5081 (64)	3981 (77)
StructLMM	2472 (41)	5092(62)	4084 (77)
Bartlett	507 (9)	3495 (33)	3976 (54)
F-killeen	101 (6)	144 (2)	109 (3)
L-mean	224 (6)	1037 (8)	606 (14)
L-median	188 (6)	808 (8)	439 (9)

Application to real data

Pig data were used to detect the SNPs affected by GEI. The herd-year-season effects, estimated using the conventional pedigree-based BLUP method, were used as environmental covariates, and the corrected phenotype were used as response variables. The calculations for the corrected phenotype value followed the method described by Song et al. (2017). Overall, 207 young and the remaining 1,571 individuals were considered as the validation and reference populations, respectively. For the maize data, an environment was randomly selected for each line to ensure that all lines could be used for analysis. Accordingly, we used 681 lines for each analysis and performed five replications of 5-fold cross-validation.

For the G × EBLUP method, the *p*-values for all markers were calculated using G × EWAS, and a threshold standard with *p*-value gradients of 1-E01–1-E05 was selected to screen the GEI-associated SNPs. The GBLUP and G × EBLUP methods were then used to estimate GEBVs. The genomic prediction accuracy on the real data was evaluated differently than that on the simulated data, using the correlation between GEBVs and the phenotypic values y (maize data) or corrected phenotypes y_c (pig data) in the validation population. MSE was computed as the average square of the difference between y or y_c and GEBVs centred on zero. The BF for each environment variable was

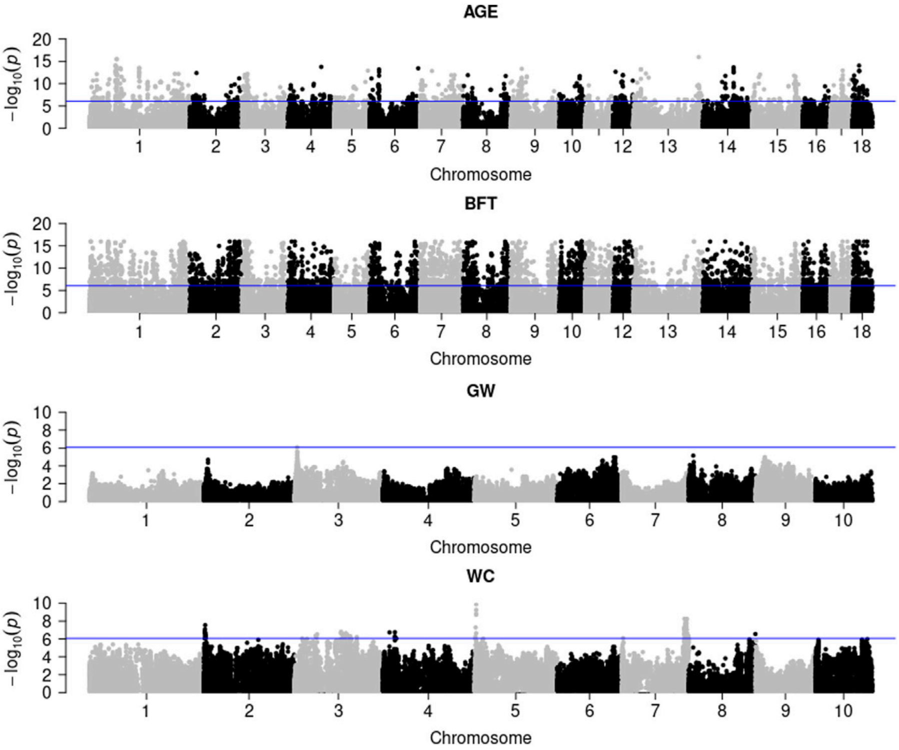


FIGURE 1
G × E marker genome-wide association analysis of pig and maize population traits. AGE: days to 100 kg; BFT: backfat thickness adjusted to 100 kg; GW: grain weight; WC: water content.

calculated to obtain the sensitive environments of markers associated with GEI. For simulated data and real data, the improvement in prediction accuracy for G × EBLUP over GBLUP was calculated by subtracting the prediction accuracy obtained by GBLUP from the prediction accuracy obtained by G × EBLUP and then dividing by the prediction accuracy obtained by GBLUP.

Results

Genome-wide G × E association analysis

Simulated data

Table 2 indicates that G × EWAS performed better than the other methods. When the GEI variance was 0.25, G × EWAS detected 2,435 significant SNPs with Bonferroni correction (0.05/45,293). Of these SNPs, 43 overlapped with the 306 simulated GEI SNPs. Although a low number of SNPs (41) overlapped with the simulated GEI SNPs, StructLMM detected a high number of significant SNPs (2,472). Similarly, G × EWAS detected a higher number of significant SNPs than those detected using the Bartlett, F-Killeen, L-mean and L-median methods, which identified 507, 101, 224 and 188 significant SNPs, respectively. Further, 9, 6, 6 and 6 SNPs overlapped with the simulated GEI QTLs, respectively. A similar trend was also observed in the scenario where GEI variance increased to 1 or 2. Larger GEI effect variances led to the identification of a higher number of real QTLs with GEI (with the exception of the F-Killeen method), as shown in Table 2 and Additional File 3 Supplementary Figures S8–S10.

Pig and maize data

Figure 1 illustrates the genome-wide G × E marker mapping on pig and maize data. Figure 1 shows that a total of 1,164 and 5,448 significant SNPs were detected in pigs for AGE and BFT traits with Bonferroni correction (0.05/56,445), respectively. For maize data, only 1 and 84 genome-wide significant SNPs were detected for the two traits, GW and WC, respectively. Additionally, the genotypic values and SNP effect values of the top most significant SNPs for each trait at different environments in pig and maize populations indicated that BFT, GW and WC were affected by the environment; however, no GEI was detected on AGE (Additional File 3: Supplementary Figures S11, S12).

TABLE 3 Genomic prediction accuracies and mean squared error (MSE) for GBLUP and G × EBLUP method under different G × E interaction *p*-values (E01~E05).

Data set	Trait	Content	GBLUP	P-value[a]				
				E01	E02	E03	E04	E05
Simulation[b]	One[c]	SNP number	45,323	23,517	14,210	8844	5543	2186
		Accuracy	0.737	0.735	0.739	0.749	0.738	0.712
		MSE	1.818	1.820	1.810	1.715	1.813	1.894
Pig	AGE	SNP number	56,445	27,762	14,117	7420	3964	2117
		Accuracy	0.225	0.223	0.226	0.226	0.226	0.224
		MSE	179.81	179.918	179.722	179.703	179.677	179.797
	BFT	SNP number	56,445	37,448	25,242	17,110	11,801	8098
		Accuracy	0.268	0.275	0.276	0.272	0.269	0.268
		MSE	2.707	2.693	2.69	2.699	2.704	2.706
Maize[b]	GW	SNP number	59,401	17,285	4279	834	421	143
		Accuracy	0.288	0.290	0.306	0.294	0.271	0.269
		MSE	46.323	46.317	46.174	46.178	46.223	46.347
	WC	SNP number	59,401	28,636	12,123	4168	2132	875
		Accuracy	0.295	0.301	0.315	0.318	0.293	0.273
		MSE	721.588	721.229	721.129	720.854	721.590	722.009

[a]Cut-off p-values for G × E interaction single nucleotide polymorphisms on G × E.
[b]One randomly selected replicate.
[c]The variance of G × E interactions was 0.25.

Accuracy and mean squared error of genomic prediction

Simulated data

The significance level was set at the *p*-value gradient of 1-E01–1-E05 to determine SNPs associated with GEI. Table 3 presents the number of SNPs affected by GEI. These SNPs along with the corresponding remaining SNPs were used in G × EBLUP. All 45,323 qualified SNPs were used in GBLUP. The genomic prediction accuracy and MSE of the simulated data were determined from a randomly selected replicate. Table 3 shows that the G × EBLUP accuracy was different under different *p*-values. G × EBLUP showed the highest genomic prediction accuracy and the lowest MSE with a *p*-value of 1-E03. Further, the prediction accuracy of G × EBLUP improved by 1.7% compared with that of GBLUP. Figure 2 shows the averaged accuracies and MSE of genomic prediction obtained using G × EBLUP and GBLUP under different GEI variances. For G × EBLUP, the average prediction accuracy was calculated by selecting the highest prediction accuracy values under different *p*-values in each repetition. When the GEI variance was 0.25, the G × EBLUP yielded 1.7% higher prediction accuracies than GBLUP. Moreover, G × EBLUP yielded lower MSE than GBLUP, with average MSE values of 1.816 and 1.942, respectively. G × EBLUP performed significantly better than GBLUP when the GEI variance was increased to 1 and 2, yielding 3.9 and 6.4% higher prediction accuracies, respectively, and a lower MSE than GBLUP.

Figure 2 also shows how the number of the environment variables influences the genomic prediction. In the scenario with a GEI variance of 0.25, when the number of environment variables was 1, 2 and 3, no significant differences were noted in the prediction accuracy and MSE among the number of different environment variables for G × EBLUP. Additionally, in all scenarios, G × EBLUP yielded 1.8% higher accuracies and a 12.4% lower MSE than GBLUP, confirming that G × EBLUP performed better.

Pig and maize data

Figure 3 and Table 3 present the accuracy and MSE of genomic prediction on pig and maize data. According to the results of G × EWAS regarding AGE trait in pigs, 27,762; 14,117; 7,420; 3964 and 2,117 SNPs were selected as G × E markers in G × EBLUP, and the accuracies of G × EBLUP under different G × E markers (1-E01–1-E05) were not significantly different. The average prediction accuracy obtained using G × EBLUP was 0.225, which was same as that obtained using GBLUP. Moreover, no differences were noted in the MSE between G × EBLUP and GBLUP. These findings indicated that AGE was not affected by GEI. However, for BFT, G × EBLUP showed the best performance at the *p*-value of 1-E02, the prediction accuracy was improved by 3% compared with that of GBLUP, and the MSE was reduced by 0.107, decreasing from 2.707 to 2.600.

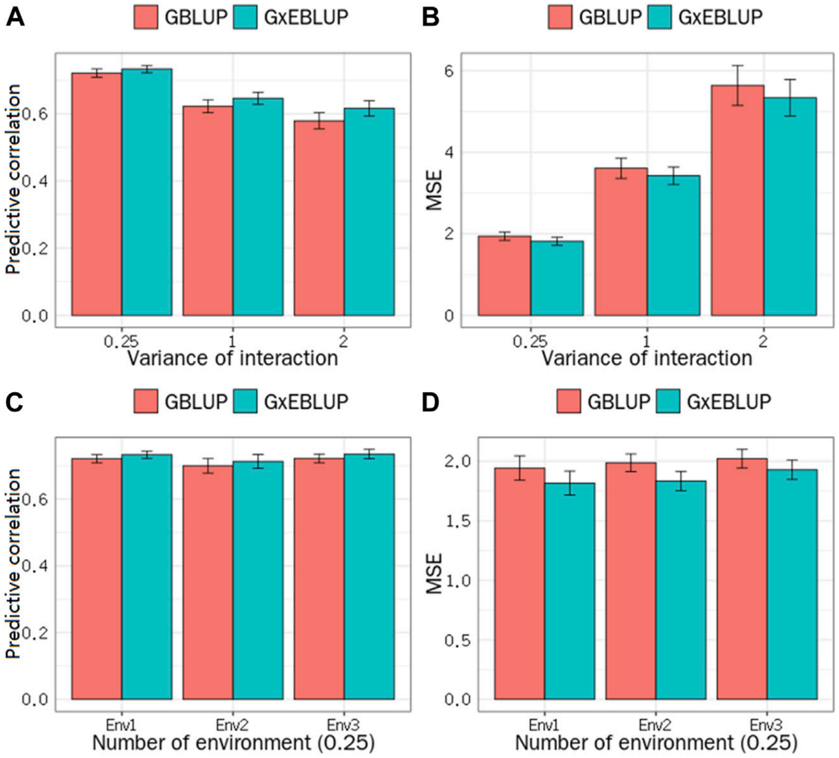


FIGURE 2
GBLUP and G × EBLUP method for the **(A)** accuracy and **(B)** mean squared error (MSE) of genomic prediction under different G × E interaction variances. Genomic prediction **(C)** accuracy and **(D)** MSE under different numbers of environment variables.

Compared with pig data, the genomic prediction of G × EBLUP showed considerable improvement in maize data. As shown in Table 3, for the trait GW from a randomly selected replicate, G × EBLUP showed the best performance at the *p*-value of 1-E02, the genomic prediction accuracy was improved by 6.25% compared with that of GBLUP, and MSE was reduced from 46.323 to 46.174. For the trait WC from a randomly selected replicate, the highest prediction accuracy of G × EBLUP was obtained at the *p*-value of 1-E03, the prediction accuracy was improved by 7.8% compared with that of GBLUP, and MSE was reduced from 721.588 to 720.854. The average prediction accuracy of 5 repetitions of 10-fold CV for GBLUP and G × EBLUP are shown in Figure 3, G × EBLUP showed approximately 5.0 and 7.7% improvement in prediction accuracy for GW and WC in maize population, respectively. Moreover, G × EBLUP also showed a lower prediction MSE than GBLUP, which further verified the advantages of G × EBLUP.

Sensitive environment detection in real data

The BF for each environmental factor was obtained, and then the sensitive environments were assessed accordingly. In pig population, for AGE and BFT, 10 SNPs each with the smallest *p*-values were selected for sensitive environmental detection. As shown in Figure 4, for AGE, the top 10 SNPs regarding season showed the largest BF, indicating that the most sensitive environmental factor for the top 10 SNPs was season. Similarly, the least sensitive environmental factors for SNPs were farm and year, as the values of BF of farm and year were equally low. For BFT, the most sensitive environmental factors for all SNPs were farm and season, and their BF values were also same; year was the least sensitive environmental factor (Figure 4). In maize population, as shown in Figure 4, the averaged log BF values indicated that the region environment variable has a strong GEI (Log(BF) > 3) with WC and GW. Additionally, the BF values of WC were higher than that of GW, which was also consistent with the superiority of genomic prediction of WC (Figure 3).

Computing time

The average computation time for G × EBLUP and GBLUP to complete each fold of CV is presented in Table 4. Running time of the methods was measured in minutes on an HP server

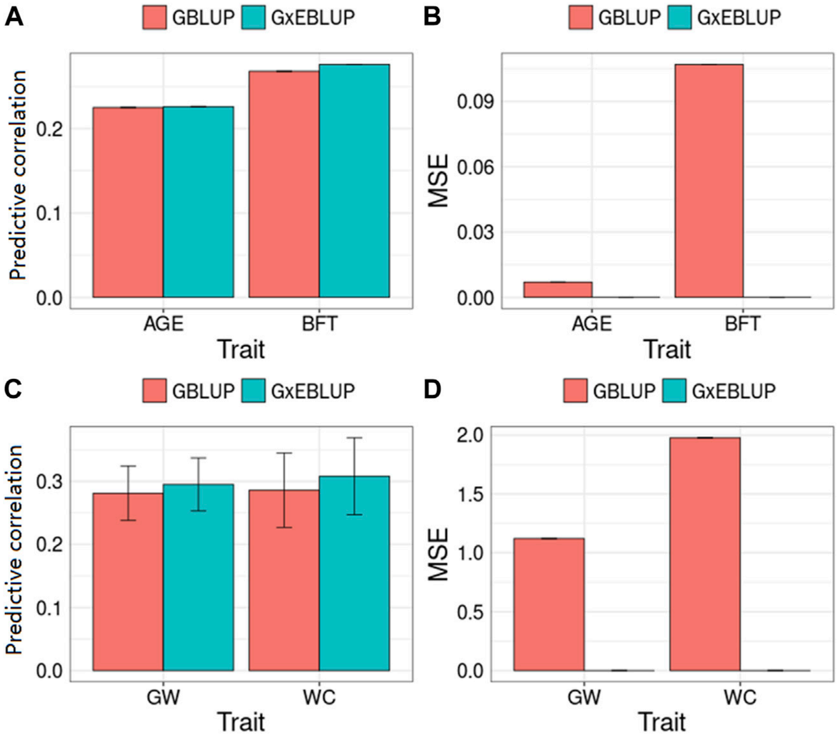


FIGURE 3
GBLUP and G × EBLUP method for the **(A,C)** accuracy and **(B,D)** mean squared error (MSE) of genomic prediction in pigs and maize. AGE: days to 100 kg; BFT: backfat thickness adjusted to 100 kg; GW: grain weight; WC: water content. MSE is a relative value by assuming that the MSE of GBLUP method is equal to 0 because of the large MSE value.

(CentOS Linux 7.9.2009, 2.5 GHz Intel Xeon processor and 515G total memory). In all scenarios, G × EBLUP runs longer than GBLUP, mainly because running G × EBLUP requires concurrently running G × EWAS, which is a single marker regression model with a long running time, e.g. G × EWAS took an average of 45.8 min in each fold of CV to complete the analysis, requiring considerably less time than GBLUP (15.05 min). In addition, there is no obvious difference here between different traits in the same population due to the same size population and number of SNPs.

Discussion

G × E interactions play an important role in livestock and plants and should be considered in breeding programmes to select elite individuals in specific environments (Crossa, 2012; Heslot et al., 2014; Jarquin et al., 2017; Perez-Rodriguez et al., 2017; Tiezzi et al., 2017; Liu et al., 2019; Zhang et al., 2019; Braz et al., 2021). However, because of their complexity, G×E interactions are usually ignored in conventional breeding and the current widely used methods of estimating genomic breeding value, e.g. GBLUP (VanRaden, 2008), single-step GBLUP (Misztal et al., 2009), BayesA (Meuwissen et al., 2001), BayesCpi (Habier et al., 2011), which could lead to biases in the estimation of breeding values and selection decisions. Although the multi-trait (Richard, 1996) and reaction norm models (Rebecka et al., 2002) are two prevalent methods for handling GEI in the estimation of genomic breeding value, our previous studies have explicated the disadvantages of these two types of methods (Song et al., 2020b). Moreover, these two methods could not detect the markers associated with GEI. In this study, we proposed G × EBLUP, which is a novel method for genomic breeding value estimation that takes GEI into account. The core of G × EBLUP is the estimation of GEI using G × EWAS by including an additional per-individual effect term that accounts for GEI; it is also powerful for the identification of the SNPs that are susceptible to GEI. Comprehensive simulation studies and real data of pigs and maize have demonstrated the superiority of the proposed method.

In the simulated data, our results indicated the superiority of the G × EBLUP method for genomic prediction in all scenarios, which was more remarkable when the variance of GEI was large (Figures 2A,B), showing that the new method can appropriately handle GEI. Our results also showed that there was no significant difference in the prediction accuracy of G × EBLUP under

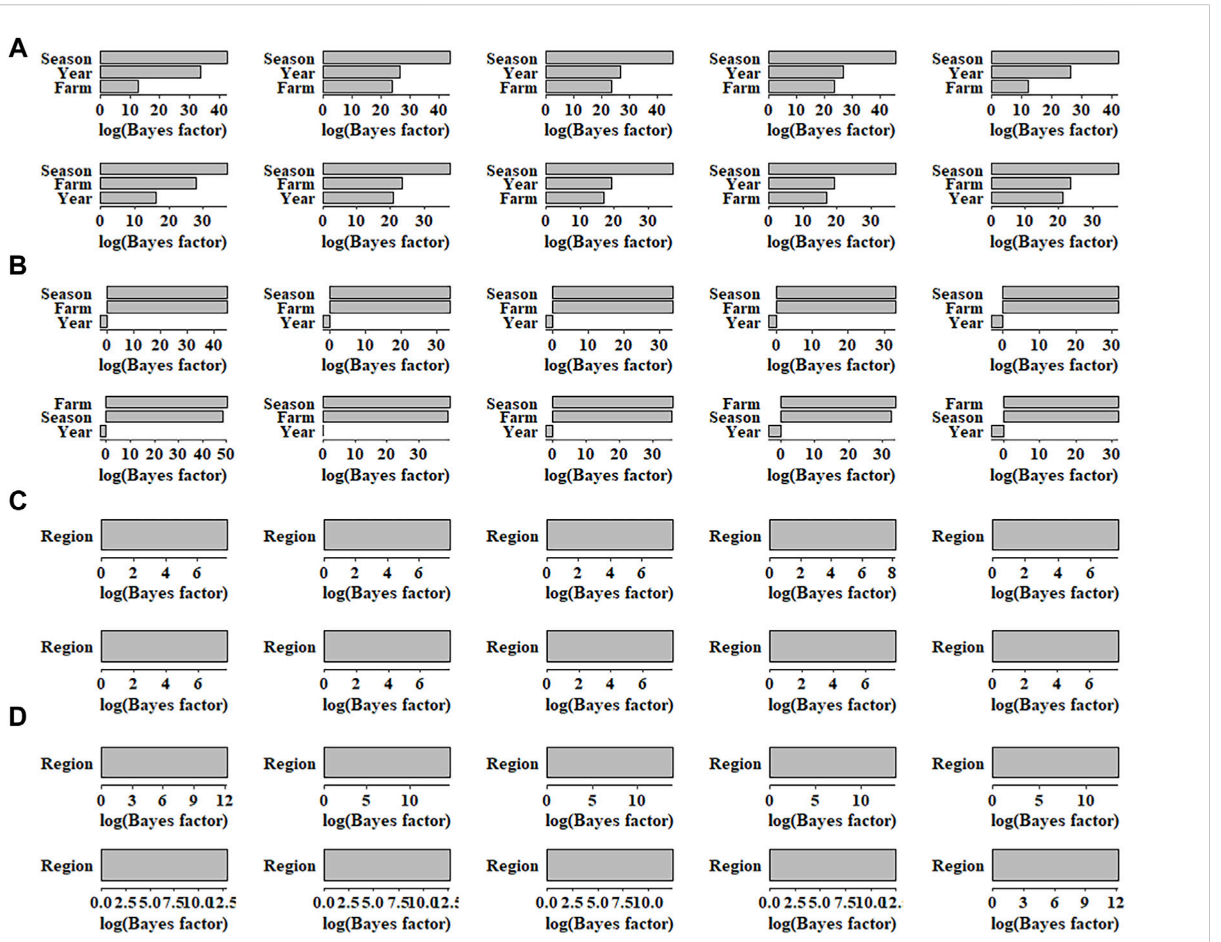


FIGURE 4
Bayes factors of 10 single nucleotide polymorphisms for **(A)** AGE, **(B)** BFT, **(C)** GW and **(D)** WC in different environmental factors. AGE: days to 100 kg; BFT: backfat thickness adjusted to 100 kg; GW: grain weight; WC: water content.

TABLE 4 Average computation time for G × EBLUP and GBLUP to complete each fold of cross-validation.

Date set	Trait[a]	G × EBLUP	GBLUP
Simulation	V0.25	45 min 48 s	15 min 3 s
	V1	46 min 27 s	15 min 12 s
	V2	46 min 13 s	15 min 9 s
Pig	AGE	30 min 9 s	2 min 14 s
	BFT	30 min 14 s	2 min 18 s
Maize	GW	26 min 13 s	1 min 7 s
	WC	26 min 8 s	1 min 10 s

[a] V0.25, V1 and V2: The traits with variance of G × E interactions of 0.25, 1 and 2, respectively, in simulated data.

different numbers of environment variables (Figures 2C,D), implying that the number of environment levels have no effect on our new method. This is a major highlight of the G × EBLUP method compared with the methods based on the multi-trait (Richard, 1996) and reaction norm models (Rebecka et al., 2002), in which the number of environment levels was the main limitation (Song et al., 2020a; Song et al., 2020b). The advantage of G × EBLUP for joint G × E analysis of multiple environment variables could be that multiple environments can interact with a single genetic locus to influence the phenotypes (Moore et al., 2019). In our new method, the interactions of genotype with different environments could be represented by one or more markers, as explicated by G × EWAS. Therefore, it was not sensitive to the number of environment levels.

In this study, we proposed G × EWAS and a score-test statistic to identify the significance of SNP affected by GEI; the details of G × EWAS and its computational complexity can be found in Additional File 1 Supplementary Material. In G × EWAS, **E** is the N × L matrix of L observed environments, and **EE**′ is used as a variance–covariance structure for G × E effects; thus, G × EWAS is a single-SNP regression model, which is different from the multi-trait (Richard, 1996) and reaction norm models (Rebecka et al., 2002). The linear covariance function (**EE'**) was primarily used because of two appealing properties. First, as the number of samples typically exceeds the number of environments in larger populations (L << N), a low-rank linear covariance is noted, which enables parameter inference with a computational complexity that scales linearly with the increasing population size. Second, a linear covariance is directly interpretable, as there is one-to-one correspondence between G × EWAS and linear regression using L covariates to account for GEI. Compared with the four methods proposed by Wang et al. (2019), our results showed the obvious advantages of G × EWAS in GEI detection (Additional File 3: Supplementary Figures S8–S10 and Table 2). The low efficiency of the other methods could be because the selection, epistasis and phantom vQTL can also cause vQTL instead of just GEI, which may lead to biases in the detection of G × E markers, e.g. the overlapped simulated QTLs were decreased for L-mean and L-median when the variance of G × E increased. Although the efficiency of G × EWAS was improved with increase in the variance of GEI, it yielded higher number of overlap QTLs with GEI. As a combination of the standard linear mixed model for genome-wide association study and StructLMM, our proposed G × EWAS performed better than StructLMM (Additional File 3: Supplementary Figures S8–S10 and Table 2). The superiority of G × EWAS was mainly because it built a genomic relationship matrix to capture the realised relationships among individuals; moreover, it can accurately capture the effect of each environment on markers by adding the GEI vector in the model, which follows a multivariate normal distribution. In the scenario of larger variance of GEI, more QTLs would contribute to the GEI, increasing the weight of per-individual effect size as described in Eq. 1; thus, it could be easily detected using G × EWAS.

Although G × EWAS could detect more significant SNPs associated with GEI using Bonferroni correction, only a small amount of the whole markers were detected. The best performance of G × EBLUP was obtained at the marker selection criterion of *p*-value of E02 (BFT in pig and WC in maize) or E03 (simulated data and WC in maize) (*p*-values < 10^{-2} or 10^{-3}) in all scenarios (Table 3). In fact, the performance of G × EBLUP at E02 and E03 was similar. Accordingly, the selected SNPs with GEI for simulated data, AGE and BFT in pig and WC and GW in maize were enough to build a genomic relationship matrix to elucidate the contribution of GEI. The number of selected SNPs with GEI in G × EBLUP was lower than that of the significant SNPs at false discovery rate of 0.05 (Additional File 2: Supplementary Table S2). Therefore, it is reasonable to use *p*-values of <10^{-3} as threshold for determination of SNPs with GEI in G × EBLUP. The advantage of G × EBLUP over GBLUP is mainly because G × EBLUP allows the assignment of different weights to the genomic variants in the different genomic relationship based on their estimated genomic parameters, which can better fit the genetic architecture of the trait, while randomly selected a subset of SNPs that are not all associated with the trait, giving more weight to these SNPs in G × EBLUP does not improve the accuracy of genomic prediction (results not shown).

Along with the mapping of G × E markers, G × EWAS could determine favourable environment using BF of SNPs on each environment. This is extremely helpful for market-specific breeding in animals and plants as it may provide further explanation to those individuals who have a higher risk of being affected by GEI in a certain environment variable (sensitive environment). In this study, farm was identified as a sensitive environment for BFT in pigs, which allows the selection of elite individuals with good performance in specified farms. Further, our results showed that GEI was different for distinct traits, e.g. similar genomic prediction accuracies were obtained for G × EBLUP and GBLUP for AGE, whereas G × EBLUP showed improvement by approximately 3% in the prediction accuracy for BFT in pigs (Figure 3). This observation is consistent with that of our previous report that showed GEI for BFT but not for AGE (Song et al., 2020b). This could be explained by values of the variance of the slope ($\sigma^2_{a_1}$) compared with those of the intercept ($\sigma^2_{a_0}$) in the reaction norm model; $\sigma^2_{a_1}$ / $\sigma^2_{a_0}$ were 0.002 and 0.348 for AGE and BFT, respectively. Thus, traits with small variance of GEI cannot improve the accuracy of genomic prediction using G × EBLUP even after the identification of more significant SNPs on AGE. Similarly, the less significant SNPs in the maize data showed larger variance and greater improvement in the genomic prediction accuracy than those in the pig data (Figure 3). Further, this might be due to the trait characteristics of plants, which are more vulnerable to GEI than livestock, the effect of environment in animal become ignorable as the industrial management. Conversely, the small sample size of the maize may have reduced the power of G × EWAS, leading to the identification of a small number of significant GEI markers, which may explain why the gain of G × EBLUP was lower than expected. In addition, using sensitive tests to detect sensitive environments is an alternative, and then fitting the overlap environment of SNPs in the G × EBLUP model based on the pre-defined environment. However, the effect of this method and how to fit it in G × EBLUP model need to be investigated in the future.

Our results showed G × EBLUP is a powerful alternative to the conventional method for the estimation of genomic breeding value. However, there are several limitations in this approach. First, G × EWAS is not a whole-genome regression model and does not account for the genome-wide contribution of all other variants, thus G × EWAS is still a single marker regression model with a long calculation time (Table 4). G × EWAS assumes all SNPs affected by GEI in the current model, and it could not differentiate between the significant SNPs with or without GEI. It might be the reason why a large number of significant SNPs were detected in the simulated data, although only few overlapped with simulated QTLs. The same phenomenon was also found in other methods, such as Bartlett, F-Killeen, L-mean, L-median and StructLMM, implying that the detection of GEI is more difficult due to the flexibility of the environment. Second, although sensitive environment can be obtained by calculating the BF value for each environment variable, the BF value of each level of an environment variable cannot be obtained (e.g., BF for each farm in pig data), which might be important for directional breeding. Third, G × EBLUP has an advantage only when the variance of the G × E interaction is relatively large, e.g., similar genomic prediction accuracies were obtained for G × EBLUP and GBLUP for AGE, as was not affected by GEI. Finally, G × EWAS cannot handle binary traits at present, as G×E tests need the estimation of nuisance parameters to capture the main effects of binary traits, and estimating these parameters requires high-dimensional integration and the inversion of a high-dimensional similarity matrix. Nevertheless, it is worth being investigated in the future.

Moreover, G × EBLUP can be extended to single-step method (ssG × EBLUP), which could improve the genomic prediction accuracy using pedigree and genomic information (Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund., 2010).

Conclusion

The G × EBLUP method proposed in this study showed the following four features: 1) genomic prediction was performed using the G × EBLUP method by considering GEI and yielded higher accuracies and lower MSE in both simulated and real pig and maize data when the variance of G × E interaction is large; 2) it could powerfully detect the genome-wide SNPs subject to GEI; 3) the number of environment levels did not influence the genomic prediction accuracy of the proposed G × EBLUP, circumventing the limitation of current methods; 4) it could determine favourable environment using SNP BF for each environment, thus being useful for market-specific animal and plant breeding.

Data availability statement

The simulated data, pig and maize data supporting the conclusions of this article are available from Figshare: https://figshare.com/articles/dataset/GXEBLUP/20347368.

Ethics statement

The animal study was reviewed and approved by Animal handling and sample collection were conducted according to protocols approved by the Institutional Animal Care and Use Committee (IACUC) at China Agricultural University. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

HS and XD conceived the new method, designed the experiment. HS, XW and YG participated in the data analysis and the result interpretation. HS and XW wrote the paper. XD revised the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Project (grant number: 2019YFE0106800); Modern Agriculture Science and Technology Key Project of Hebei Province (grant number: 19226376D); Beijing Natural Science Foundation (grant number: 6222014) and China Agriculture Research System of MOF and MARA.

Acknowledgments

The authors gratefully acknowledge Lai Jinsheng and Song Weibin at China Agricultural University, Beijing Tongzhou International Seed Co., Ltd., Beijing Liuma Pig Breeding Co., Ltd. for providing maize data and pig blood 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

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.2022.972557/full#supplementary-material

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93 (2), 743–752. doi:10.3168/jds.2009-2730

Braz, C. U., Rowan, T. N., Schnabel, R. D., and Decker, J. E. (2021). Genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. *Sci. Rep.* 11 (1), 13335. doi:10.1038/s41598-021-92455-x

Browning, B. L., and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84 (2), 210–223. doi:10.1016/j.ajhg.2009.01.005

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., Lee, J. J., et al. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2 (25), 4–7. doi:10.1186/s13742-015-0047-8

Christensen, O., and Lund, M. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42 (1), 2. doi:10.1186/1297-9686-42-2

Crossa, J. (2012). From genotype × environment interaction to gene × environment interaction. *Curr. Genomics* 13 (3), 225–244. doi:10.2174/138920212800543066

Davies, R. (1980). Algorithm as 155: The distribution of a linear combination of χ2 random variables. *Appl. Stat.* 29 (3), 323–333. doi:10.2307/2346911

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinforma.* 12, 186. doi:10.1186/1471-2105-12-186

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127 (2), 463–480. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.]. doi:10.1007/s00122-013-2231-5

Huan, L., Yongqiang, T., and Hao, H. Z. (2008). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Statistics Data Analysis* 53 (4), 853. doi:10.1016/j.csda.2008.11.025

Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127 (3), 595–607. N.I.H., Extramural; Research Support, Non-U.S. Gov't]. doi:10.1007/s00122-013-2243-1

Jarquin, D., Lemes, D. S. C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing Genomic-Enabled prediction accuracy by modeling genotype x environment interactions in Kansas wheat. *The plant genome* 10 (2). doi:10.3835/plantgenome2016.12.0130

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90 (430), 773–795. doi:10.1080/01621459.1995.10476572

Kerin, M., and Marchini, J. (2020). Inferring gene-by-environment interactions with a bayesian whole-genome regression model. *Am. J. Hum. Genet.* 107 (4), 698–713. [Journal Article; Research Support, Non-U.S. Gov't]. doi:10.1016/j.ajhg.2020.08.009

Lippert, C., Xiang, J., Horta, D., Widmer, C., Kadie, C., Heckerman, D., et al. (2014). Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* 30 (22), 3206–3214. N.I.H., Extramural; Research Support, Non-U.S. Gov't]. doi:10.1093/bioinformatics/btu504

Liu, A., Su, G., Hoglund, J., Zhang, Z., Thomasen, J., Christiansen, I., et al. (2019). Genotype by environment interaction for female fertility traits under conventional and organic production systems in Danish Holsteins. *J. Dairy Sci.* 102 (9), 8134–8147. doi:10.3168/jds.2018-15482

Madsen, P., Sorensen, P., Su, G., Damgaard, L. H., Thomsen, H., and Labouriau, R. (2006). "Dmu - a package for analyzing multivariate mixed models," in the proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Brasil, 11–27.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4), 1819–1829. doi:10.1093/genetics/157.4.1819

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

Moore, R., Casale, F. P., Jan, B. M., Horta, D., Franke, L., Barroso, I., et al. (2019). A linear mixed-model approach to study multivariate gene-environment interactions. *Nat. Genet.* 51 (1), 180–186. doi:10.1038/s41588-018-0271-0

Perez-Rodriguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., et al. (2017). Single-Step genomic and pedigree genotype x environment interaction models for predicting wheat lines in international environments. *Plant Genome-US* 10 (2), 28724079. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.]. doi:10.3835/plantgenome2016.09.0089

Rebecka, K., Erling, S., Per, M., Just, J., and Hossein, J. (2002). Genotype by environment interaction in nordic dairy cattle studied using reaction norms. *Acta Agric. Scand. Sect. A - Animal Sci.* 52 (1), 11–24. doi:10.1080/09064700252806380

Richard, F. (1996). Introduction to quantitative genetics (4th edn). *Trends Genet.* 12 (7), 280. doi:10.1016/0168-9525(96)81458-2

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14 (6), R55. doi:10.1186/gb-2013-14-6-r55

Song, H., Zhang, J., Jiang, Y., Gao, H., Tang, S., Mi, S., et al. (2017). Genomic prediction for growth and reproduction traits in pig using an admixed reference population. *J. Anim. Sci.* 95 (8), 3415–3424. doi:10.2527/jas.2017.1656

Song, H., Zhang, Q., and Ding, X. (2020a). The superiority of multi-trait models with genotype-by-environment interactions in a limited number of environments for genomic prediction in pigs. *J. Anim. Sci. Biotechnol.* 11 (1), 88. doi:10.1186/s40104-020-00493-8

Song, H., Zhang, Q., Misztal, I., and Ding, X. (2020b). Genomic prediction of growth traits for pigs in the presence of genotype by environment interactions using single-step genomic reaction norm model. *J. Anim. Breed. Genet.* 137 (6), 523–534. doi:10.1111/jbg.12499

Tiezzi, F., de Los, C. G., Parker, G. K., and Maltecca, C. (2017). Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle. *J. Dairy Sci.* 100 (3), 2042–2056. doi:10.3168/jds.2016-11543

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92 (1), 16–24. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.]. doi:10.3168/jds.2008-1514

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* 5 (8), w3538. doi:10.1126/sciadv.aaw3538

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89 (1), 82–93. doi:10.1016/j.ajhg.2011.05.029

Zhang, Z., Kargo, M., Liu, A., Thomasen, J. R., Pan, Y., and Su, G. (2019). Genotype-by-environment interaction of fertility traits in Danish Holstein cattle using a single-step genomic reaction norm model. *Hered. (Edinb)* 123 (2), 202–214. [Journal Article; Research Support, Non-U.S. Gov't]. doi:10.1038/s41437-019-0192-4

Zhong, S., Dekkers, J. C., Fernando, R. L., and Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* 182 (1), 355–364. U.S. Gov't, Non-P.H.S.]. doi:10.1534/genetics.108.098277

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 14 October 2022
DOI 10.3389/fgene.2022.922369

Multifactorial methods integrating haplotype and epistasis effects for genomic estimation and prediction of quantitative traits

Yang Da*, Zuoxiang Liang and Dzianis Prakapenka

Department of Animal Science, University of Minnesota, Saint Paul, MN, United States

The rapid growth in genomic selection data provides unprecedented opportunities to discover and utilize complex genetic effects for improving phenotypes, but the methodology is lacking. Epistasis effects are interaction effects, and haplotype effects may contain local high-order epistasis effects. Multifactorial methods with SNP, haplotype, and epistasis effects up to the third-order are developed to investigate the contributions of global low-order and local high-order epistasis effects to the phenotypic variance and the accuracy of genomic prediction of quantitative traits. These methods include genomic best linear unbiased prediction (GBLUP) with associated reliability for individuals with and without phenotypic observations, including a computationally efficient GBLUP method for large validation populations, and genomic restricted maximum estimation (GREML) of the variance and associated heritability using a combination of EM-REML and AI-REML iterative algorithms. These methods were developed for two models, Model-I with 10 effect types and Model-II with 13 effect types, including intra- and inter-chromosome pairwise epistasis effects that replace the pairwise epistasis effects of Model-I. GREML heritability estimate and GBLUP effect estimate for each effect of an effect type are derived, except for third-order epistasis effects. The multifactorial models evaluate each effect type based on the phenotypic values adjusted for the remaining effect types and can use more effect types than separate models of SNP, haplotype, and epistasis effects, providing a methodology capability to evaluate the contributions of complex genetic effects to the phenotypic variance and prediction accuracy and to discover and utilize complex genetic effects for improving the phenotypes of quantitative traits.

KEYWORDS

multifactorial model, epistasis, haplotype, SNP, GBLUP, GREML

Check for updates

OPEN ACCESS

EDITED BY
Kefei Chen,
Curtin University, Australia

REVIEWED BY
Tianjing Zhao,
University of California, Davis,
United States
José Marcelo Soriano Viana,
Universidade Federal de Viçosa, Brazil

*CORRESPONDENCE
Yang Da,
yda@umn.edu

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

RECEIVED 17 April 2022
ACCEPTED 12 September 2022
PUBLISHED 14 October 2022

CITATION
Da Y, Liang Z and Prakapenka D (2022), Multifactorial methods integrating haplotype and epistasis effects for genomic estimation and prediction of quantitative traits.
Front. Genet. 13:922369.
doi: 10.3389/fgene.2022.922369

Introduction

Genomic estimation of variance components and associated heritabilities and genomic prediction for quantitative traits using single nucleotide polymorphism (SNP) markers and mixed models have become a widely used approach for genetic improvement in livestock and crop species. The rapid growth in genomic selection data provides unprecedented opportunities to discover and utilize complex genetic mechanisms, but methodology and computing tools are lacking for investigating complex genetic mechanisms using the approach of genomic estimation and prediction. The integration of global low-order epistasis effects and local high-order epistasis effects contained in haplotypes for genomic estimation and prediction is a step forward for the discovery and application of complex genetic mechanisms to improve the phenotypes of quantitative traits.

The theory of genetic partition of two-locus genotypic values defines four types of epistasis values: additive × additive (A×A), additive × dominance (A×D), dominance × additive (D×A), and dominance × dominance (D×D) epistasis values (Cockerham, 1954; Kempthorne, 1954). The Cockerham method defines each epistasis coefficient as the product of the coefficients of the two interacting effects that each can be an additive or dominance effect (Cockerham, 1954). This definition of epistasis coefficient is the basis for defining epistasis model matrices in terms of the model matrices of additive and dominance effects. Cockerham also defines a pedigree epistasis relationship as the product between the pedigree additive and dominance relationships (Cockerham, 1954), and this definition is the theoretical basis for Henderson's approach to express epistasis relationship matrices as the Hadamard products of the additive and dominance relationship matrices (Henderson, 1985). Henderson's Hadamard products for epistasis relationship matrices were suggested for genomic prediction using epistasis effects by replacing the pedigree additive and dominance relationship matrices with the genomic additive and dominance relationship matrices calculated from SNP markers (Su et al., 2012; Muñoz et al., 2014; Vitezica et al., 2017). This genomic version of Henderson's Hadamard products avoids the use of large epistasis model matrices that can be difficult or impossible to compute but contains intra-locus epistasis effects that are not present in the epistasis model (Martini et al., 2020). For this reason, the genomic version of Henderson's Hadamard products could be described as approximate genomic epistasis relationship matrices (AGERM). Formulations have been developed to obtain the exact genomic epistasis relationship matrices (EGERM) that remove the intra-locus epistasis effects in AGERM by modifying Henderson's Hadamard products without creating the epistasis model matrices (Jiang and Reif, 2015; Martini et al., 2016; Jiang and Reif, 2020; Martini et al., 2020). The difference between AGERM and EGERM tends to diminish as the number of SNPs increases (Jiang and Reif, 2020). Henderson's Hadamard products and hence AGERM are applicable to any order of epistasis effects, and EGERM also has a general formula for any order of epistasis effects (Jiang and Reif, 2020). However, limited tests showed that fourth-order global epistasis contributed virtually nothing to the phenotypic variance but generated considerable computing difficulty (Liang et al., 2021), raising questions about the value of global epistasis effects beyond the third-order. Methods of genomic estimation and prediction of global epistasis effects up to the third-order should have wide-range applications, given that the number of reported epistasis effects lags far behind the number of single-point effects (Carlborg and Haley, 2004; Phillips, 2008; Ritchie and Van Steen, 2018) even though epistasis effects are important genetic effects (Cordell, 2002; Segre et al., 2005; Mackay, 2014). In contrast to the computing difficulty and uncertain impact of global high-order epistasis effects beyond the third-order, local high-order epistasis effects in haplotypes with potentially many SNPs were responsible for the increased accuracy of predicting phenotypic values of certain traits (Liang et al., 2020; Bian et al., 2021). The integration of haplotype and epistasis effects provides an approach to investigate the contributions of global low-order epistasis effects and local high-order epistasis effects to the phenotypic variance and the accuracy of genomic prediction under the same model.

An epistasis GWAS in Holstein cattle showed that intra- and inter-chromosome epistasis effects affected different traits differently, for example, the daughter pregnancy rate was mostly affected by inter-chromosome epistasis effects, whereas milk production traits were mostly affected by intra-chromosome epistasis effects (Prakapenka et al., 2021), and genomic heritability estimates of intra- and inter-chromosome heritabilities for the daughter pregnancy rate using methods in this article showed that inter-chromosome A×A heritability was much higher than the intra-chromosome A×A heritability (Liang et al., 2022). Therefore, dividing pairwise epistasis effects into intra- and inter-chromosome epistasis effects allows the investigation of the contributions of intra- and inter-chromosome pairwise epistasis effects to the phenotypic variance and prediction accuracy.

The purpose of the multifactorial model in this article is to integrate haplotype effects and epistasis effects up to the third-order for genomic estimation of variance components and associated heritabilities, as well as genomic prediction of genetic and phenotypic values of quantitative traits, to provide a general and flexible methodology framework for genomic prediction and estimation using complex genetic mechanisms and to provide methodology details of the EPIHAP computer package that implements the integration of haplotype and epistasis effects (Liang et al., 2021, 2022). The methodology in this article will facilitate the discovery and utilization of global low-order and local high-order epistasis effects relevant to the phenotypic variances and prediction accuracies of quantitative traits, and obtain new knowledge of complex genetic mechanisms underlying quantitative traits.

Materials and methods

Quantitative genetics model with single nucleotide polymorphism, haplotype, and epistasis effects and values

The mixed model with single-SNP additive and dominance effects, haplotype additive effects, and pairwise SNP epistasis effects in this article is based on the quantitative genetics (QG) model resulting from the genetic partition of single-SNP genotypic values (Da et al., 2014; Wang and Da, 2014), haplotype genotypic values (Da, 2015), and pairwise genotypic values (Cockerham, 1954). An advantage of this QG model is the readily available quantitative genetics interpretations of SNP additive and dominance effects, values, and variances; haplotype additive effects, values, and variances; epistasis effects, values, and variances; and the corresponding SNP, haplotype, and epistasis heritability estimates. Two QG models are developed: Model-I with 10 effect types, including SNP additive and dominance effects, haplotype additive effects, and epistasis effects up to the third-order; and Model-II with 13 effect types resulting from replacing the pairwise epistasis effects of Model-I with intra- and inter-chromosome epistasis effects. Detailed descriptions of the effects, values, model matrices, the coding of the model matrices, as well as the precise definition of each term in the two QG models, are provided in Supplementary Text S1 and Supplementary Table S1. With these precise definitions of genetic effects, values, and model matrices in the QG models, a concise multifactorial QG model covering both Model-I and Model-II can be established, that is

$$\mathbf{g} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{W}_i\boldsymbol{\tau}_{io} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{u}_i \tag{1}$$

$$\mathbf{u}_i = \mathbf{W}_i\boldsymbol{\tau}_{io} \tag{2}$$

where $\boldsymbol{\tau}_{io}$ = genetic effects of the i^{th} effect type from the original QG model based on genetic partition, $\mathbf{W}_i$ = model matrix of $\boldsymbol{\tau}_{io}$, $\mathbf{u}_i$ = genetic values of the i^{th} effect type from the original QG model, and f = number of effect types. For Model-I, subscripts $i = 1, \ldots, 10$ represent SNP additive (A), SNP dominance (D), haplotype additive, A×A, A×D, D×D, A×A×A, A×A×D, A×D×D, and D×D×D effects sequentially. For Model-II, subscripts $i = 1, \ldots, 13$ represent SNP additive, SNP dominance, haplotype additive, intra-chromosome A×A, intra-chromosome A×D, intra-chromosome D×D, inter-chromosome A×A, inter-chromosome A×D, inter-chromosome D×D, A×A×A, A×A×D, A×D×D, and D×D×D effects sequentially. The variance–covariance matrix of the genetic values of Eqs 1 and 2 is

$$\mathbf{G} = \mathrm{var}\left(\sum_{i=1}^{f} \mathbf{W}_i\boldsymbol{\tau}_{io}\right) = \sum_{i=1}^{f} \mathrm{Var}(\mathbf{u}_i) = \sum_{i=1}^{f} \mathbf{G}_i = \sum_{i=1}^{f} \sigma^2_{io}\mathbf{W}_i\mathbf{W}_i' \tag{3}$$

$$\mathrm{Var}(\boldsymbol{\tau}_{io}) = \sigma^2_{io}\mathbf{I} \tag{4}$$

$$\mathbf{G}_i = \mathrm{Var}(\mathbf{u}_i) = \mathbf{W}_i\mathrm{Var}(\boldsymbol{\tau}_{io})\mathbf{W}_i' = \sigma^2_{io}\mathbf{W}_i\mathbf{W}_i' \tag{5}$$

where $\sigma^2_{io} = \mathrm{Var}(\tau_{ijo})$ genetic variance of the i^{th} effect type under the original QG model is common to all individuals (all j values). It is of note that $\mathbf{W}_i\mathbf{W}_i'$ is not a genomic relationship matrix but is the primary information for calculating each genomic relationship matrix. The structure of the **G** matrix of Eqn. 3 assumes independence between the genetic values of different effect types. However, the GBLUP values of different effect types using the **G** matrix of Eqn. 3 could be correlated. Under the Hardy–Weinberg equilibrium (HWE) and LE assumptions, additive, dominance, and epistasis effects are independent of each other (Cockerham, 1954; Kempthorne, 1954). For genome-wide SNPs, the LE assumption generally does not hold for closely linked loci, and nonzero Hardy–Weinberg disequilibrium (HWD) may exist numerically. These and other unknown factors in real data may result in the existence of correlations between different effect types. Haplotype additive values are correlated with SNP additive effects because a haplotype additive value is the sum of all SNP additive values and an epistasis value within the haplotype block plus a potential haplotype loss (Da et al., 2016). In two recent haplotype studies for genomic prediction, the integration of SNP and haplotype effects increased the prediction accuracy for four of the seven traits in the human study (Liang et al., 2020) and for three of the eight traits in the swine study (Bian et al., 2021), showing that SNP and haplotype additive values compensated each other for prediction accuracy and that the correlation between SNP and haplotype additive values were incomplete for those traits. The correlation between haplotype and epistasis values can be complex. The correlation should be nonexistent if the A×A values are inter-chromosome A×A values or intra-chromosome A×A values involving distal SNPs not covered by the haplotypes, but the correlation could be strong if the A×A values are intra-chromosome A×A values involving proximal SNPs covered by the haplotypes.

The reparametrized and equivalent quantitative genetics model for genomic estimation and prediction

Genomic relationship matrices are used for genomic estimation and prediction, and the use of genomic relationship matrices results in a reparametrized and equivalent model of the original QG model for genetic values, to be referred to as the RE-QG model, where "reparametrized" refers to the reparameterization of the genetic effects, model matrix, and genetic variance of each effect type; and "equivalent" refers to the requirement of the same first and second moments for the original QG model (Eqs 1–5) and the RE-QG model. This RE-QG model of genetic values can be expressed as

$$\mathbf{g} = \mu\mathbf{I} + \sum\nolimits_{i=1}^{f} \mathbf{T}_i\boldsymbol{\tau}_i = \mu\mathbf{I} + \sum\nolimits_{i=1}^{f} \mathbf{u}_i \tag{6}$$

$$\mathbf{G} = \mathrm{var}\left(\sum\nolimits_{i=1}^{f} \mathbf{u}_i\right) = \sum\nolimits_{i=1}^{f} \mathbf{G}_i = \sum\nolimits_{i=1}^{f} \sigma_i^2 \mathbf{T}_i\mathbf{T}_i' = \sum\nolimits_{i=1}^{f} \sigma_i^2 \mathbf{S}_i$$
$$= \sum\nolimits_{i=1}^{f} \sigma_{io}^2 \mathbf{W}_i\mathbf{W}_i' \tag{7}$$

where

$$\boldsymbol{\tau}_i = \sqrt{k_i}\boldsymbol{\tau}_{io} = \text{genetic effects of the } i^{th} \text{ effect type} \tag{8}$$

$$\mathbf{T}_i = \mathbf{W}_i \big/ \sqrt{k_i} = \text{model matrix of } \boldsymbol{\tau}_i \tag{9}$$

$$\sigma_i^2 = \mathrm{Var}\left(\tau_{ij}\right) = \mathrm{tr}\left(\mathbf{G}_i\right)/n = \sum\nolimits_{j=1}^{n} G_i^{jj} \big/ n = k_i\sigma_{io}^2 \tag{10}$$

= variance of the genetic effects of the i^{th} effect type common to all individuals

= average variance of all individuals for the genetic values of the i^{th} effect type

$$\mathbf{u}_i = \mathbf{T}_i\boldsymbol{\tau}_i = \mathbf{W}_i\boldsymbol{\tau}_{io} = \text{genetic values of the } i^{th} \text{ effect type} \tag{11}$$

$$\mathbf{G}_i = \mathrm{Var}\left(\mathbf{u}_i\right) = \sigma_i^2\mathbf{T}_i\mathbf{T}_i' = \sigma_i^2\mathbf{S}_i = \sigma_{io}^2\mathbf{W}_i\mathbf{W}_i' \tag{12}$$

= variance–covariance matrix of the genetic values of the i^{th} effect type

$$\mathbf{S}_i = \mathbf{T}_i\mathbf{T}_i' = \mathbf{W}_i\mathbf{W}_i' \big/ k_i$$
$$= \text{genomic relationship matrix of the } i^{th} \text{ effect type} \tag{13}$$

$$k_i = \mathrm{tr}\left(\mathbf{W}_i\mathbf{W}_i'\right)\big/ n$$
$$= \text{average of the diagonal elements of } \mathbf{W}_i\mathbf{W}_i'. \tag{14}$$

Equations 8–10 are the reparametrization of the genetic effects, model matrices, and genetic variances of the original QG model, whereas Eqs 11 and 12 show the genetic values and the variance–covariance matrix of the genetic values are the same under the RE-QG and QG models. In Eq.10, G_i^{jj} = the genetic variance of the j^{th} individual for the i^{th} effect type = the j^{th} diagonal element of the $\mathbf{G}_i$ matrix defined by Eq. 12. The k_i formula of Eq. 14 as the average of the diagonal elements of $\mathbf{W}_i\mathbf{W}_i'$ was originally proposed for genomic additive relationships (Hayes and Goddard, 2010) and was used for genomic dominance relationships (Da et al., 2014; Wang and Da, 2014), haplotype additive genomic relationships (Da, 2015), and pairwise epistasis genomic relationships (Vitezica et al., 2017). The need for this RE-QG model is due to the use of the genomic relationship matrices (e.g., Eq. 13) because the QG model does not contain genomic relationship matrices (Eq. 3). Detailed notations of the QG model of Eqs 1–5 in reference to the RE-QG model described by Eqs 6–14 are summarized in Supplementary Table S1.

The formula of the genomic relationship matrix ($\mathbf{S}_i$ of Eq. 13) is based on the model matrix of each effect type and can be difficult or impossible to compute if epistasis model matrices are used. This computing difficulty of epistasis model matrices is removed by calculating $\mathbf{S}_i$ based on the model matrices of SNP additive and dominance effects without creating the epistasis model matrices using either AGERM or EGERM. AGERM refers to the genomic version of Henderson's Hadamard products between pedigree additive and dominance relationship matrices (Henderson, 1985), with the pedigree additive and dominance relationship matrices replaced by the genomic additive and dominance relationship matrices (Su et al., 2012; Muñoz et al., 2014; Vitezica et al., 2017). AGERM contains intra-locus epistasis that should not exist (Martini et al., 2020), and EGERM removes intra-locus epistasis from AGERM based on products between genomic additive and dominance relationship matrices (Jiang and Reif, 2020; Martini et al., 2020).

The QG and RE-QG models have the same prediction accuracy due to the equivalence between these two models. The genetic values ($\mathbf{u}_i$, Eqs 2, 11) and the variance–covariance matrix of the genetic values ($\mathbf{G}_i$, Equations 5 and 12) under the QG and RE-QG models are identical, although $\mathbf{u}_i$ and $\mathbf{G}_i$ have different expressions under the QG and RE-QG models. Consequently, the QG model without using genomic relationship matrices and the RE-QG model using genomic relationship matrices have identical accuracy of genomic prediction. The choice of the k_i formula for defining the genomic relationship matrix does not affect the accuracy of genomic prediction but affects the interpretation and application of the genetic variance and genomic relationships for each effect type. Since the interpretation of each genetic variance is a focus, whereas the interpretation of the genomic relationships is not a focus in this study, the interpretation of the genetic variance and associated heritability is the consideration in choosing the k_i formula of Eq.14.

The RE-QG model using genomic relationships (Equations 6–14) has two major advantages over the QG model without using genomic relationship matrices (Equations 1–5), although the two models have the same prediction accuracy. First, the use of genomic relationships, originally proposed for genomic additive relationships (VanRaden, 2008), provides a genomic version of the traditional theory and methods of best linear unbiased prediction (BLUP) that uses pedigree relationships, and this genomic version can utilize a wealth of BLUP-based theory, methods, and computing strategies. Second, the genetic variance of the genetic effects of each effect type under the RE-QG model can be used for estimating genomic heritability, whereas the genetic variance of the genetic effects under the QG model cannot be used for estimating genomic heritability. With the k_i value defined by Eq. 14, the variance of the genetic effects of the i^{th} effect type, $\sigma_i^2 = k_i\sigma_{io}^2$ (Eq.10), has the unique interpretation as the average variance of the genotypic values of all individuals and is a common variance to all individuals. Moreover, $\sigma_i^2 = k_i\sigma_{io}^2$ is unaffected by the number of levels for each effect type, unless the number of levels such as the number of SNPs is too small to provide sufficient coverage of the genome (Da et al., 2014; Tan et al., 2017; Liang et al., 2020). In contrast, the QG model does not have a method to estimate genetic variance components for calculating genomic heritabilities because σ^2_{io} is an inverse function of the number of effect levels. As the number of effect levels such as the number of SNPs increases or decreases, the value of each element in $\mathbf{W}_i\mathbf{W}_i^{'}$ changes in the same direction and the σ^2_{io} estimate changes in the opposite direction, that is, as the number of effect levels increases or decreases, σ^2_{io} decreases or increases. Consequently, the σ^2_{io} estimate does not have a unique interpretation and cannot be used for estimating genomic heritability (Da et al., 2014). Moreover, the variance of the genetic value of an individual ($\sigma^2_{io}\,(\mathbf{W}_i\mathbf{W}_i^{'})^{jj}$) cannot be used for calculating genomic heritability because of the individual specificity of the $(\mathbf{W}_i\mathbf{W}_i^{'})^{jj}$ values, as shown as follows.

The exact relationship between the genetic variance for the ith effect type of the jth individual under the RE-QG model and the QG model can be described based on the $\mathbf{G}_i$ matrix defined by Eq. 12:

$$G_i^{jj} = \mathrm{Var}\left(u_{ij}\right) = \sigma_i^2\,(\mathbf{S}_i)^{jj} = \sigma^2_{io}\left(\mathbf{W}_i\mathbf{W}_i^{'}\right)^{jj} \tag{15}$$

where G_i^{jj} = the jth diagonal element of the $\mathbf{G}_i$ matrix defined by Eq.12 = the genetic variance of the jth individual for the genotypic value of the ith effect type, and u_{ij} = the jth element of $\mathbf{u}_i$ defined by Eq.11. Equation 15 shows that different individuals do not have a common variance of the genetic values (G_i^{jj}) unless all diagonal elements of $\mathbf{S}_i$ or $\mathbf{W}_i\mathbf{W}_i^{'}$ are identical, which could not happen with genome-wide SNP data in the absence of identical twins because genome-wide SNPs have a high degree of individual specificity. Consequently, G_i^{jj} is not a common variance to all individuals and cannot be used for calculating the genomic heritability of the ith effect type. In contrast, σ_i^2 of Eq.10 under the RE-QG model as the average variance of the genotypic values of all individuals is common to all individuals and can be used for calculating the heritability of each effect type. For the example of Model-I, the exact genetic interpretation of G_i^{jj} is $G_i^{jj} = \sigma^2_{aj}$ = the variance of the genomic additive (breeding) value of the jth individual for i = 1, $G_i^{jj} = \sigma^2_{dj}$ = the variance of the genomic dominance value of the jth individual for i = 2, $G_i^{jj} = \sigma^2_{ahj}$ = the variance of the genomic haplotype additive value of the jth individual for i = 3, $G_i^{jj} = \sigma^2_{aaj}$ = the variance of the A×A value of the jth individual for i = 4, $G_i^{jj} = \sigma^2_{adj}$ = the variance of the A×D value of the jth individual for i = 5, $G_i^{jj} = \sigma^2_{ddj}$ = the variance of the D×D value of the jth individual for i = 6, $G_i^{jj} = \sigma^2_{aaaj}$ = the variance of the A×A×A value of the jth individual for i = 7, $G_i^{jj} = \sigma^2_{aadj}$ = the variance of the A×A×D value of the jth individual for i = 8, $G_i^{jj} = \sigma^2_{addj}$ = the variance of the A×D×D value of the jth individual for i = 9, and $G_i^{jj} = \sigma^2_{dddj}$ = the variance of the D×D×D value of the jth individual for i = 10. These genetic interpretations, along with those for intra- and inter-chromosome pairwise epistasis effects of Model-II under the QG and RE-QG models, are summarized in Supplementary Table S1.

Results and discussion

The multifactorial model of phenotypic values

Based on the RE-QG model of Eqs 6–14, the multifactorial model for phenotypic values is

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb} + \mathbf{Zg} + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}\sum\nolimits_{i=1}^{f}\mathbf{T}_i\boldsymbol{\tau}_i + \mathbf{e}\\ &= \mathbf{Xb} + \mathbf{Z}\sum\nolimits_{i=1}^{f}\mathbf{u}_i + \mathbf{e}\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbf{V} &= \mathbf{ZGZ}' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}\left(\sum\nolimits_{i=1}^{f}\mathbf{G}_i\right)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N\\ &= \mathbf{Z}\left(\sum\nolimits_{i=1}^{f}\sigma_i^2\mathbf{T}_i\mathbf{T}_i^{'}\right)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}\left(\sum\nolimits_{i=1}^{f}\sigma_i^2\mathbf{S}_i\right)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N\end{aligned} \tag{17}$$

where $\mathbf{y}$ = N×1 column vector of phenotypic observations, $\mathbf{Z}$ = N × n incidence matrix allocating phenotypic observations to each individual = identity matrix for one observation per individual (N = n), N = number of observations, n = number of individuals, $\mathbf{b}$ = c × l column vector of fixed effects such as heard-year-season in dairy cattle, c = number of fixed effects, $\mathbf{X}$ = N × c model matrix, $\mathbf{b,e}$ = N × 1 column vector of random residuals, σ_e^2 = residual variance, and $\mathbf{G} = \sum_{i=1}^{f}\mathbf{G}_i$ (Eq. 7). The phenotypic values ($\mathbf{y}$) are assumed to follow a normal distribution with mean $\mathbf{Xb}$ and variance–covariance matrix of $\mathbf{V}$. The methods described below for genomic estimation and prediction are based on the conditional expectation (CE) method, which is more efficient computationally than the methods based on mixed-model equations (MME) when the number of genetic effects is greater than the number of individuals (Da et al., 2014; Da, 2015).

For Model-I with 10 effect types, the genomic epistasis relationship matrices can be calculated using either EGERM or AGERM. However, EGERM or AGERM did not consider intra- and inter-chromosome genomic epistasis relationship matrices that are required by Model-II with 13 effect types. This research derives intra- and inter-chromosome genomic epistasis relationship matrices for both EGERM and AGERM.

Intra- and inter-chromosome genomic epistasis relationship matrices

The main derivation of the intra- and inter-chromosome genomic epistasis relationship matrices is the partition of the numerator of a genomic epistasis relationship matrix into intra- and inter-chromosome numerators. The first step is to derive the intra-chromosome numerator, and the second step is to derive the inter-chromosome numerator as the difference between the whole-genome numerator and the intra-chromosome numerator. The last step is to divide the intra-chromosome numerator by the average of the diagonal elements of the intra-chromosome numerator and to divide the inter-chromosome numerator by the average of the diagonal elements of the inter-chromosome numerator. Using this procedure, intra- and inter-chromosome epistasis relationship matrices were derived for both EGERM and AGERM (Supplementary Text S1).

Genomic best linear unbiased prediction and reliability

Based on the multifactorial genetic model of Eqs 16 and 17, the GBLUP of the genetic values of the ith effect type ($\hat{\mathbf{u}}_i$) and the best linear unbiased estimator (BLUE) or generalized least squares (GLS) estimator of fixed effect ($\hat{\mathbf{b}}$) are

$$\hat{\mathbf{u}}_i = \sigma_i^2\mathbf{S}_i\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) = \sigma_i^2\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{y},\ \ i = 1, \ldots, f \tag{18}$$

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{19}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$. The GBLUP of total genetic values of the n individuals is the summation of all types of genetic values:

$$\hat{\mathbf{g}} = \sum\nolimits_{i=1}^{f}\hat{\mathbf{u}}_i. \tag{20}$$

Reliability of GBLUP is the squared correlation between the GBLUP of a type of genetic value and the unobservable true genetic value being predicted by the GBLUP. The expected accuracy of predicting the genetic values by the GBLUP is the square root of reliability or the correlation between the GBLUP of a type of genetic effect and the unobservable true genetic effects being predicted by the GBLUP. In the absence of validation studies for observed prediction accuracy, reliability or the expected prediction accuracy is the measure of prediction accuracy of the GBLUP. The reliability of the GBLUP of the total genetic value (Eq. 2) of the jth individual is

$$R_{gj}^2 = \left[\mathbf{G}\left(\mathbf{Z}'\mathbf{P}\mathbf{Z}\right)\mathbf{G}\right]^{jj}/\mathbf{G}^{jj} \tag{21}$$

where $\mathbf{G} = \sum_{i=1}^{f}\mathbf{G_i} = \sum_{i=1}^{f}\sigma_i^2\mathbf{T}_i\mathbf{T}_i' = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i$ (Eq. 7), $\mathbf{G}^{jj} = \sum_{i=1}^{f}\mathbf{G}_i^{jj} = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i^{jj}$, and subscript or superscript jj denotes the jth diagonal element. The reliability formula for any or a combination of genetic values can be readily derived from Eq. 21, for example, the reliability of $\hat{\mathbf{u}}_3$ (GBLUP of haplotype additive values) is obtained from Eq. 21 by deleting all terms except $\mathbf{G}_3(\mathbf{Z}'\mathbf{P}\mathbf{Z})\mathbf{G}_3$ in the numerator and $\sigma_3^2\mathbf{S}_3^{jj}$ in the denominator, with changes in the $\mathbf{V}$ and $\mathbf{P}$ matrices accordingly.

Calculation of genomic best linear unbiased prediction and reliability for individuals with and without phenotypic observations separately

Two strategies are available for calculating GBLUP and the reliability of Eqs 20 and 21. Strategy-1 is a one-step strategy that includes all individuals with and without phenotypic observations in the same system of equations so that GBLUP and reliability are calculated simultaneously for all individuals. This strategy essentially augments the mixed model for individuals with phenotypic observations with a set of null equations consisting of "0"s but uses each genomic relationship matrix for all individuals, and these null equations and the use of the relationship matrix for all individuals do not affect the GBLUP, reliability, and heritability of individuals with phenotypic observations. The advantage of this one-step strategy is the simplicity of data preparation. For example, for a k-fold cross validation study, the phenotypic input file only needs to have k columns of the trait observations, with one column for each validation where the phenotypic observations for the validation individuals are set as "missing," and the **X** and **Z** model matrices for the "missing" observations are set to zero. With this strategy, the genotypic data need to be processed only once. As the number of traits increases for validation studies, this one-step strategy becomes more appealing due to the savings in data preparation work. This strategy has been implemented in our computing tools of GVCBLUP (Wang et al., 2014), GVCHAP (Prakapenka et al., 2020), and EPIHAP (Liang et al., 2021, 2022). However, when the number of validation individuals or individuals without phenotypic values is large, each genomic relationship matrix ($\mathbf{S}_i$ matrix) is large, and the one-step strategy becomes more difficult as the number of individuals increases.

For large numbers of individuals without phenotypic observations, calculating GBLUP for individuals with and without phenotypic values separately is more efficient computationally than calculating GBLUP for all individuals in the same system of equations by applying Henderson's BLUP for animals without phenotypic observations (Henderson, 1977) to GBLUP. Let n_1 = number of individuals with phenotypic observations, n_0 = number of individuals without phenotypic observations, $n = n_1 + n_0$, and let the $\mathbf{S}_i$ matrix be partitioned as

$$\mathbf{S}_i = \begin{bmatrix} \mathbf{S}_{i11} & \mathbf{S}_{i10} \\ \mathbf{S}_{i01} & \mathbf{S}_{i00} \end{bmatrix},\ \ i = 1, \ldots, f \tag{22}$$

where $\mathbf{S}_{i11} = n_1 \times n_1$ genomic relationship matrix of the genetic values of the ith effect type for individuals with phenotypic observations, $\mathbf{S}_{i01} = n_0 \times n_1$ = genomic relationship matrix of the genetic values of the ith effect type between individuals without phenotypic observations and individuals with phenotypic observations, $\mathbf{S}_{i10} = \mathbf{S}'_{i01} = n_1 \times n_0 =$ genomic relationship matrix between individuals with phenotypic observations and individuals without phenotypic observations, and $\mathbf{S}_{i00} = n_0 \times n_0$ genomic relationship matrix of the genetic values of the i^{th} effect type for individuals without phenotypic observations. In Eqs 16 and 17, $\mathbf{y} = \mathbf{y}_1$, the $\mathbf{Z}$ matrix needs to be changed to $\mathbf{Z} = [\,\mathbf{Z}_1 \;\; \mathbf{0}\,]$, the $\mathbf{u}_i$ vector partitioned as $\mathbf{u}_i = [\mathbf{u}'_{i1}, \mathbf{u}'_{i0}]'$, and the $\mathbf{g}$ vector partitioned as $\mathbf{g} = [\mathbf{g}'_1, \mathbf{g}'_0]'$, where $\mathbf{Z}_1 = N \times n_1$ incidence matrix allocating phenotypic observations to individuals with phenotypic observations, $\mathbf{0} = N \times n_0$ incidence matrix with elements "0" connecting phenotypic observations to individuals without phenotypic observations. With these changes and Eq. 22, the $\mathbf{V}$ matrix of Eq. (17) can be re-written as

$$\mathbf{V} = \mathbf{Z}_1\left(\sum_{i=1}^{f} \mathbf{G}_i\right)\mathbf{Z}'_1 + \sigma_e^2\mathbf{I}_N = \mathbf{Z}_1\left(\sum_{i=1}^{f} \sigma_i^2\mathbf{S}_{i11}\right)\mathbf{Z}'_1 + \sigma_e^2\mathbf{I}_N \quad (23)$$

and the GBLUP and reliability for individuals with and without phenotypic observations can be calculated as

$$\hat{\mathbf{u}}_{i1} = \sigma_i^2\mathbf{S}_{i11}\mathbf{Z}'_1\mathbf{V}^{-1}\left(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{b}}\right) = \sigma_i^2\mathbf{S}_{i11}\mathbf{Z}'_1\mathbf{P}\mathbf{y}_1, \;\; i = 1, \ldots, f \quad (24)$$

$$\hat{\mathbf{g}}_1 = \sum_{i=1}^{f} \hat{\mathbf{u}}_{i1} \quad (25)$$

$$R_{g1j}^2 = \left[\mathbf{G}_{11}\left(\mathbf{Z}'_1\mathbf{P}\mathbf{Z}'_1\right)\mathbf{G}_{11}\right]^{jj} / \mathbf{G}_{11}^{jj} \quad (26)$$

$$\hat{\mathbf{u}}_{i0} = \sigma_i^2\mathbf{S}_{i01}\mathbf{Z}'_1\mathbf{V}^{-1}\left(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{b}}\right) = \sigma_i^2\mathbf{S}_{i01}\mathbf{Z}'_1\mathbf{P}\mathbf{y}_1, \;\; i = 1, \ldots, f \quad (27)$$

$$= \sigma_i^2\mathbf{S}_{i01}\mathbf{S}_{i11}^{-1}\mathbf{S}_{i11}\mathbf{Z}'_1\mathbf{P}\mathbf{y}_1 = \mathbf{G}_{i01}\mathbf{G}_{i11}^{-1}\mathbf{G}_{i11}\mathbf{Z}'_1\mathbf{P}\mathbf{y}_1 = \mathbf{G}_{i01}\mathbf{G}_{i11}^{-1}\hat{\mathbf{u}}_{i1} \quad (28)$$

$$\hat{\mathbf{g}}_0 = \sum_{i=1}^{f} \hat{\mathbf{u}}_{i0} \quad (29)$$

$$R_{g0j}^2 = \left[\mathbf{G}_{01}\left(\mathbf{Z}'_1\mathbf{P}\mathbf{Z}'_1\right)\mathbf{G}_{10}\right]^{jj} / \mathbf{G}_{00}^{jj} \quad (30)$$

where $\hat{\mathbf{u}}_{i1} = n_1 \times 1$ column vector of the GBLUP of the genetic values of the i^{th} effect type for individuals with phenotypic observations, $\hat{\mathbf{g}}_1 = n_1 \times 1$ column vector of the GBLUP of the total genetic values for individuals with phenotypic observations, $R_{g1j}^2 =$ reliability for the j^{th}individuals with phenotypic observations, $\hat{\mathbf{u}}_{i0} = n_0 \times 1$ column vector of the GBLUP of the genetic values of the i^{th} effect type for individuals without phenotypic observations, $\hat{\mathbf{g}}_0 = n_0 \times 1$ column vector of the GBLUP of the total genetic values for individuals without phenotypic observations, $R_{g0j}^2 =$ reliability for the j^{th} individuals without phenotypic observations, $\mathbf{G}_{11} = \sum_{i=1}^{f}\mathbf{G}_{i11} = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_{i11}$, $\mathbf{G}_{01} = \sum_{i=1}^{f}\mathbf{G}_{i01} = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_{i01}$, $\mathbf{G}_{10} = \sum_{i=1}^{f}\mathbf{G}_{i10} = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_{i10}$, $\mathbf{G}_{11}^{jj} = \sum_{i=1}^{f}\mathbf{S}_{i11}^{jj}\sigma_i^2$, and $\mathbf{G}_{00}^{jj} = \sum_{i=1}^{f}\mathbf{S}_{i00}^{jj}\sigma_i^2$.

Equations 27 and 28 yield identical results if $\mathbf{S}_{i11}^{-1}$ exists. However, when the number of individuals is greater than the number of effect levels, such as the number of SNPs, $\mathbf{S}_{i11}^{-1}$ in Eq. 28 does not exist, and Eq. 27 still can calculate the GBLUP. The usefulness of Eq. 28 is that it shows the GBLUP of individuals without phenotypic observations is the regression of the genetic values of individuals without phenotypic observations on the genetic values of individuals with phenotypic observations. The advantage of Eq. 27 is that it does not calculate $\mathbf{S}_{i11}^{-1}$ and hence is unaffected by the singularity of $\mathbf{S}_{i11}$. Therefore, Eq. 27 is recommended for calculating GBLUP for individuals without phenotypic observation when the number of such individuals is large. The GBLUP calculations of Eqs 24, 27, and 28 do not involve the genomic relationship matrix among individuals without phenotypic observations $\mathbf{S}_{i00}$, which is much larger than $\mathbf{S}_{i11}$ when n_1 is much larger than n_0. The reliability calculation for individuals without phenotypic observations (Eq. 30) only uses the diagonal elements of $\mathbf{S}_{i00}$ and not the entire $\mathbf{S}_{i00}$.

Advantage of the integrated model over separate models

The multifactorial model of Eqs 16 and 17 integrating SNP, haplotype, and epistasis effects has the advantage of using more effect types and assessing each effect type based on the phenotypic values adjusted for all remaining effect types over separate models for SNP, haplotype, and epistasis effects that do not have a mechanism to adjust for effect types not in the model, and each uses a smaller number of genetic effects in the model.

This advantage of the multifactorial model assessing each effect type based on the phenotypic values adjusted for all remaining effect types can be shown using the MME version of the GBLUP for the i^{th} effect type:

$$\hat{\mathbf{u}}_i = \left(\mathbf{Z}'_i\mathbf{Z}_i + \mathbf{G}_i^{-1}\right)^{-1}\left[\mathbf{Z}'_i\mathbf{y} - \left(\mathbf{Z}'_i\mathbf{X}\hat{\mathbf{b}} + \sum_{\substack{j=1 \\ j \neq i}}^{f} \mathbf{Z}'_i\mathbf{Z}_j\hat{\mathbf{u}}_j\right)\right]$$

$$= \left(\mathbf{Z}'_i\mathbf{Z}_i + \mathbf{G}_i^{-1}\right)^{-1}\mathbf{Z}'_i\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_{\substack{j=1 \\ j \neq i}}^{f} \mathbf{Z}_j\hat{\mathbf{u}}_j\right) = \left(\mathbf{Z}'_i\mathbf{Z}_i + \mathbf{G}_i^{-1}\right)^{-1}\mathbf{Z}'_i\mathbf{y}_{bu}^{*} \quad (31)$$

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-}\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i\right)$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\left(\mathbf{y} - \sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y}_u^{*} \quad (32)$$

where $\mathbf{y}_{bu}^{*} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_{\substack{j=1 \\ j \neq i}}^{f} \mathbf{Z}_j\hat{\mathbf{u}}_j$ = phenotypic observations adjusted for the fixed effects and all random genetic values except those of $\hat{\mathbf{u}}_i$, $\mathbf{y}_u^{*} = \mathbf{y} - \sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i$ = phenotypic observations adjusted for all random genetic values, and $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Eq. 31 shows the MME version of $\hat{\mathbf{u}}_i$ uses the phenotypic values adjusted for the GBLUP of all other effect types in the model. Since the MME version of $\hat{\mathbf{u}}_i$ (Eq. 31) and $\hat{\mathbf{b}}$ (Eq. 32) are identical to the CE version of $\hat{\mathbf{u}}_i$ (Eq. 18) and $\hat{\mathbf{b}}$ (Eq. 19), the CE version of $\hat{\mathbf{u}}_i$ (Eq. 18) uses the phenotypic values adjusted for the GBLUP of all other effect types in the model even though the CE version does not do such adjustments explicitly.

Genomic restricted maximum estimation (GREML) of variances and heritabilities

The estimation of variance components uses GREML and a combination of EM-REML and AI-REML algorithms of iterative solutions. EM-REML is slow but converges, whereas AI-REML is fast but fails for zero heritability estimates. In our GVCBLUP, GVCHAP, and EPIHAP computing packages that implement these two algorithms (Wang et al., 2014; Prakapenka et al., 2020; Liang et al., 2021), EM-REML is used automatically when AI-REML fails. The EM-REML iterative algorithm for the multifactorial model of Eqs 16 and 17 is

$$\sigma_i^{2(j+1)} = \sigma_i^{2(j)}\mathbf{yP}^{(j)}\mathbf{ZS}_i\mathbf{Z}'\mathbf{P}^{(j)}\mathbf{y}/\mathrm{tr}\left(\mathbf{P}^{(j)}\mathbf{ZS}_i\mathbf{Z}'\right),\ \ i = 1, \ldots, f \tag{33}$$

$$\sigma_e^{2(j+1)} = \sigma_e^{2(j)}\mathbf{yP}^{(k)}\mathbf{P}^{(j)}\mathbf{y}/\mathrm{tr}\left(\mathbf{P}^{(j)}\right) \tag{34}$$

where j = iteration number. The AI-REML iterative algorithm is an extension of the early formulations (Johnson and Thompson, 1995; Lee and van der Werf, 2006) to the multifactorial model of Eqs 16 and 17:

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \left(\mathbf{AI}^{(j)}\right)^{-1}\Delta^{(j)} \tag{35}$$

where $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, ..., \sigma_f^2, \sigma_{f+1}^2)' = (f+1) \times 1$ column vector of variance–covariance components, $\sigma_{f+1}^2 = \sigma_e^2$ = residual variance, $\Delta = (\Delta_1, \Delta_2, ..., \Delta_f, \Delta_{f+1})' = (f+1) \times 1$ column vector of the partial derivatives of the log residual likelihood function with respect to each variance component, and j = iteration number. A typical term in Δ (Δ_i) and a typical term in **AI** (AI_{ik}) are

$$\begin{aligned}\Delta_i &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\sigma_i^2}\right) + \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial\mathbf{V}}{\partial\sigma_i^2}\mathbf{Py}\\ &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{PZS}_i\mathbf{Z}'\right) + \frac{1}{2}\mathbf{y}'\mathbf{PZS}_i\mathbf{Z}'\mathbf{Py}, \quad i = 1, ..., f+1\end{aligned} \tag{36}$$

$$\begin{aligned}AI_{ik} &= \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial\mathbf{V}}{\partial\sigma_i^2}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\sigma_k^2}\mathbf{Py}\\ &= \frac{1}{2}\mathbf{y}'\mathbf{PZ}'\mathbf{S}_i\mathbf{Z}'\mathbf{PZS}_k\mathbf{Z}'\mathbf{Py}, \qquad i, k = 1, \ldots, f+1\end{aligned} \tag{37}$$

where $\mathbf{S}_{f+1} = \mathbf{I}_N$. For the full Model-I or Model-II, some effect types inevitably may have zero variances. In those cases, AI-REML (Equations 35–37) fails and EM-REML (Equations 33 and 34) still converges, although a slow convergence rate can be expected for the full Model-I or Model-II. Once the effect types with zero variances are removed from the model, AI-REML converges, and a fast convergence rate can be expected. The estimate of the genomic heritability for each type of genetic effects (h_i^2) and the total heritability of all types of genetic effects (H^2) are

$$h_i^2 = \sigma_i^2/\sigma_y^2 \quad i = 1, \ldots, f \tag{38}$$

$$H^2 = \sum\nolimits_{i=1}^{f} h_i^2 \tag{39}$$

where $\sigma_y^2 = \sum_{i=1}^{f}\sigma_i^2 + \sigma_e^2$ = phenotypic variance.

The heritability estimates of Eq. 38 can be used for model selection by removing effect types with heritability estimates below a user-determined threshold value from the prediction model. Since different traits may have different genetic architectures, we hypothesize that some traits may involve only a small number of the effect types, and some traits are more complex and involve more effect types; global epistasis may be more important than local high-order epistasis effects of haplotypes for some traits, whereas the reverse may be true for other traits, and some traits may be affected by both global high-order and local high-order epistasis effects. The heritability estimates from Eq. 37 provide an approach to evaluate these hypotheses and identify effect types relevant to the phenotypic variance, whereas the total heritability of Eq. 38 provides an estimate of the total genetic contribution to the phenotypic variance. In addition to the use of heritability estimates, prediction accuracy based on GBLUP can be used for model selection by requiring a threshold accuracy level for the effect type to be included in the prediction model, for example, we identified the A + A×A model to have the same accuracy of predicting the phenotypic values of daughter pregnancy rate as the full Model-I in U.S. Holstein cows (Liang et al., 2022).

Estimation of pairwise epistasis effect and heritability

The heritability of an SNP, haplotype block, or pairwise epistasis effect is the contribution of the genetic effect to the phenotypic variance and is also the contribution to the heritability of the effect type, and is estimated through the GBLUP of the corresponding genetic effects. These heritability estimates can be used to identify genome locations with large contributions to the phenotypic variance. The estimation of pairwise epistasis effects and heritability is the most demanding computation because the pairwise epistasis model matrices must be creased and are no longer avoidable. Estimating the effects and heritabilities for third-order epistasis effects is computationally unfeasible and is not considered. GBLUPs of SNP, haplotype, and pairwise epistasis effects of Model-I (Supplementary Table S1) are calculated as

$$\hat{\boldsymbol{\tau}}_i = \sigma_i^2\mathbf{T}_i'\mathbf{Z}'\mathbf{Py} = \mathbf{T}_i'\mathbf{S}_i^{-1}\hat{\mathbf{u}}_i \tag{40}$$

where $\hat{\boldsymbol{\tau}}_i$ is the m × 1 column vector of SNP additive effects for i = 1, or SNP dominance effects for i = 2; or b × 1 column vector of haplotype additive effects for i = 3; or $\binom{m}{2} \times 1$ column vector of A×A epistasis effects for i = 4, or $2\binom{m}{2} \times 1$ column vector of A×D epistasis effects for i = 5, or $\binom{m}{2} \times 1$ column vector of D×D epistasis effects for i = 6. For i = 5; the order of A×D and D×A effects is determined by the order of the model matrices of those effects, that is, $\hat{\boldsymbol{\tau}}_5 = (\hat{\boldsymbol{\tau}}_{\alpha\delta}{}', \hat{\boldsymbol{\tau}}_{\delta\alpha}{}')'$ if $\mathbf{T}_5 = (\mathbf{T}_{\alpha\delta}, \mathbf{T}_{\delta\alpha})$, or $\hat{\boldsymbol{\tau}}_5 = (\hat{\boldsymbol{\tau}}_{\delta\alpha}{}', \hat{\boldsymbol{\tau}}_{\alpha\delta}{}')'$ if $\mathbf{T}_5 = (\mathbf{T}_{\delta\alpha}, \mathbf{T}_{\alpha\delta})$. The heritability of the jth effect

TABLE 1 Genomic heritability estimates of additive, dominance, and epistasis effects up to the third-order for five traits in a swine population.

	Trait				
	T1	T2	T3	T4	T5
Effect	Exact genomic epistasis relationship matrices (EGERM)				
A	0.023	0.217	0.131	0.336	0.366
D	0.000	0.013	0.000	0.000	0.052
A×A	0.046	0.186	0.278	0.017	0.054
A×D	0.000	0.000	0.091	0.000	0.000
D×D	0.000	0.000	0.091	0.000	0.000
A×A×A	0.000	0.000	0.000	0.000	0.000
A×A×D	0.000	0.000	0.079	0.000	0.000
A×D×D	0.000	0.000	0.102	0.000	0.000
D×D×D	0.000	0.000	0.117	0.000	0.000
Total heritability	0.069	0.416	0.889	0.354	0.471
Effect	Approximate genomic epistasis relationship matrices (AGERM)				
A	0.022	0.215	0.139	0.329	0.360
D	0.000	0.013	0.000	0.000	0.051
A×A	0.043	0.176	*0.280*	0.016	0.050
A×D	0.000	0.000	0.091	0.000	0.000
D×D	0.000	0.000	0.090	0.000	0.000
A×A×A	0.000	0.000	0.000	0.000	0.000
A×A×D	0.000	0.000	0.075	0.000	0.000
A×D×D	0.000	0.000	0.095	0.000	0.000
D×D×D	0.000	0.000	0.109	0.000	0.000
Total heritability	0.065	0.404	0.879	0.346	0.461

TABLE 2 Observed prediction accuracy of epistasis models relative to the additive model for five traits in a swine population.

	Trait				
	T1	T2	T3	T4	T5
Prediction accuracy of SNP model					
A	0.066	0.495	0.326	0.468	0.493
A + D	0.056	0.495	0.326	0.468	0.496
Epistasis model	A + AA	A + D + AA	A + AA + AD + DD+		
AAD + ADD + DDD	A + AA	A + D + AA			
EGERM					
Prediction accuracy	0.063	0.498	0.336	0.468	0.497
Accuracy increase (%)	−4.545	0.606	3.067	0.000	0.202
AGERM					
Prediction accuracy	0.063	0.498	0.336	0.468	0.497
Accuracy increase (%)	−4.545	0.606	3.067	0.000	0.202

"Prediction accuracy" is the observed prediction accuracy calculated as the correlation between the GBLUP of genotypic values and the phenotypic values in each validation population and then averaged over all 10 validation populations. "Accuracy increase" is the percentage increase of the observed prediction accuracy of the epistasis model over the observed prediction accuracy of the best SNP model, which was the additive model (A) for T1–T4 and the A + D model for T5. A = additive effects, D = dominance effects, AA = A×A effects, AD = A×D effects, DD = D×D effects, AAA = A×A×A effects, AAD = A×A×D effects, ADD = A×D×D dominance effects, and DDD = D×D×D dominance effects.

TABLE 3 Computing time (in seconds) for the construction of exact and approximate genomic epistasis relationship matrices for a swine population with 3,534 pigs and 52,843 SNPs using 20 threads of the Mangi supercomputer of the Minnesota Supercomputer Institute at the University of Minnesota.

Genomic epistasis relationship matrices	Pairwise	Third-order	Fourth-order
EGERM	666	796	1,256
AGERM	70	96	133
EGERM/AGERM	9.51	8.29	9.44

of the ith effect type ($\hat{h}^2_{ij}$) is estimated as a faction of the genomic heritability of the ith effect type ($\hat{h}^2_i$):

$$\hat{h}^2_{ij} = \left(\hat{\tau}^2_{ij}\Big/\sum\nolimits_{i=1}^{m}\hat{\tau}^2_{ij}\right)\hat{h}^2_i = \left(\hat{\tau}^2_{ij}\Big/\hat{\boldsymbol{\tau}}'_i\hat{\boldsymbol{\tau}}_i\right)\hat{h}^2_i = \hat{\sigma}^2_{ij}\Big/\hat{\sigma}^2_y \tag{41}$$

where $\hat{\boldsymbol{\tau}}_{ij}$ = the jth effect of $\hat{\boldsymbol{\tau}}_i$, $\hat{\sigma}^2_i$ = estimated variance of the ith effect type, $\hat{\sigma}^2_{ij}$ = estimated variance of the jth effect of the ith effect type, and $\hat{h}^2_i$ = genomic heritability of the ith effect type defined by Equation (52). For proving Equation 57, $\hat{\sigma}^2_i$ and $\hat{\sigma}^2_{ij}$ can be formulated based on the method of mixed-model equations (MME):

$$\begin{aligned}\hat{\sigma}^2_i &= \hat{\boldsymbol{\tau}}_i{}'\hat{\boldsymbol{\tau}}_i/[m_i-, \mathrm{tr}(\mathbf{C}^{ii})\lambda_i] = \sum\nolimits_{j=1}^{m_i}\tau^2_{ij}/[m_i - \mathrm{tr}(\mathbf{C}^{ii})\lambda_i] \\ &= \sum\nolimits_{j=1}^{m_i}\hat{\sigma}^2_{ij}\end{aligned} \tag{42}$$

$$\hat{\sigma}^2_{ij} = \hat{\boldsymbol{\tau}}^2_{ij}/[m_i - \mathrm{tr}(\mathbf{C}^{ii})\lambda_i] \tag{43}$$

where $\mathbf{C}^{ii}$ is the submatrix in the inverse or generalized inverse of the coefficient matrix of the MME corresponding to the ith effect type, m_i = number of effects of the ith effect type, and $\lambda_i = \hat{\sigma}^2_e/\hat{\sigma}^2_i$. Dividing Eq. 43 by $\hat{\sigma}^2_y$ and multiplying by $\hat{\sigma}^2_i/\hat{\sigma}^2_i$ yield Eq. 41:

$$\begin{aligned}\hat{h}^2_{ij} &= \left(\hat{\sigma}^2_{ij}\big/\hat{\sigma}^2_i\right)\left(\hat{\sigma}^2_i\big/\hat{\sigma}^2_y\right) = \left(\hat{\sigma}^2_{ij}\big/\hat{\sigma}^2_i\right)\left(\hat{\sigma}^2_i\big/\hat{\sigma}^2_y\right) = \left(\hat{\tau}^2_{ij}\Big/\sum\nolimits_{i=1}^{m}\hat{\tau}^2_{ij}\right)\hat{h}^2_i \\ &= \left(\hat{\tau}^2_{ij}\big/\hat{\boldsymbol{\tau}}'_i\hat{\boldsymbol{\tau}}_i\right)\hat{h}^2_i = \left(\hat{\sigma}^2_{ij}\big/\hat{\sigma}^2_y\right).\end{aligned}$$

It is readily seen that the sum of all heritability estimates of the ith effect type is the genomic heritability of the ith effect type: $\sum_{i=1}^{m_i}\hat{h}^2_{ij} = \hat{h}^2_i$. It is of note that Eqs 42 and 43 using MME are only for proving Eq. 41. The MME method is computationally prohibitive for estimating genetic effects and their variances under the multifactorial model, although the MME method yields results identical to the CE method, which is computationally feasible for genomic estimation and prediction under the multifactorial model.

Comparison between exact and approximate genomic epistasis relationship matrices

We evaluated the differences between AGERM and EGERM in genomic heritability estimates and prediction accuracies using a publicly available swine genomics data set that had 3,534 animals from a single PIC nucleus pig line with five anonymous traits and 52,842 genotyped and imputed autosome SNPs after filtering by requiring minor allele frequency (MAF) > 0.001 and proportion of missing SNP genotypes < 0.100 (Cleveland et al., 2012). The EGERM followed the method used by Jiang and Reif (2020), and the AGERM methods are described in Supplementary Text S1. The heritability results showed that EGERM had slightly higher heritability estimates than AGERM except for the A×A heritability of T3, where AGERM had a slightly higher estimate than EGERM (0.280 vs. 0.278, Table 1). From Table 1, effect type with nonzero heritability estimates was included in the prediction model for evaluating the observed prediction accuracy as the correlation between the GBLUP of genotypic values and the phenotypic values in each validation population and then averaged over all 10 validation populations. The results showed that AGERM and EGERM had the same prediction accuracy for this swine sample (Table 2). A disadvantage of EGERM is the computing time for the construction of EGERM, about 9.51 times as much time for pairwise relationship matrices, 8.29 as much time for third-order and 9.44 times as much time for fourth-order as required for AGERM (Table 3). However, computing time is not the deciding factor for choosing between the exact and approximate methods because the multi-node approach that calculates each genomic relationship matrix in pieces and adds those pieces together can reduce the computing time to an acceptable level when multiple threads/cores are available, and the two-step strategy can be used so that each genomic relationship is calculated only once for different traits and validation populations (Prakapenka et al., 2020). Prediction accuracy is the ultimate deciding factor in choosing between different methods. We reported results of comparing AGERM and EGERM using 60,671 SNPs and 22,022 first-lactation Holstein cows with phenotypic observations of daughter pregnancy rates, showing that AGERM and EGERM had the same heritability estimates and prediction accuracy, but EGERM required 21 times as much computing time as that required by AGERM, which required 1.32 times as much time for the genomic additive relationship matrix (Liang et al., 2022). The combined results of the swine and Holstein samples indicated that EGERM and AGERM had similar results and that the computing difficulty of EGERM over AGERM increased rapidly as the sample size increased. Given the computing difficulty of

EGERM and the negligible differences between EGERM and AGERM in prediction accuracy, AGERM should be favored for its mathematical simplicity and computing efficiency, at least for samples with 50,000 SNPs or more.

Numerical demonstration

The methods of genomic epistasis relationship matrices based on the additive and dominance model matrices, GREML, GBLUP and reliability, and estimation of effect heritability are demonstrated using an R program (DEMO.R) and a small artificial sample for the convenience of reading the numerical results (Supplementary Text S2 and R program). Because of the artificial nature and the extremely small sample size, this numerical demonstration does not have any genetic and methodology implications and is for showing calculations of the methods only. This R program is an extension of the R demo program of GVCHAP that integrates SNP and haplotype effects and has a computing pipeline for producing the input haplotype data from the SNP data (Prakapenka et al., 2020).

Conclusion

The multifactorial methods with SNP, haplotype, and epistasis effects up to the third-order provide an approach to investigate the contributions of global low-order and local high-order epistasis effects to the phenotypic variance and the accuracy of genomic prediction. Genomic heritability of each effect type from GREML and prediction accuracy from validation studies using GBLUP can be used jointly to identify effect types contributing to the phenotypic variance and the accuracy of genomic prediction, and the GBLUP for the multifactorial model with selected effect type can be used for genomic evaluation. With many capabilities, including the use of intra- and inter-chromosome separately, the multifactorial methods offer a significant methodology capability to investigate and utilize complex genetic mechanisms for genomic prediction and for understanding the complex genome–phenome relationships.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: https://academic.oup.com/g3journal/article/2/4/429/6026060#supplementary-data.

Author contributions

YD conceived this study and derived the formulations. ZL contributed to formulations of the epistasis genomic relationships, implemented the epistasis methods in EPIHAP, and validated and evaluated the methods. DP contributed to the data processing for methodology evaluation. YD and ZL prepared the manuscript.

Funding

This research was supported by the National Institutes of Health's National Human Genome Research Institute, grant R01HG012425, as part of the NSF/NIH Enabling Discovery through Genomics (EDGE) Program; grant 2020-67015-31133 from the USDA National Institute of Food and Agriculture; and project MIN-16-124 of the Agricultural Experiment Station at the University of Minnesota. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The use of the Ceres and Atlas computers in this research was supported by USDA-ARS projects 8042-31000-002-00-D and 8042-31000-001-00-D.

Acknowledgments

The supercomputer of Minnesota Supercomputer Institute at the University of Minnesota and the Ceres and Atlas high-performance computing system of USDA-ARS were used for the evaluation and testing of the methods and EPIHAP computing package. The authors thank Paul VanRaden, Steven Schroeder, and Ransom Baldwin for help with the use of the USDA-ARS computing facilities.

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.2022.922369/full#supplementary-material

References

Bian, C., Prakapenka, D., Tan, C., Yang, R., Zhu, D., Guo, X., et al. (2021). Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. *Genet. Sel. Evol.* 53 (1), 78–19. doi:10.1186/s12711-021-00661-y

Carlborg, Ö., and Haley, C. S. (2004). Epistasis: Too often neglected in complex trait studies? *Nat. Rev. Genet.* 5 (8), 618–625. doi:10.1038/nrg1407

Cleveland, M. A., Hickey, J. M., and Forni, S. (2012). A common dataset for genomic analysis of livestock populations. *G3* 2 (4), 429–435. doi:10.1534/g3.111.001453

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39 (6), 859–882. doi:10.1093/genetics/39.6.859

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* 11 (20), 2463–2468. doi:10.1093/hmg/11.20.2463

Da, Y. (2015). Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genet.* 16 (1), 144. doi:10.1186/s12863-015-0301-1

Da, Y., Tan, C., and Parakapenka, D. (2016). 0336 Joint SNP-haplotype analysis for genomic selection based on the invariance property of GBLUP and GREML to duplicate SNPs. *J. Animal Sci.* 94 (5), 161–162. doi:10.2527/jam2016-0336

Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* 9 (1), e87666. doi:10.1371/journal.pone.0087666

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53 (11), 876–883. doi:10.1139/G10-076

Henderson, C. (1977). Best linear unbiased prediction of breeding values not in the model for records. *J. Dairy Sci.* 60 (5), 783–787. doi:10.3168/jds.s0022-0302(77)83935-0

Henderson, C. (1985). Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *J. Animal Sci.* 60 (1), 111–117. doi:10.2527/jas1985.601111x

Jiang, Y., and Reif, J. C. (2020). Efficient algorithms for calculating epistatic genomic relationship matrices. *Genetics* 216 (3), 651–669. doi:10.1534/genetics.120.303459

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201 (2), 759–768. doi:10.1534/genetics.115.177907

Johnson, D., and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* 78 (2), 449–456. doi:10.3168/jds.s0022-0302(95)76654-1

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. B Biol. Sci.* 143 (910), 102–113.

Lee, S. H., and van der Werf, J. H. (2006). An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* 38 (1), 25–43. doi:10.1186/1297-9686-38-1-25

Liang, Z., Prakapenka, D., and Da, Y. (2022). Comparison of two methods of genomic epistasis relationship matrices using daughter pregnancy rate in U.S. Holstein cattle. Abstract 2466V, page 409 of ADSA2022 Abstracts. Available at: https://www.adsa.org/Portals/0/SiteContent/Docs/Meetings/2022ADSA/Abstracts_BOOK_2022.pdf?v=20220613 (Accessed September 27, 2022).

Liang, Z., Prakapenka, D., and Da, Y. (2021). Epihap: A computing tool for genomic estimation and prediction using global epistasis effects and haplotype effects. Abstract P167, page 223 of ADSA2021 Abstracts, ADSA 2021 Virtual Annual Meeting. Available at: https://www.adsa.org/Portals/0/SiteContent/Docs/Meetings/2021ADSA/ADSA2021_Abstracts.pdf (Accessed September 27, 2022).

Liang, Z., Tan, C., Prakapenka, D., Ma, L., and Da, Y. (2020). Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. *Front. Genet.* 11 (1461), 588907. doi:10.3389/fgene.2020.588907

Mackay, T. F. (2014). Epistasis and quantitative traits: Using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* 15 (1), 22–33. doi:10.1038/nrg3627

Martini, J. W., Toledo, F. H., and Crossa, J. (2020). On the approximation of interaction effect models by Hadamard powers of the additive genomic relationship. *Theor. Popul. Biol.* 132, 16–23. doi:10.1016/j.tpb.2020.01.004

Martini, J. W., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: How gene interaction translates into genomic relationship. *Theor. Appl. Genet.* 129 (5), 963–976. doi:10.1007/s00122-016-2675-5

Muñoz, P. R., Resende, M. F., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198 (4), 1759–1768. doi:10.1534/genetics.114.171322

Phillips, P. C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9 (11), 855–867. doi:10.1038/nrg2452

Prakapenka, D., Liang, Z., Jiang, J., Ma, L., and Da, Y. (2021). A Large-scale genome-wide association study of epistasis effects of production traits and daughter pregnancy rate in US Holstein cattle. *Genes* 12 (7), 1089. doi:10.3390/genes12071089

Prakapenka, D., Wang, C., Liang, Z., Bian, C., Tan, C., and Da, Y. (2020). Gvchap: A computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. *Front. Genet.* 11, 282. doi:10.3389/fgene.2020.00282

Ritchie, M. D., and Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: Challenges in abundance of methods, practical considerations, and biological interpretation. *Ann. Transl. Med.* 6 (8), 157. doi:10.21037/atm.2018.04.05

Segre, D., DeLuna, A., Church, G. M., and Kishony, R. (2005). Modular epistasis in yeast metabolism. *Nat. Genet.* 37 (1), 77–83. doi:10.1038/ng1489

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PloS One* 7 (9), e45293. doi:10.1371/journal.pone.0045293

Tan, C., Wu, Z., Ren, J., Huang, Z., Liu, D., He, X., et al. (2017). Genome-wide association study and accuracy of genomic prediction for teat number in Duroc pigs using genotyping-by-sequencing. *Genet. Sel. Evol.* 49 (1), 35. doi:10.1186/s12711-017-0311-8

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206 (3), 1297–1307. doi:10.1534/genetics.116.199406

Wang, C., and Da, Y. (2014). Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. *PLoS One* 9 (12), e114484. doi:10.1371/journal.pone.0114484

Wang, C., Prakapenka, D., Wang, S., Pulugurta, S., Runesha, H. B., and Da, Y. (2014). Gvcblup: A computer package for genomic prediction and variance component estimation of additive and dominance effects. *BMC Bioinforma.* 15 (1), 270. doi:10.1186/1471-2105-15-270

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/fgene.2022.1070783

OPEN ACCESS

EDITED BY
Li Ma,
University of Maryland, College Park,
United States

REVIEWED BY
Guiguiggbaza-Kossigan Dayo,
Centre International de Recherche-Développement sur l'Elevage en Zone Subhumide, Burkina Faso
Kefei Chen,
Curtin University, Australia

*CORRESPONDENCE
Mirte Bosse,
✉ mirte.bosse@wur.nl

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

RECEIVED 15 October 2022
ACCEPTED 19 December 2022
PUBLISHED 13 January 2023

CITATION
Peng Y, Derks MFL, Groenen MAM, Zhao Y and Bosse M (2023), Distinct traces of mixed ancestry in western commercial pig genomes following gene flow from Chinese indigenous breeds.
Front. Genet. 13:1070783.
doi: 10.3389/fgene.2022.1070783

Distinct traces of mixed ancestry in western commercial pig genomes following gene flow from Chinese indigenous breeds

Yebo Peng[1], Martijn FL Derks[2,3], Martien AM Groenen[2], Yiqiang Zhao[1] and Mirte Bosse[2,4]*

[1]State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China, [2]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, Netherlands, [3]Topigs Norsvin Research Center, Beuningen, Netherlands, [4]Amsterdam Insitute of Life and Environment (A-Life), VU University Amsterdam, Amsterdam, Netherlands

Studying gene flow between different livestock breeds will benefit the discovery of genes related to production traits and provide insight into human historical breeding. Chinese pigs have played an indispensable role in the breeding of Western commercial pigs. However, the differences in the timing and volume of the contribution of pigs from different Chinese regions to Western pigs are not yet apparent. In this paper, we combine the whole-genome sequencing data of 592 pigs from different studies and illustrate patterns of gene flow from Chinese pigs into Western commercial pigs. We describe introgression patterns from four distinct Chinese indigenous groups into five Western commercial groups. There were considerable differences in the number and length of the putative introgressed segments from Chinese pig groups that contributed to Western commercial pig breeds. The contribution of pigs from different Chinese geographical locations to a given western commercial breed varied more than that from a specific Chinese pig group to different Western commercial breeds, implying admixture within Europe after introgression. Within different Western commercial lines from the same breed, the introgression patterns from a given Chinese pig group seemed highly conserved, suggesting that introgression of Chinese pigs into Western commercial pig breeds mainly occurred at an early stage of breed formation. Finally, based on analyses of introgression signals, allele frequencies, and selection footprints, we identified a ~2.65 Mb Chinese-derived haplotype under selection in Duroc pigs (CHR14: 95.68–98.33 Mb). Functional and phenotypic studies demonstrate that this *PRKG1* haplotype is related to backfat and loin depth in Duroc pigs. Overall, we demonstrate that the introgression history of domestic pigs is complex and that Western commercial pigs contain distinct traces of mixed ancestry, likely derived from various Chinese pig breeds.

KEYWORDS

introgression, hybridization, selection, commercial pigs, gene flow

1 Introduction

Introgression and hybridization played a distinct role in the evolutionary diversification of plants and animals (Dowling and Secor, 1997; Mallet, 2005; Arnold et al., 2008; Stukenbrock, 2016; Grant and Grant, 2019). Genetic material introgressed from sister lineages has often been adaptive in plant and animal evolution (Dowling et al., 2016; Burgarella et al., 2019; Janzen et al., 2019; Cao et al., 2021). In wild animals and plants, adaptive introgression played an essential role in disease resistance and environmental adaptation. Examples entail introgression in *P. trichocarpa* (Suarez-Gonzalez et al., 2016; Suarez-Gonzalez et al., 2018), *Zea mays* (Hufford et al., 2013), and sheep (Cao et al., 2021). Sometimes morphological characteristics changed, for example, wing patterning in *Heliconius* butterflies (Pardo-Diaz et al., 2012; Enciso-Romero et al., 2017). In modern humans, a variant of the *EPAS1* gene was introduced from Denisovans into Tibetans, which has proven beneficial to the adaptation of Tibetans to high altitudes (Huerta-Sanchez et al., 2014; Zhang W. et al., 2020). However, introgressed haplotypes can also have adverse effects. Examples are risk factors for type 2 diabetes, lupus, biliary cirrhosis (Sankararaman et al., 2014), and even COVID-19 inherited from Neanderthals (Zeberg and Pääbo, 2020).

Human activities have impacted over 75% of the global land area over the past ten thousand years (Venter et al., 2016; Bullock et al., 2018). Domestication and dispersal of pets, plants, and livestock have strongly altered the worldwide distribution of flora and fauna (Wichmann et al., 2009; Ottoni et al., 2013; Koch et al., 2015; Bullock et al., 2018). During the first industrial revolution, humans deliberately promoted crossbreeding of local animal and plant breeds to accelerate the process of breeding. Human-mediated hybridization between different breeds has been an important factor in shaping domestic plants and animals' genomic and phenotypic diversity (Larson and Burger, 2013; Meng et al., 2018). The hybridization from bovine ancestors improved Mongolian yak management and breeding (Medugorac et al., 2017). Likewise, haplotypes introgressed from Holstein and Brown Swiss affect protein and fat content of milk, calving traits, body conformation, feed efficiency, carcass, and fertility traits (Zhang et al., 2018).

Pigs have a long history of admixture. In the genus *Sus*, post-divergence interspecific admixture occurred before the domestication of *Sus scrofa* (Frantz L. A. F. et al., 2013; Frantz et al., 2014; Frantz et al., 2016; Liu et al., 2019). For *Sus. scrofa*, *Sus. cebifrons*, and *Sus. verrucosus*, around 23% of their genomes have been affected by admixture during the later Pleistocene climatic transition (Frantz et al., 2014). Gene flow also happened extensively between domesticated pigs to their wild ancestors during the domestication process (Giuffra et al., 2000; Frantz A. C. et al., 2013; Zhu et al., 2017). Hybridization between China and Western animals may date back to the 1st—fourth century AD (Wang et al., 2011). Historical records report that Chinese pigs were repeatedly introduced into Europe to improve the local pig breeds from the 18th century (Giuffra et al., 2000; Wang et al., 2011; White, 2011), followed by introduction into America from the 19th century onwards (Wang et al., 2011; White, 2011). Vice versa, Western commercial pigs were introduced into China since the start of the 20th century (Wang et al., 2011; White, 2011). The complex hybrid history between China and Western pigs has shaped the present genomic landscape in pigs.

There are 118 native pig breeds in China (Megens et al., 2007) with diverse phenotypic characteristics. Characteristic for Eastern Chinese pigs is early sexual maturity, higher ovulation number, and higher litters size (>15 for some breeds) (Wang et al., 2011). South Chinese pigs have inferior reproductive performance (8–10 piglets per parity for Luchan pigs), thinner skin, and excellent heat resistance (Wang et al., 2011; Chen et al., 2020).

In recent years, genomic studies have revealed some Chinese haplotypes in Western pig breeds that were likely introgressed and selected for. *AHR* is a toxicity- and fertility-related gene (Denison et al., 2011; Onteru et al., 2012). It is introgressed from a Chinese breed into Dutch Large White pigs (Bosse et al., 2014). Based on the Illumina Porcine 60 K SNP Beadchip dataset of Erhualian, White Duroc × Erhualian F2 population, Duroc and Landrace pigs. Yang et al. found a mutation in *VRTN* that increased vertebra number, carcass length, and teat number in Western pigs and was inherited from Chinese Erhualian pigs (Yang et al., 2016). The meat quality-related genes (*SAL1*, *ME1*) and fertility-related genes (*GNRHR*, *GNRH1*), are reported as being introgressed into Duroc from Meishan pigs by Zhao et al., using whole-genome re-sequencing data of 32 Chinese Meishan and 31 Duroc pigs (Zhao et al., 2018). Recently, Chen et al. also explored whole-genome sequencing data from 266 Eurasian wild boars and domestic pigs. They found that the *GOLM1-NAA35*, a gene that is responsible for cytokine interleukin 6 (IL-6) production in human immune cells (Li et al., 2016), is inherited from south Chinese pigs (SCN) in French Large White (LWHFR) (Chen et al., 2020). They also found a haplotype spanning *KATNAL1* that originated from east Chinese pigs (ECN) pigs and has been selected to increase the fertility in LWHFR pigs. Although only LWHFR and two Chinese native pig groups were included, their study provided the novel perspective that introgression from Chinese pigs to commercial breeds may vary considerably. Therefore, in the current study we extensively explore source of introgression and genomic regions that contained introgressed segments on a large scale, including multiple Western pig breeds and a broad sampling of Asian breeds.

Thus, many genomic segments from local Chinese pigs that contributed to favorable characteristics of Western commercial breeds have been identified (Bosse et al., 2014; Frantz et al., 2015; Chen et al., 2017; Chen et al., 2020; Wang et al., 2020), but these records are sporadic, and no systematic survey has been conducted. How extensive these episodes of introgression and improvement of Western domesticated pigs with animals from Asia have been, and where in China these pigs originated, are still unanswered questions. Although Chinese pigs are highly polymorphic (Amaral et al., 2008; Frantz A. C. et al., 2013; Zhao et al., 2019), they form a close genetic group, and there has been an extensive genetic exchange between (local) breeds (Huang et al., 2020). Disentangling the sources of the introgressed haplotypes will shed new light on historical breeding practices, help understand the molecular mechanisms underlying phenotype change, and be of great significance to future breeding.

Even though the overall level of introgression from Chinese pigs to Western commercial breeds seems relatively stable across breeds, the underlying haplotypes, genomic loci, and breed origins may vary (Bosse et al., 2014). In this paper, we present a comprehensive study of the gene flow of pigs from different Chinese origins into five distinct Western commercial lines and illustrate the difference of global haplotype introgression patterns between donor-recipient combinations.

2 Materials and methods

2.1 SNP calling, phasing, and imputation

The datasets analyzed during the current study are available from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) under project PRJEB1683 (Groenen et al., 2012), PRJEB29465 (Grahofer et al., 2019), PRJEB9922 (Frantz et al., 2015), PRJNA186497 (Li et al., 2013), PRJNA213179 (Ai et al., 2015), PRJNA231897, PRJNA238851 (Wang et al., 2015), PRJNA254936, PRJNA255085 (Ramirez et al., 2015), PRJNA260763 (Choi et al., 2015), PRJNA273907, PRJNA305081, PRJNA305975, PRJNA309108 (Li et al., 2017), PRJNA314580, PRJNA320525 (Bianco et al., 2015), PRJNA320526, PRJNA320527, PRJNA322309, PRJNA369600, PRJNA378496 (Zhao et al., 2018), PRJNA398176 (Zhu et al., 2017), PRJNA438040, PRJNA488327 (Yan et al., 2018), PRJNA488960 (Zhang Y. et al., 2020), PRJNA524263 (Zhang W. et al., 2020), and PRJNA550237 (Chen et al., 2020).

A total of 730 samples were included with Asian, Western, domestic and wild backgrounds (See Table S1). Raw reads were aligned to the Sscrofa11.1 reference genome (Warr et al., 2020) using the bwa-mem algorithm (Li and Durbin, 2009). Samtools-v1.8 (Li, 2011) was used for sorting, merging, and marking potential PCR duplications. Finally, haplotype-based variant detection was conducted with freeBayes-v1.1 (--min-base-quality 10 --min-mapping-quality 20 --min-alternate-fraction 0.2 --haplotype-length 0 --pooled-continuous--ploidy 2 --min-alternate-count 2) (Garrison and Marth, 2012). After SNP calling, SNP loci were screened and retaining with a quality value greater than 20 (vcffilter -f "QUAL> 20"). Further quality control was conducted with the following criteria: minor allele frequency (MAF) > 0.01, missing rate <0.01, call rate >90%, sequencing depth of sample >4. Individuals and loci satisfying the above criteria were retained for futher analyses, and assigned to their specific background (193 Chinese indigenous pigs, 30 Asian wild boars, 13 Yucatan mini-pigs, 40 Western wild boars, 298 Western commercial pigs and 18 wild suidae; Supplementary Table S1). Finally, phasing and imputation were performed based on this data set with Beagle 5.1 (Browning et al., 2018; Browning et al., 2021) (window = 20 overlap = 4 gp = true ap = true).

2.2 Genetic structure analysis

T-SNE dimensionality reduction was first conducted by *sklearn. manifold.TSNE* (*n_components* = 2, *perplexity* = 24) in scikit-learn-0.23.1 python package on high-quality *Sus. Scrofa* samples. To construct the Neighbor-joining tree (NJ-tree), we calculated the IBS distance matrix by plink-1.9 on phased SNP data with default parameters. Then the NJ-tree was constructed by fastME-v2.15 (-D 1 -m N -b 10000 -T 10 -s -I) (Lefort et al., 2015) with *Sus cebifrons* as the outgroup. The tree was plotted by the iTOl-v5 online tool (Letunic and Bork, 2021). Model-based global ancestry estimation was conducted with Admixture-1.3 (-B10 -c10) (Alexander et al., 2009) with cross-validation to assess the best fitting K-value.

2.3 Local introgression detection

Western wild boars and Yucatan minipigs were combined as the Western haplotypes background for the introgression study. Chinese groups were set as the donor population for every commercial line. Then putative introgression segments were detected for every donor-recipient combination with relative Identity-by-descent (rIBD) method using whole-genome sequencing data (Bosse et al., 2014). Identity-by-descent (IBD) detection was performed with the refinedIBD algorithm (length = 0.1 trim = 0.01 lod = 1) (Browning and Browning, 2013). These parameters were adjusted to detect not only segments that are identical, but segments with similar origins (i.e., Western or Asian) that show higher similarity than expected between Chinese and European ancestries.

The rIBD values were calculated on non-overlapping bins of 10 kb along the genome. For every bin, we calculated rIBD values with the following formula: r$IBD = nIBD_{R,D} - nIBD_{R,B}$. $nIBD_{R,D}$ denotes the normalized IBD (nIBD) value of the recipient-donor pair, $nIBD_{R,B}$ denotes the nIBD value of the recipient-background pair. $nIBD = \frac{Count\_IBD}{Total\_IBD}$, Count_IBD = shared IBD counts between group1 and group2, $Total\_IBD = N_1{}^*N_2$. N_1 and N_2 are the sample size of group1 and group2, respectively. That way, rIBD >0 indicates that commercial breeds (recipient) shares more IBD traces with Chinese indigenous (donor) than Western background, and thus denotes introgression from the donor into the recipient population. In contrast, a negative rIBD value indicates that the number of haplotypes shared by the recipient and the background population is greater than that shared with the donor at that locus. We then performed a Z-transformation of the rIBD values with the mean and standard deviation values of overall IBD from all donor-recipient pairs. We independently selected the presumed introgression bins with a Z-rIBD threshold of μ+2σ for every pair, where μ and σ are the mean and standard deviation of Z-rIBD values. Positive significant Z-rIBD values are thus indicative of the presumed introgression from Chinese breeds into the Western commercial pig.

2.4 Overlapping ratio of Z-rIBD segments

To measure the coincidence of significant positive/negative Z-rIBD fragments between different donor recipients, we calculated the overlapping ratio by the following formula:

$$Overlapping\, ratio_{C,A:\,B} = 1 - \left(\frac{Noneoverlapped\, Counts_{C,A} + Noneoverlapped\, Counts_{C,B}}{Total\, Counts_{C,A} + Total\, Counts_{C,B}} \right)$$

Where A, B, C denote the three populations. When we compare the overlapping level between "C - > A" and "C - > B", population C denotes one donor population while A and B denote two different recipients. To compare the overlapping level between "A- > C" and "B- > C", population C denote on recipient while A and B denote two different donors. $Noneoverlapped\, Counts_{C,A}$: the number of positive/negative fragments/bins shared between C and A but not shared with B. $Total\, Counts_{C,A}$: the total number of positive/negative fragments shared by C and A.

2.5 Selective sweep analysis

We performed a genome scan to detect recent adaptive introgression events using polymorphism data from the recipient populations only, using the VolcanoFinder-v.1.0 tool (Setter et al., 2020). The ancestral genome was constructed with 16-way Enredo-Pecan-Ortheu multiple alignments files, downloaded from the Ensembl v.103 databases (https://www.ensembl.org/). We obtained the allele frequency and the unnormalized site frequency spectrum files required for Volcanofinder, and performed the analysis according to the standard workflow from VolcanoFinder (https://doi.org/10.5061/dryad.7h44j0zr7) (Setter et al., 2020). Finally, variants were polarized by the ancestor alleles status and used as the input for VolcanoFinder-v1.0 (-big 30000, -1 1 1) (Setter et al., 2020).

2.6 Selection of introgression segments for further analysis

To locate important introgressed segments, we merged the consecutive significant ZrIBD bins into one introgression segment. We ranked introgression segments by segment length as the first criterion and average rIBD value as the second. Then, we selected the segments with a length larger than 11 Kb and Log (10) likelihood ratio of selective sweep footprint >11 (the 0.95 quantile). After that, we computed the length of the introgressed segment, selective sweep footprints, average ZrIBD value and average minor allele frequency for every introgression segment and selected the segments that matched all criteria as top candidates for further analysis.

2.7 Haplotype origin tracing

To trace the origin of haplotypes, alleles were first joined into a "FASTA" format sequence from the phased VCF file by an in-house python script. For a genomic region of interest, variants belonging to the same haplotype were joined to a sequence. These haplotypes thus consist of a string of variants derived from the phased VCF. Subsequently, the SNP distance matrix between haplotypes was calculated with SNP-dists v0.7.0 (https://github.com/tseemann/snp-dists). Finally, hierarchical clustering was conducted in R using the gplots package. Paterson's D-statistics (Patterson et al., 2012) were computed by Dtrios (-j100) of the Dsuite v0.4 (Malinsky et al., 2021) tool package.

2.8 Determination of Chinese-derived alleles

We refer to an allele as a "Chinese-derived allele" when it occurs in Duroc and Chinese pigs, but is nearly absent in European wild boars. So the "Chinese-derived allele" should match the following criteria: 1) allele frequency in Duroc pigs≥0.1.2). Allele frequency in any of the Chinese local pig groups≥0.1.3) allele frequency in European wild boars≤0.0125 (i.e., only one European wild boar among 40 boars has that allele and is heterozygous).

2.9 Candidate variants selection and LD calculation

Chinese-derived variants were annotated by snpEff-v5.0 (Cingolani et al., 2012). To pinpoint potential causal variants with a high effect on the phenotype, the variants were then ranked using pCADD's PHRED score. Briefly, the pCADD is the "pig combined annotation dependent depletion", a model to score single nucleotide variants in pig genomes in terms of their putative deleteriousness, or effect on phenotypes, based on a combination of annotations, see (Gross et al., 2020). The pCADD model is a pig-specific variant of the original CADD model that was developed for human aimd aims to discriminate neutral variants from variants with high impact. Then, "Candidate variants" were selected by the following principles: 1) PHRED score >4.3 (the whole genome mean value). 2) missense variant, 3′UTR variant, or 5′UTR variant.

The LD level of the proxy SNPs with other variants from the sequence data was calculated by plink v1.90b6 (--ld-snp new14_97387849 --ld-window 3000 --ld-window-kb 3000 --r2 --ld-window-r2 0) in the Duroc population. The mean r2 values between proxy SNP and other variants in every block were used as the LD level of the proxy SNP and that block.

2.10 SNP selection from the illumina 50 K SNP array dataset

To be able to test phenotypic effects of the Chinese-derived introgressed haplotypes on chromosome 14, we wanted to expand our sample size by incorporating genotype data from commercial Duroc pigs. The genotype data was obtained from routinely screened pigs from Topigs Norsvin pigs that were genotyped by the (Illumina) Geneseek custom 50 K SNP chip with 50,689 SNPs (50 K) (Lincoln, NE, USA). The chromosomal positions are based on the *Sscrofa11.1* reference assembly. In our set of re-sequenced Duroc pigs, SNPs were filtered using the following requirements: Each marker had a MAF greater than 0.01, a call rate greater than 0.85, and an animal call rate >0.7. SNPs with a p-value below 1×10^{-5} for the Hardy-Weinberg equilibrium exact test were also discarded. All pre-processing steps were performed using plink v1.90b3.30 (Chang et al., 2015). SNPs on chromosome 14 were retained for further LD analyses to identify the SNP in highest LD with the candidate variants on the introgressed haplotypes. We tested LD between the candidate SNPs from the sequence in the Asian derived haplotype and SNPs on the 50 K chip by usingPlink-1.9 (--ld-snp new14_97387849 --ld-window 3000 --ld-window-kb 3000 --r2 --ld-window-r2 0).

2.11 Phenotype-genotype association

To estimate the impact of genotypes on production traits, we used the genotype data for the candidate SNP from 11,255 Duroc animals (not all animals have all phenotypes) to test the association of our introgressed allele with the following traits: daily gain from birth to Tstart (25 kg) for 9,921 animals, daily gain from Tstart to the Tend (25–120 kg) in 10,986 animals, backfat at 120 kg (Tend) in 7,192 animals, lean meat percentage and loin depth at the end (120 kg) in 7,688 and 7,192 animals respectively. The corrected phenotypes for all traits of each animal were obtained from the routine genetic evaluation by Topigs Norsvin. Then, for each trait, we conducted a Welch's t-test (significance threshold $p < 0.05$) to test for differences in phenotypes of the different genotypes at our candidate SNP, that were assigned to either European or Asian background.

3 Results

3.1 Data collection

We collected 730 samples (Supplementary Table S1) from NCBI (https://www.ncbi.nlm.nih.gov/) and conducted SNP Calling with freeBayes-v1.1 (Garrison and Marth, 2012). After strict quality control, 19, 656, 271 SNPs and 592 samples were retained for further analyses, including 193 Chinese indigenous pigs, 30 Asian wild boars, 13 Western local pigs (Yucatan mini-pigs), 40 Western wild boars, 298 Western commercial pigs, and 18 samples from suidae in Southeast Asian islands (Supplementary Table S1).

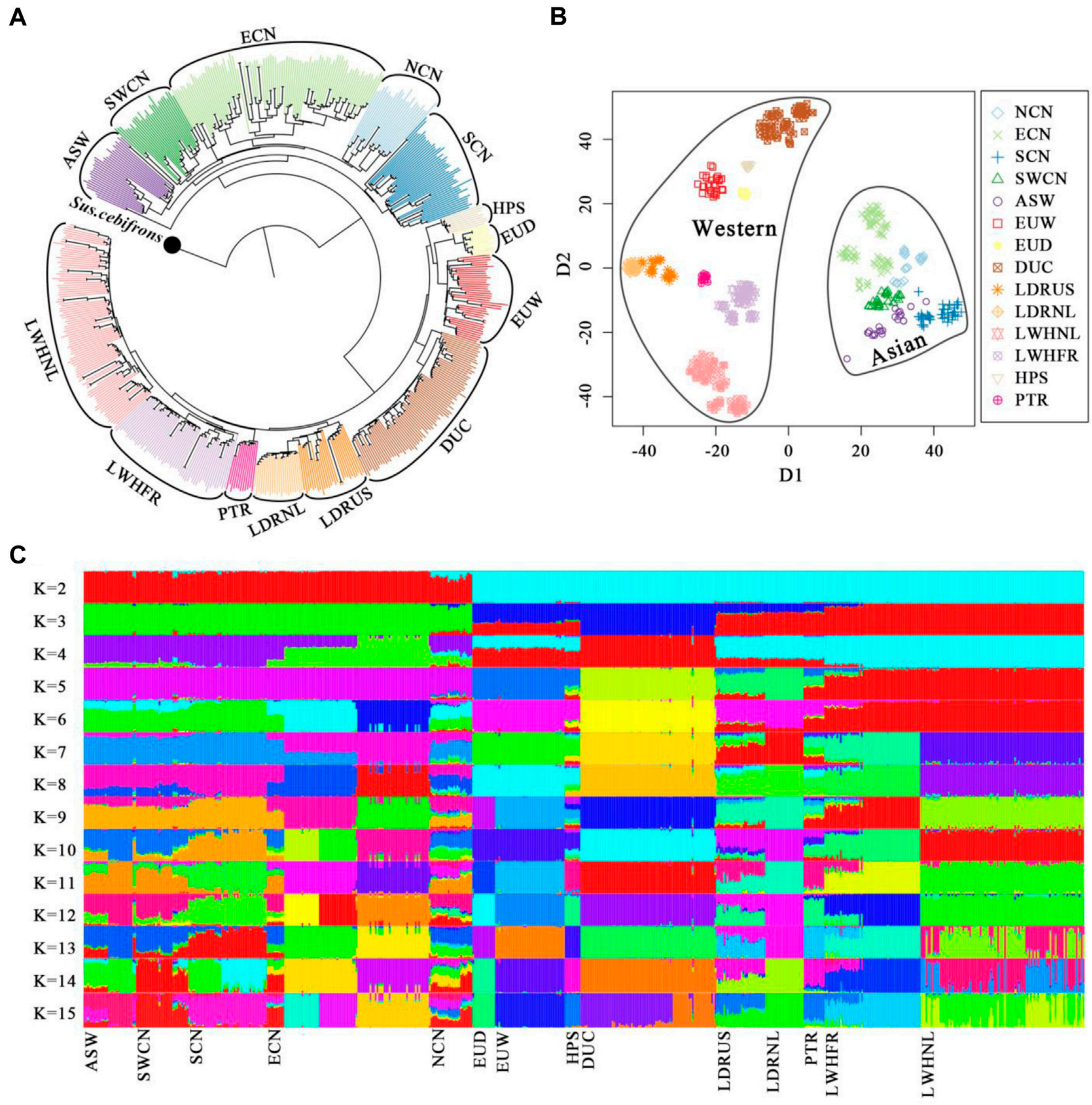


FIGURE 1
Genetic structure of pigs in this study. **(A)**. The Neighbor-joining tree was constructed by fastME-v2.1.5 based on the IBS-distance matrix and set *Sus. cebifrons* form Southeast Asian islands as the outgroup. **(B)**. Dimensionality reduction of whole-genome SNPs with the t-SNE algorithm. **(C)**. Global ancestry inference of Chinese and Western pigs conducted with ADMIXTURE-v1.3.0. NCN, North Chinese pigs; ECN, East Chinese pigs; SCN, South Chinese pigs; SWCN, Southwest Chinese pigs; ASW, Asian Wild boars; EUW, European Wild boars; EUD, European local pigs; DUC, Duroc pigs; LDRUS, American Landrace pigs; LDRNL, Dutch Landrace pigs; LWHNL, Dutch Large White pigs; LWHFR, French Large White pigs; HPS, Hampshire pigs; PTR, Pietrain pigs.

3.2 Genetic structure analysis

We predefined the groups of Chinese pigs according to our previous analysis (Peng et al., 2022) and their geographical origins (Wang et al., 2011) (Supplementary Figure S1). The NJ-tree, t-SNE dimensionality and global ancestry analysis were used to dissect the genetic structure of our samples. The NJ-tree and ancestry inference separate Western and Chinese-derived pigs (Figures 1A–C). Chinese animals clustered into a monophyletic clade, and different Chinese origins clustered into sub-clades except for Chinese Northern pigs (Figure 1A). For Western pigs, every commercial line clustered into a monophyletic clade (Figure 1A) and breeds were clearly distinguished in the t-SNE plot (Figure 1B). Admixture analysis was consistent with this pattern, with increasing values of K above six indicating local ancestry for European pigs, and Asian substructure was best captured with K = 14, when the cross-validation reached a plateau (Figure 1C and Supplementary Figure S2). After we removed a few of the eastern Chinese samples that showed hybrid ancestral components in the admixture result, we assigned the China and Western pig breeds to specific clusters according to geographical sources and genetic relationships.

Finally, Duroc (DUC), Dutch Large White (LWHNL), French Large White (LWHFR), Dutch Landrace (LDRNL), and American Landrace (LDRUS) were recognized as five distinct Western

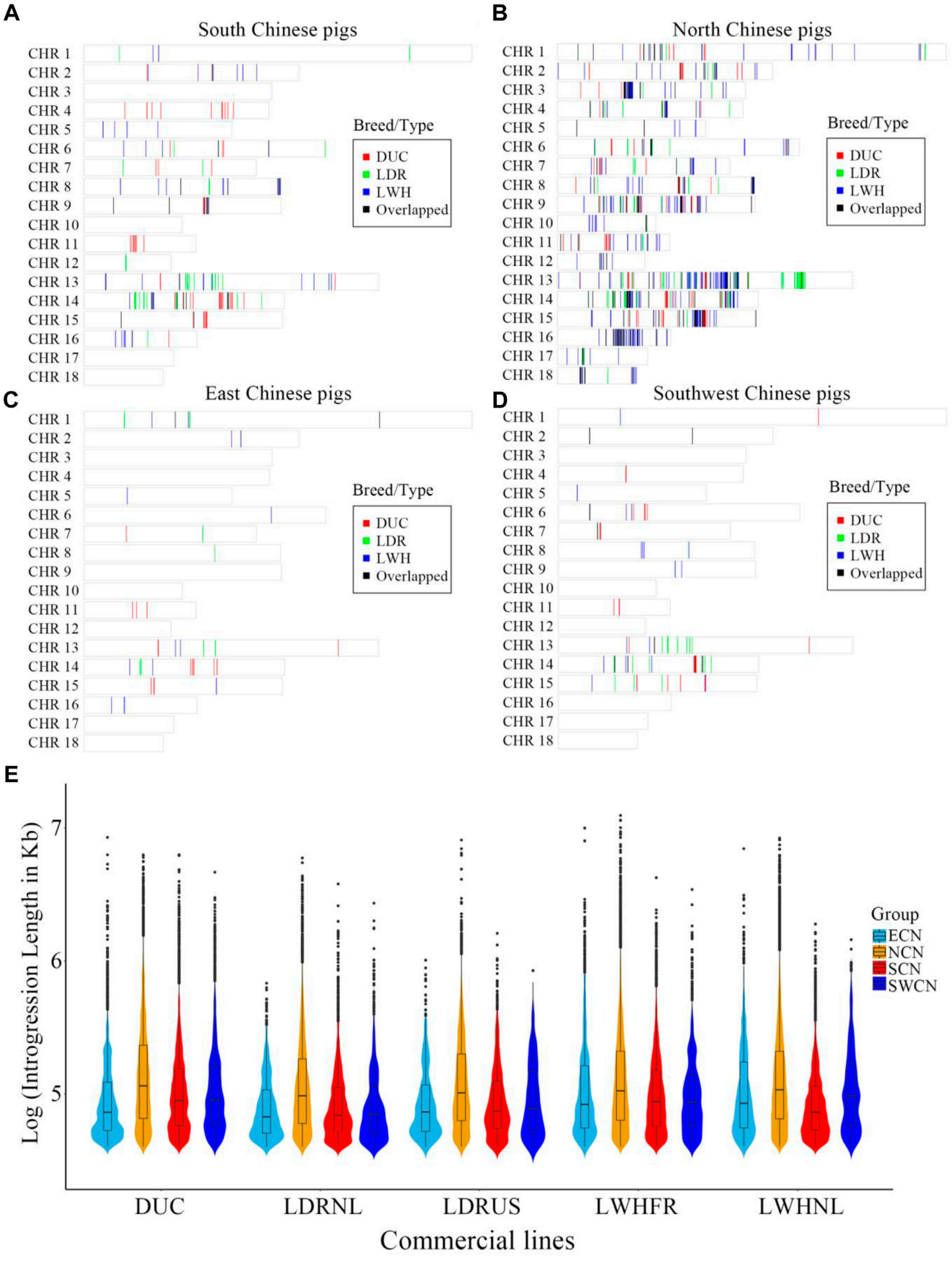
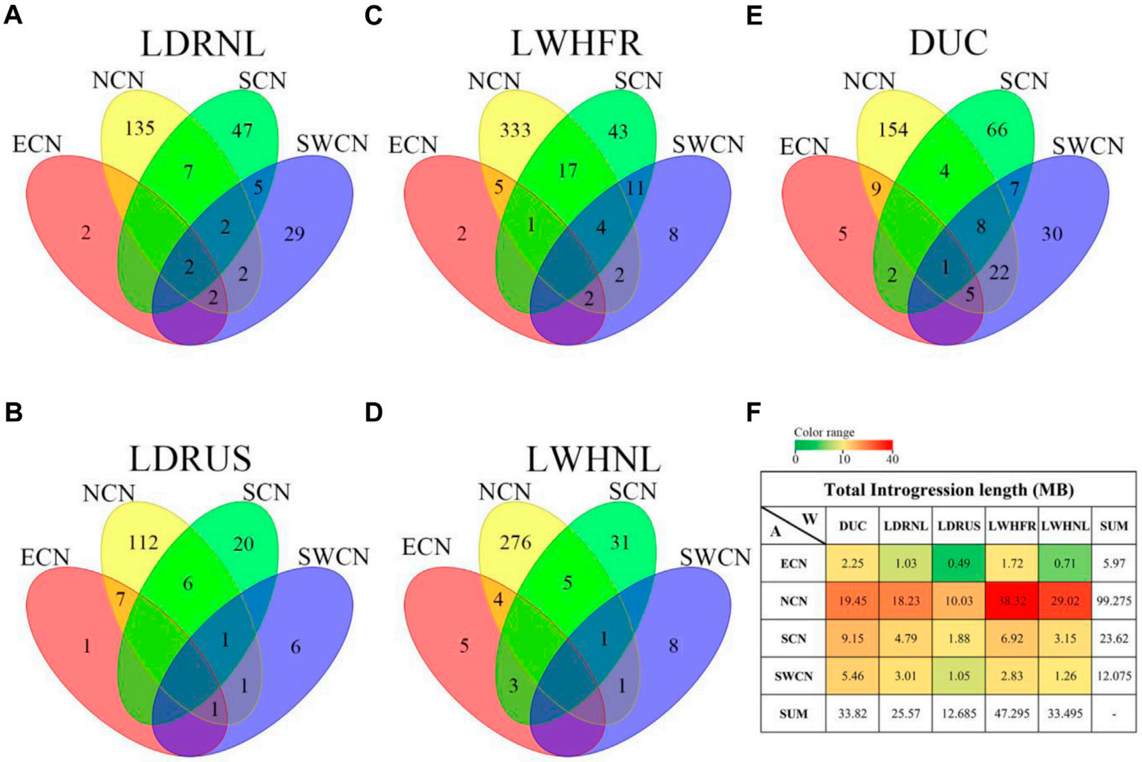


FIGURE 2
The distribution of genomic regions with introgression signature from **(A)** South Chinese pigs, **(B)** North Chinese pigs, **(C)** East Chinese pigs, and **(D)** Southwest Chinese pigs to different Western commercial breeds. DUC: Duroc, LDR: American and Dutch Landrace pigs, LWH: French and Dutch Large White pigs, Overlapped: the overlapped introgressed region between any two pairs. **(E)**. The features of natural logarithms transformed introgressed fragment lengths (in Kb) from China to Western pigs. ECN, East Chinese pigs; NCN, North Chinese pigs; SCN, South Chinese pigs; SWCN, Southwest Chinese pigs. DUC, Duroc; LDRUS, American Landrace pigs; LDRNL, Dutch Landrace pigs; LWHFR, French Large White pigs; LWHNL, Dutch Large White pigs.

commercial lines (WS). European wild boars (EUW) plus local European pigs (EUD) were defined as the Western background population (WB). Moreover, Southern (SCN), Eastern (ECN), Northern (NCN), and Southwestern (SWCN) Chinese pigs were defined as the four Chinese local groups (AB). We combined Chinese wild boars (CNW), Korean wild boars (KRW), and Thai wild boars (THW) into the Chinese background population (AB) (Supplementary Table S1).



Total Introgression length (MB)						
A \ W	DUC	LDRNL	LDRUS	LWHFR	LWHNL	SUM
ECN	2.25	1.03	0.49	1.72	0.71	5.97
NCN	19.45	18.23	10.03	38.32	29.02	99.275
SCN	9.15	4.79	1.88	6.92	3.15	23.62
SWCN	5.46	3.01	1.05	2.83	1.26	12.075
SUM	33.82	25.57	12.685	47.295	33.495	-

FIGURE 3
Venn diagram of gene counts on the putative introgression fragments from Chinese groups to Western commercial breed lines. **(A)**. LDRNL as the recipient. **(B)**. LWHFR as the recipient. **(C)**. DUC as the recipient. **(D)**. LDRUS as the recipient. **(E)**. LWHNL as the recipient. **(F)**. Total length (in Mb) of putative introgression segments (The overlapped introgression has been masked in the "SUM" column and row.).

3.3 Introgression landscape from Chinese to western pigs

We assessed local signatures of introgression in the Western pig genomes using an IBD haplotype sharing method. There are large introgressed fragments and introgression clusters from China to Western pigs (Figures 2A–D). On a genome-wide scale, the proportion and local regions of putative introgression are highly diverse between different donor-recipient pairs. The highest proportion of introgression into Western commercial breeds is NCN, followed by SCN (Figures 2A–D). The amount of introgression varied between the European breeds, with most putative introgression segments from Chinese pigs found in Large White breeds and the French Large White line in particular (Total length of Chinese-derived segments is 33.82 Mb for DUC, 25.57 Mb for LDRNL, 12.68 Mb for LDRUS, 47.29 Mb for LWHFR, 33.49 Mb for LWHNL. Figures 2A–D and Figure 3F). The putative introgression fragments also varied in length and number (Figures 2E, 3E). The longest introgression fragments reach ~1.2 Mb between LWHFR and NCN (Total length: ~38 Mb), but only ~0.34 Mb between LDRNL and ECN (Total length: ~1 Mb). For any Western commercial line, the average introgressed segment length from NCN is longer than from other Chinese populations (Figure 2E), suggesting a relatively recent genetic exchange between NCN and Western pigs.

We studied the number of genes affected by the introgression fragments for different donor-recipient pairs. Results are consistent with the total introgression length (Figure 3F). Most of the genes affected by introgression from local Chinese pigs into Western commercial pigs are specific for every donor-recipient pair (Figures 3A–E). Furthermore, most introgressed genes are from NCN to Large White (LWH), especially LWHFR (428 genes, Figure 3C).

3.4 Various segments from different Chinese groups are introgressed into specific western breeds

We further compared the degree of overlap of positive/negative Z-rIBD segments among donor-recipient combinations to study the global introgression differences. We found that the overall positive Z-rIBD (introgression footprint, see method) patterns are less similar than negative Z-rIBD patterns in different donor-recipient pairs (Figure 4, And Supplementary Figure S3). For donor groups from different sources in China, the degree of overlap significant positive Z-rIBD segments between donor-recipient pairs range from 3.3% to 20.56% (Supplementary Table S2), but for negative Z-rIBD segments, it is 69.73%–93.51% (Supplementary Table S2). The degree of overlap of positive Z-rIBD segments is much lower than the negative Z-rIBD segments (Supplementary Table S2), suggesting specific introgression.

Here we take Duroc as an example. The difference in the positive peaks of Z-rIBD is noticeable (Figure 4). There are peaks at different locations or heights on chromosome nine for ECN, NCN, and SCN (Figures 4A–C), but there is no significant Z-rIBD peak for the SWCN-DUC pair (Figure 4D). On chromosome 11, there are peaks located at 34–39 Mb with different heights or widths (Figures 4A–D). Likewise, on

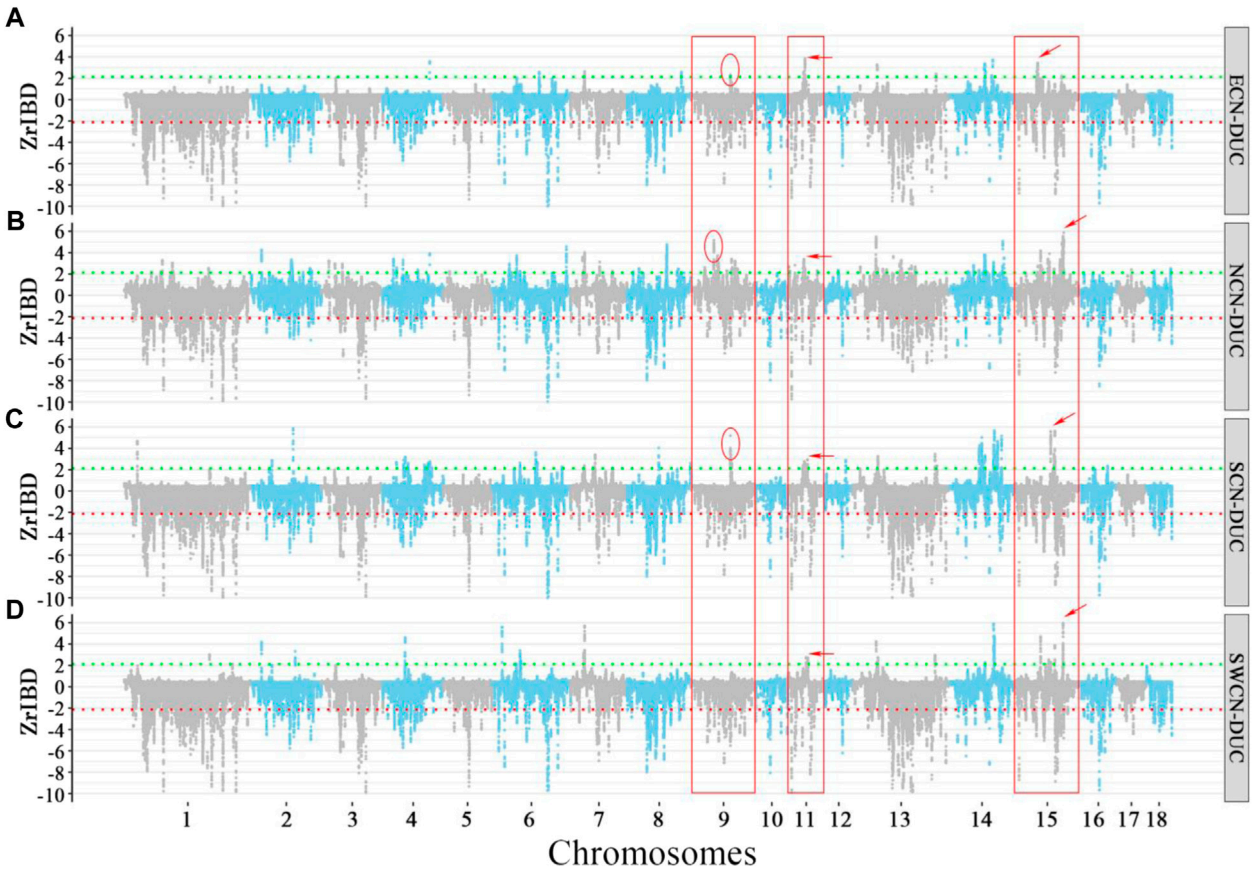


FIGURE 4
Manhattan plot of Z-rIBD values of Duroc *versus* different Chinese indigenous groups with European wild boars and Yucatan minipig as the background population. A positive Z-rIBD value indicates an introgression signal from a Chinese group to Duroc. In contrast, a negative value indicates Duroc shared more IBD fragments with a Western background population than the Chinese group (See method). Green and red dash lines are positive or negative significance levels (*mean* ± *2sd*). **(A)**. The Z-rIBD dot plot with ECN as the donor. **(B)**. The Z-rIBD dot plot with NCN as the donor. **(C)**. The Z-rIBD dot plot with SCN as the donor. **(D)**. The Z-rIBD dot plot with SWCN as the donor.

chromosome 15, the highest peak is located at different positions for the Chinese groups (Figures 4A–D) except for NCN and SWCN (Figures 4B–D). Suggesting that pigs from different regions in China contributed differently to Western commercial pig breeds.

3.5 Hybridization occurred in the early breeding process of commercial pigs

There is a large difference in the introgression patterns between specific Chinese groups and Western commercial lines (Figure 5, And Supplementary Figure S4). The degree of overlap of positive Z-rIBD segments for the specific Chinese local pigs to different European commercial pigs ranges from 0% to 34.83% (Supplementary Table S3). In contrast, for the negative Z-rIBD segments, it ranges from 27.48% to 49.81% (Supplementary Table S3). However, the differences in the introgression patterns within related breeds from a specific Chinese group are smaller. Dutch and French Large White breeds show a similar introgression pattern from North Chinese pigs compared with other commercial pigs (Figures 5D,E, and Supplementary Table S3). The degree of overlap of positive Z-rIBD for these breeds is as high as 34.83% (Supplementary Table S3). In contrast, this is only around 10% compared to the other Western commercial lines (Supplementary Table S3). A broad peak on chromosome 3 (chr3:48–52 Mb) was found in Dutch Large White and French Large White (Figures 5D,E), but not in the other breeds (Figures 5A–C).

In the Landrace breed, the degree of overlap of positive Z-rIBD is as high as 26.55% (Figures 5B,C, and Supplementary Table S3). A significant introgression signal on chromosome 17 (CHR17: 17–18 Mb) is observed in both Dutch and American Landrace (Figures 5B,C) but not seen in the other breeds (Figures 5A,D,E). Besides, the Z-rIBD peaks mentioned above are different in the different pig lines. The observed difference in introgression signal from specific Chinese groups to related Western commercial lines reflects a difference in the extent of introgression. This suggests that gene flow occurred mainly in the early stages of commercial pig breeding rather than after the differentiation of the lines. However, the tendency of artificial selection caused changes in signal strength.

3.6 A Chinese-derived haplotype introgressed into duroc genomes

We observed a cluster of Duroc-specific introgression signatures spanning ~2.65 Mb on chromosome 14 (chr14: 95.68–98.33 Mb)

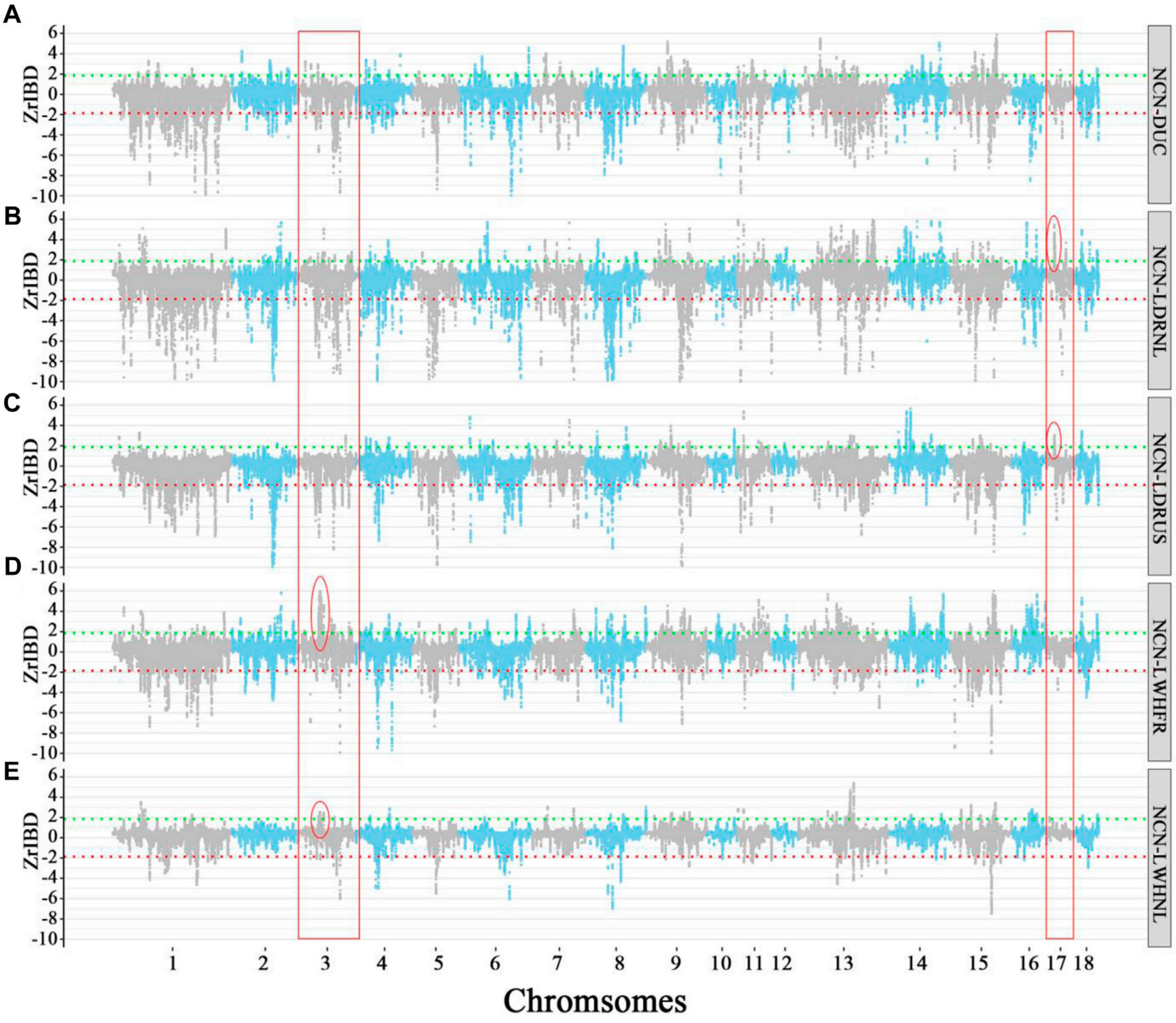


FIGURE 5
The Manhattan plot of Z-rIBD values of Northern China pigs *versus* Western pigs with European wild boars and Yucatan minipig as the background population. A positive Z-rIBD value indicates an introgression signal from a Chinese group to a Western commercial line. In contrast, a negative value indicates a Western commercial line shared more IBD fragments with the Western background population than a Chinese group (See method). Green and red dash lines are positive or negative significance levels (*mean ± 2sd*). **(A)**. The Z-rIBD dot plot with DUC as the recipient. **(B)**. The Z-rIBD dot plot with LDRNL as the recipient. **(C)**. The Z-rIBD dot plot with LDRUS as recipient. **(D)**. The Z-rIBD dot plot with LWHFR as the recipient. **(E)**. The Z-rIBD dot plot with LWHNL as the recipient.

(Figure 6). Such a strong introgression and selection signal is not seen for the other commercial pigs at that region (Supplementary Figures S5, S6). This introgressed region appears to be a set of segments derived from Chinese pigs in the Duroc population. The Z-rIBD value for SCN-DUC is up to 5.65 for segment 3 (the mean Z-rIBD value is 2.48 for segment 1, 3.90 for segment 2, 2.19 for segment 3 and 2.77 for segment 4, Figure 6A). Except for SWCN-DUC in segment 1 (mean Z-rIBD = 5.89, Figure 6D) and segment 3 (mean Z-rIBD = 4.74, Figure 6D), the mean Z-rIBD values is highest in the SCN-DUC pair (Figures 6A–D). We also observed a lower minor allele frequency than expected by chance (0.04 for this region but 0.12 for whole-genome) in Duroc (Figure 6E). These signatures are located within a strong adaptive selection region (chr14: 92–101 Mb) on the Duroc genome (Figure 6F). Moreover, there are five candidate genes in this region: *PCDH15*, *MBL2*, *DKK1*, *PRKG1*, and *CSTF2T* (Figure 6G).

Additionally, PCA plots of Duroc and Chinese pigs from SNPs across the full genome and local SNPs in this region display a strong discondancy (Supplementary Figure S7). The clustering of the Duroc and Chinese pigs in this region hint at introgression, evident from the big difference between the global and local PCA analyses. Combining the above results, we suspect that this haplotype in the Duroc genome was inherited from SCN or SWCN pigs.

To trace the sources of the haplotype, we then calculated a distance matrix between individuals for every segment by SNP-dists v0.7.0 followed by hierarchical clustering in R-4.0 using the gplots package (Warnes, 2020). The results (Figure 7) show that most Duroc pigs clustered together with Chinese pigs (especially with ECN, SCN, and SWCN), in sharp contrast to LWH and Landrace (LDR). The LWH and LDR clustered with Western background populations on segment 1 and segment 4 (Figures 7A,D). In the clustering of segments

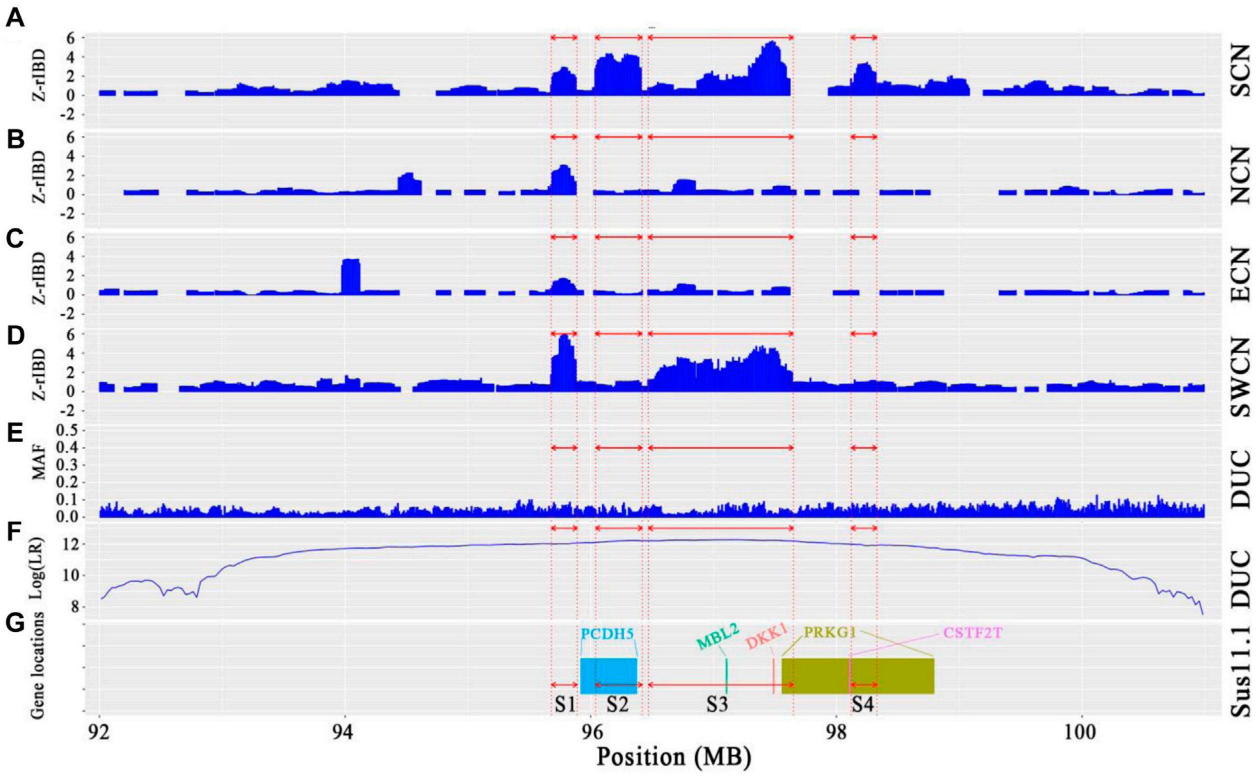


FIGURE 6
Local genome features on chr14:93.98–98.19 Mb of Duroc pigs. **(A–D)**. Z-rIBD values were calculated with Duroc as the recipient, European wild boars and Yucatan minipigs as the background together, and Chinese groups as the donor. **(E)**. Minor allele frequency of the Duroc population. **(F)**. Log-likelihood ratio of selection footprints calculated from VolcanoFinder-v1.0 tool. **(G)**. Candidate gene locations on *Sus. scrofa 11.1* reference genome. S1 denotes segment 1, which locate in chr14:95.68–95.89 Mb; S2 denotes segment 2, which locate in chr14:96.04–96.42 Mb; S3 denotes segment 3, which locate in chr14:96.47–97.65 Mb; S4 denotes segment 4, which locates in chr14:98.12–98.33 Mb.

2 and 4, more SCN pigs are located witnin Duroc clades (Figures 7B,D), suggesting that fragment 2 is more likely derived from SCN pigs. The results of the ABBA-BABA test (D-statistics) highlights that Duroc shares more derived alleles with SCN than other Chinese pigs for segment 2, segment 3, and segment 4 (Table 1). Moreover, a high degree of linkage disequilibrium (LD) in this region (r2 = 0.56, Supplementary Figure S8) was found. According to the dist trees, D-statistics and the degree of LD, we believe that the haplotype (chr14: 95.68–98.33 Mb) in the Duroc genome is derived from SCN pigs.

3.7 Prioritizing causal variants within introgressed haplotypes

We further investigated SNP allele frequencies in the four segments in different pig populations (Table 2, And Supplementary Table S4). There are many alleles with low (≤0.0125) frequencies in Western wild boars but high frequencies in Duroc pigs for each of the segments (123 variants in segment 1,118 variants in segment 2,436 variants in segment 3,383 variants in segment 4, Supplementary Table S4). Furthermore, the derived alleles in Duroc pigs at these loci seem to have undergone strong selection (Figure 6F). We believe these are candidate alleles derived from Chinese pigs due to their moderate allele frequencies in Chinese pigs (Table 2, And Supplementary Table S4). Seven candidate mutations (Supplementary Table S5) were selected from the putative Chinese-derived set of alleles that potentially have a high functional impact (see methods). These variants are likely to have a strong impact on the phenotype as derived from the pCADD model, with the strongest located within the three prime UTR region of *PRKG1*.

3.8 Association of *PRKG1-haplotype* with production traits

We analyzed genotype and phenotype data of 11,255 animals from a commercial Duroc population to assess the potential phenotypic impact of the introgressed haplotypes. We screened the (Illumina) Geneseek custom 50 K SNP array for SNPs in highest LD with the introgressed haplotypes, and a SNP (INRA0045978) was selected as a proxy for the introgressed segment due to its high LD (r2 range from 0.65 to 0.73) with the seven candidate alleles in the Duroc population (Supplementary Table S6, Supplementary Figure S9). Next, we used the genotypes for this selected SNP from 11,255 Duroc animals from the same commercial breed to test the association of INRA0045978 with a set of production traits (See methods).

We found a significant association with backfat (genotype “0/0” *versus* “1/1; *t*-test *p*-value 0.016; Figure 8C) with a and with loin

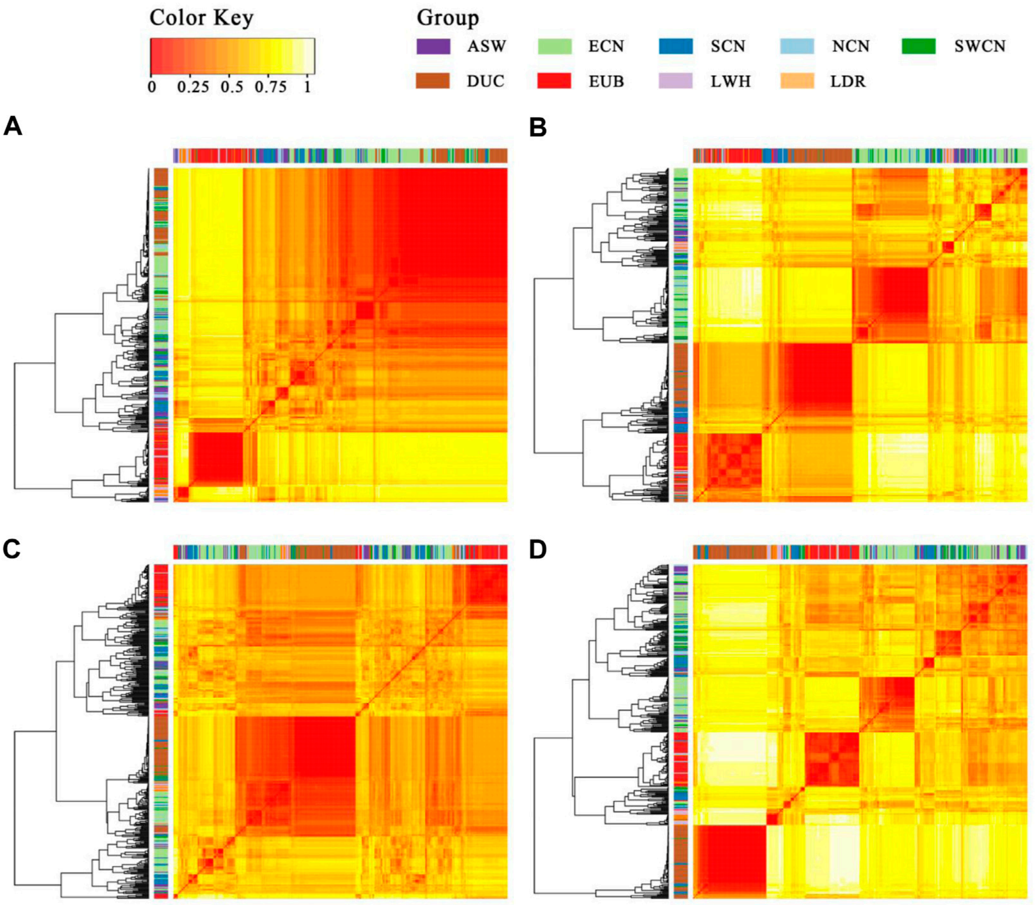


FIGURE 7
Heatmap and hierarchical clustering of the SNP distance matrix. **(A)**. Segment 1 (1343 SNPs were included); **(B)**. Segment 2 (1734 SNPs were included); **(C)**. Segment 3 (7210 SNPs were included); **(D)**. Segment 4 (1761 SNPs were included). SNP distance matrix was calculated with SNP-dists v0.7.0.

depth (genotype “0/1” *versus* “1/1; *t*-test *p*-value 0.028; Figure 8E and Supplementary Table S7). The INRA0045978 SNP has a low Duroc reference allele frequency in Western wild boar (0.0125) but higher in Chinese pigs (0.5517) and Duroc (0.8782). These results suggest that the *PRKG1-haplotype* may decrease backfat (mean difference of 2.3 mm) and increase loin depth (mean difference of 6.1 mm) in Duroc pigs.

4 Discussion

We conducted a comprehensive analysis of the introgression from China to Western commercial pigs. The complexity of the commercial pig breeding process caused unforeseen scenarios. Our findings reveal the distribution and quantity of Chinese pig genetic components in major Western commercial pig breeds.

Interestingly, we found that the overall positive introgression patterns across breeds are less similar than negative patterns. The high degree of overlap for negative Z-rIBD segments was caused by the close genetic relationship among local Chinese pigs. The lower degree of overlap for the positive Z-rIBD segments indicates specific contributions from different Chinese local pigs into Western pigs. This could indicate that some genomic regions in Western pigs do not allow introgression from such distantly related pig populations and that purifying selection is at play. By contrast, breed-specific traits requirements could promote introgression reserved at specific loci, wherein other breeds, these Chinese-derived haplotypes, are undesired. Therefore, we hypothesize that genomic regions lacking Chinese introgression in all Western pigs contain genes that contribute to traits shared across all Western pigs and identify this as an exciting avenue for future research.

Introgressed sequences from different Chinese pig groups were found for a given Western breed. This may have been influenced by the opening of foreign trade ports in China hundreds of years ago and by the traits of pigs in different places (Chen et al., 2020). Western commercial breeds have retained different proportions and different specific loci of introgression. We believe different Chinese pig breeds were introduced for crossbreeding before current Western breeds were established. After establishing Western commercial breeds, these breeds were selected in different directions. We show introgression signals at the same genomic positions but with different introgression intensities for different lines from the same breed. This suggests the influence of directional selection on the gene flow. These results show that the variation in

TABLE 1 D-statistics result of four segments.

Segment	P1	P2	P3	D-statistic	Z-score	*p*-value
S1 (14:95.68–95.89 Mb)	EAS	DUC	EUW	0.322572	3.0973	0.001
	DUC	NCN	EUW	0.189092	2.7052	0.0034
	SCN	DUC	EUW	0.160682	1.1969	0.1157
	SWCN	DUC	EUW	0.158532	2.2859	0.0111
S2 (14:96.04–96.42 Mb)	EUW	DUC	ECN	0.132401	1.1937	0.1163
	DUC	EUW	NCN	0.0728634	0.8777	0.1901
	EUW	DUC	SCN	0.447761	6.6222	2E-11
	EUW	DUC	SWCN	0.187129	1.8819	0.0299
S3 (14:96.47–97.65 Mb)	ECN	DUC	EUW	0.266232	3.7459	9E-05
	DUC	NCN	EUW	0.0031214	0.0414	0.4835
	EUW	DUC	SCN	0.354104	9.3692	0
	SWCN	DUC	EUW	0.180207	2.7965	0.0026
S4 (14:98.12–98.33 Mb)	EUW	ECN	DUC	0.0861798	0.8438	0.1994
	EUW	NCN	DUC	0.0833274	0.9529	0.1703
	EUW	SCN	DUC	0.291517	3.3918	0.0003
	EUW	SWCN	DUC	0.114714	1.3727	0.0849

Note: S1, S2, S3, and S4 indicate segments 1–4 introgressed from Chinese pigs into commercial pigs. D=(ABBA-BABA)/(ABBA + BABA), with closely related *Sus* species from Southeast Asian islands as the outgroup. P1-P3, are the combination of EUW, and Chinese native pig groups (There are four valid combinations according to the formula of D-statistics).

TABLE 2 Average allele frequency of the Chinese-derived alleles within the four segments, in every population.

Segment	DUC	LDRNL	LDRUS	LWHFR	LWHNL	EUD	EUW	ECN	NCN	SCN	Secn	ASW
S1	0.8894	0.0022	0.0172	0.0264	0.0372	0.0003	0.0122	0.7238	0.5020	0.4289	0.7300	0.3062
S2	0.8827	0.0016	0.0058	0.0576	0.1149	0.0007	0.0123	0.1060	0.1123	0.4429	0.1874	0.2032
S3	0.8961	0.3528	0.3401	0.2664	0.2177	0.0004	0.0081	0.3453	0.3090	0.3500	0.5444	0.2241
S4	0.8818	0.0551	0.0602	0.0675	0.0448	0.0006	0.0046	0.1958	0.2003	0.3029	0.2138	0.1974

Note: S1, S2, S3, and S4 indicate segments 1–4 introgressed from Chinese pigs into commercial pigs. In the four segments, there are high allele frequencies in Duroc but low allele frequencies in Landrace and Large White, while these allele frequencies levels in Chinese pigs are high or moderate.

phenotypes of Western commercial breeds is caused by i) their initial variety, ii) different Chinese pigs used for introgression, iii) different directions and strength of selection after introgression. For different commercial lines of the same breed, the variation in phenotypes was most likely mainly caused by variation in the strength of selection. An illustration is the identified novel introgression haplotype from Southern China to Duroc pigs on chromosome 14 harboring the *PRKG1* gene. The *PRKG1* gene straddles the two introgressed segments (segment 3 and segment 4). Considering the high degree of LD in this region, it is very likely that they are derived from a single gene flow event. *PRKG1* has previously been reported to have undergone positive selection in Duroc (Kim et al., 2015) and is related to fatty acid composition. The gene showed copy number variation in Iberian - Landrace crosses (Revilla et al., 2017) and is related to average daily gain in Large White pigs (Wu et al., 2019). Furthermore, we showed that this introgressed *PRKG1*-haplotype significantly affects the thickness of the pig backfat and loin depth (Figures 8C–E), indicating its relevance for commercial breeding.

We also found other genes with essential functions in this region (Table 3). *PCDH15* is related to backfat thickness according to a GWAS result of Landrace and Yorkshire population (Lee and Shin, 2018). Porcine *MBL2* is one of the mannose-binding lectins; it is the central component of innate immunity, facilitating phagocytosis and inducing the lectin activation pathway of the complement system (Phatsara et al., 2007; Bergman et al., 2014). *DKK1* is one of the Wnt signaling inhibitors. Upregulation of *DKK1* expression can be observed in the endometrium in pigs during the pre-implantation period (Zeng et al., 2019). *CSTF2T* plays a potential role in infertility as a mutation in this gene caused male infertility in humans (Gorukmez and Gorukmez, 2020). In conclusion, the introgressed segment contains a set of genes with potential impact on backfat thickness, immunity, daily gain and reproduction.

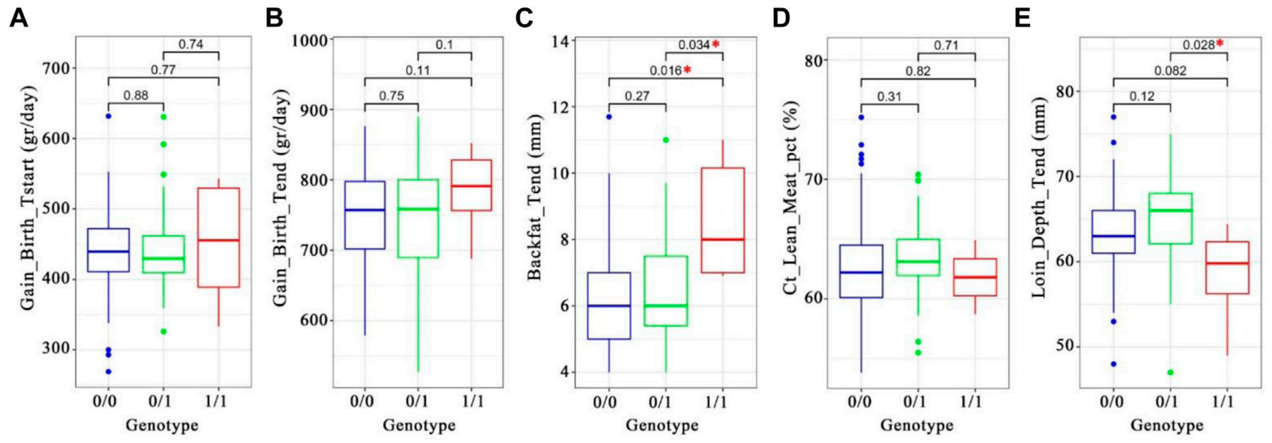


FIGURE 8
Box-plot of phenotype-genotype associations of the introgressed haplotype tagging SNP INRA0045978 (chr14:97387849) in ~11,000 Duroc pigs. A–E, the t. test *p*-values were written on the plots. A star in red denotes significant difference between two genotypes. **(A)** daily gain from birth to starting (Grams per day). **(B)**. Daily gain from start to the end (Grams per day). **(C)**. Backfat at the end (Millimeters). **(D)**. Lean meat percentage (Percentage of lean meat). **(E)**. Loin depth at the end (millimeters).

TABLE 3 Genes overlapping with the four segments within the introgressed region.

Segment lable	Position (BP)	Name	Description
S1 & S2	chr14:95,920,700–96,372,532	PCDH15	Protocadherin related 15
S 3	chr14:97,103,926–97,107,635	MBL2	*Sus scrofa* mannose-binding lectin 2
S 3	chr14:97,487,117–97,490,450	DKK1	Dickkopf WNT Signaling Pathway Inhibitor 1
S 3 & S 4	chr14:97,558,535–98,793,356	PRKG1	Protein Kinase CGMP-Dependent 1
S 4	chr14:98,105,772–98,110,358	CSTF2T	Cleavage Stimulation Factor Subunit 2 Tau Variant

Note: S1, S2, S3, and S4 indicate segments 1–4 introgressed from Chinese pigs into commercial pigs. The name of the gene is GeneCards (https://www.genecards.org/) Symbol. Description information is from GeneCards.

We also observed a large number of introgressed haplotypes in commercial Western pig breeds derived from NCN. However, we did not find any relevant written records of such an introduction of NCN into Europe or America. A general view is that ECN/SCN has been introduced to Europe to improve Western commercial pig breeds (Chen et al., 2017; Zhao et al., 2018; Chen et al., 2020). We, therefore, assume that NCN did not participate in the crossbreeding with Western commercial pigs directly but that the haplotypes introgressed and retained in Western pigs are more conserved in NCN than SCN/ECN. This suggests that current NCN pigs resemble the local breeds introduced centuries ago. This assumption should, however, be confirmed in future studies. Furthermore, it is known that Western commercial pigs contributed to NCN after the 20th century. Ai et al. (2015) found an extreme divergence between the northern and southern Chinese pig haplotypes in the 14-Mb region on the X chromosome. These haplotypes found in NCN were also found in European pigs. Therefore, a reciprocal introgression from European-related boars to NCN and *vice versa* cannot be ruled out. Therefore, care should be taken when assessing the direction of selection and interpretation of the results.

5 Conclusion

A comprehensive analysis of the genetic introgression from Chinese pigs of different regions into different Western commercial lines was studied with 592 re-sequencing pigs. Our analysis revealed different Chinese pig haplotypes' complex introgression patterns and characteristics into Western commercial pig breeds. The results showed that the amount and origin of haplotypes introgressed from different Chinese pig sources to specific Western pigs vary greatly. The impact of Chinese haplotypes from specific sources on different commercial breeds is very different. The introgression likely occurred in the early stages of breed development. Breeding selection tendency experienced by different lines likely led to the observed differences in gene introgression. LWH pigs are most affected by Chinese haplotypes and the haplotypes were better retained in LWHFR. We also found that a ~2.65 Mb Chinese-derived haplotype in Duroc pigs significantly affects the thickness of the pig backfat and the increase of loin depth.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

MB and YZ conceived the idea, YB and MD performed analyses, MG, YZ, and MB provided supervision, YP wrote the manuscript with input from all authors.

Funding

This work was funded by China Scholarship Council (CSC) File No. 201906350013.

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.2022.1070783/full#supplementary-material

References

Ai, H., 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, 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, 1655–1664. doi:10.1101/gr.094052.109

Amaral, A. J., Megens, H. J., Crooijmans, R. P., Heuven, H. C., and Groenen, M. A. (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 179, 569–579. doi:10.1534/genetics.107.084277

Arnold, M. L., Sapir, Y., and Martin, N. H. (2008). Review. Genetic exchange and the origin of adaptations: Prokaryotes to primates. *Philos. Trans. R. Soc. Lond B Biol. Sci.* 363, 2813–2820. doi:10.1098/rstb.2008.0021

Bergman, I. M., Edman, K., van As, P., Huisman, A., and Juul-Madsen, H. R. (2014). A two-nucleotide deletion renders the mannose-binding lectin 2 (MBL2) gene nonfunctional in Danish Landrace and Duroc pigs. *Immunogenetics* 66, 171–184. doi:10.1007/s00251-014-0758-5

Bianco, E., Soto, H. W., Vargas, L., and Pérez-Enciso, M. (2015). The chimerical genome of Isla del Coco feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity. *Mol. Ecol.* 24, 2364–2378. doi:10.1111/mec.13182

Bosse, M., Megens, H. J., Frantz, L. A., 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

Browning, B. L., and Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194, 459–471. doi:10.1534/genetics.113.150029

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

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

Bullock, J. M., Bonte, D., Pufal, G., da Silva Carvalho, C., Chapman, D. S., García, C., et al. (2018). Human-mediated dispersal and the rewiring of spatial networks. *Trends Ecol. Evol.* 33, 958–970. doi:10.1016/j.tree.2018.09.008

Burgarella, C., Barnaud, A., Kane, N. A., Jankowski, F., Scarcelli, N., Billot, C., et al. (2019). Adaptive introgression: An untapped evolutionary mechanism for crop adaptation. *Front. Plant Sci.* 10, 4. doi:10.3389/fpls.2019.00004

Cao, Y. H., Xu, S. S., Shen, M., Chen, Z. H., Gao, L., Lv, F. H., et al. (2021). Historical introgression from wild relatives enhanced climatic adaptation and resistance to pneumonia in sheep. *Mol. Biol. Evol.* 38, 838–855. doi:10.1093/molbev/msaa236

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, H., Huang, M., Yang, B., Wu, Z., Deng, Z., Hou, Y., et al. (2020). Introgression of Eastern Chinese and Southern Chinese haplotypes contributes to the improvement of fertility and immunity in European modern pigs. *Gigascience* 9, giaa014–13. doi:10.1093/gigascience/giaa014

Chen, M., Su, G., Fu, J., Zhang, Q., Wang, A., Sandø Lund, M., et al. (2017). Population admixture in Chinese and European *Sus scrofa*. *Sci. Rep.* 7, 13178. doi:10.1038/s41598-017-13127-3

Choi, J. W., Chung, W. H., Lee, K. T., Cho, E. S., Lee, S. W., Choi, B. H., et al. (2015). Whole-genome resequencing analyses of five pig breeds, including Korean wild and native, and three European origin breeds. *DNA Res.* 22, 259–267. doi:10.1093/dnares/dsv011

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

Denison, M. S., Soshilov, A. A., He, G., DeGroot, D. E., and Zhao, B. (2011). Exactly the same but different: Promiscuity and diversity in the molecular mechanisms of action of the aryl hydrocarbon (dioxin) receptor. *Toxicol. Sci. official J. Soc. Toxicol.* 124, 1–22. doi:10.1093/toxsci/kfr218

Dowling, T. E., Markle, D. F., Tranah, G. J., Carson, E. W., Wagman, D. W., and May, B. P. (2016). Introgressive hybridization and the evolution of lake-adapted catostomid fishes. *PLoS One* 11, e0149884. doi:10.1371/journal.pone.0149884

Dowling, T. E., and Secor, C. L. (1997). The role of hybridization and introgression in the diversification of animals. *Annu. Rev. Ecol. Syst.* 28, 593–619. doi:10.1146/annurev.ecolsys.28.1.593

Enciso-Romero, J., Pardo-Diaz, C., Martin, S. H., Arias, C. F., Linares, M., McMillan, W. O., et al. (2017). Evolution of novel mimicry rings facilitated by adaptive introgression in tropical butterflies. *Mol. Ecol.* 26, 5160–5172. doi:10.1111/mec.14277

Frantz, A. C., Zachos, F. E., Kirschning, J., Cellina, S., Bertouille, S., Mamuris, Z., et al. (2013a). Genetic evidence for introgression between domestic pigs and wild boars (*Sus scrofa*) in Belgium and Luxembourg: A comparative approach with multiple marker systems. *Biol. J. Linn. Soc.* 110, 104–115. doi:10.1111/bij.12111

Frantz, L. A. F., Madsen, O., Megens, H. J., Groenen, M. A., and Lohse, K. (2014). Testing models of speciation from genome sequences: Divergence and asymmetric admixture in island south-east asian Sus species during the plio-pleistocene climatic fluctuations. *Mol. Ecol.* 23, 5566–5574. doi:10.1111/mec.12958

Frantz, L. A. F., Meijaard, E., Gongora, J., Haile, J., Groenen, M. A., and Larson, G. (2016). The evolution of suidae. *Annu. Rev. animal Biosci.* 4, 61–85. doi:10.1146/annurev-animal-021815-111155

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H. J., Bosse, M., Paudel, Y., et al. (2013b). Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus. Genome Biol.* 14, R107. doi:10.1186/gb-2013-14-9-r107

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H. J., Cagan, A., Bosse, M., et al. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148. doi:10.1038/ng.3394

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Available at: https://arxivorg/abs/12073907 (Accessed December 16, 2022).

Giuffra, E., Kijas, J. M., Amarger, V., Carlborg, Ö., Jeon, J. T., and Andersson, L. (2000). The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics* 154, 1785–1791. doi:10.1093/genetics/154.4.1785

Gorukmez, O., and Gorukmez, O. (2020). First infertile case with CSTF2TGene mutation. *Mol. Syndromol.* 11, 228–231. doi:10.1159/000509686

Grahofer, A., Letko, A., Hafliger, I. M., Jagannathan, V., Ducos, A., Richard, O., et al. (2019). Chromosomal imbalance in pigs showing a syndromic form of cleft palate. *BMC Genomics* 20, 349. doi:10.1186/s12864-019-5711-4

Grant, P. R., and Grant, B. R. (2019). Hybridization increases population variation during adaptive radiation. *Proc. Natl. Acad. Sci. U. S. A.* 116, 23216–23224. doi:10.1073/pnas.1913534116

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622

Gross, C., Derks, M., Megens, H. J., Bosse, M., Groenen, M. A. M., Reinders, M., et al. (2020). pCADD: SNV prioritisation in *Sus scrofa*. *Genet. Sel. Evol.* 52, 4. doi:10.1186/s12711-020-0528-9

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, 458–475. doi:10.1111/eva.12887

Hufford, M. B., Lubinksy, P., Pyhajarvi, T., Devengenzo, M. T., Ellstrand, N. C., and Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genet.* 9, e1003477. doi:10.1371/journal.pgen.1003477

Janzen, G. M., Wang, L., and Hufford, M. B. (2019). The extent of adaptive wild introgression in crops. *New Phytol.* 221, 1279–1288. doi:10.1111/nph.15457

Kim, H., Caetano-Anolles, K., Seo, M., Kwon, Y. J., Cho, S., Seo, K., et al. (2015). Prediction of genes related to Positive selection using whole-genome resequencing in three commercial Pig breeds. *Genomics Inf.* 13, 137–145. doi:10.5808/GI.2015.13.4.137

Koch, K., Algar, D., Searle, J. B., Pfenninger, M., and Schwenk, K. (2015). A voyage to terra australis: Human-mediated dispersal of cats. *BMC Evol. Biol.* 15, 262. doi:10.1186/s12862-015-0542-7

Larson, G., and Burger, J. (2013). A population genetics view of animal domestication. *Trends Genet.* 29, 197–205. doi:10.1016/j.tig.2013.01.003

Lee, Y. S., and Shin, D. (2018). Genome-wide association studies associated with backfat thickness in Landrace and Yorkshire Pigs. *Genomics Inf.* 16, 59–64. doi:10.5808/GI.2018.16.3.59

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference Program. *Mol. Biol. Evol.* 32, 2798–2800. doi:10.1093/molbev/msv150

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi:10.1093/nar/gkab301

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

Li, M., Chen, L., Tian, S., Lin, Y., Tang, Q., Zhou, X., et al. (2017). Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* 27, 865–874. doi:10.1101/gr.207456.116

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* 45, 1431–1438. doi:10.1038/ng.2811

Li, Y., Oosting, M., Deelen, P., Ricano-Ponce, I., Smeekens, S., Jaeger, M., et al. (2016). Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat. Med.* 22, 952–960. doi:10.1038/nm.4139

Liu, L., Bosse, M., Megens, H. J., Frantz, L. A. F., Lee, Y. L., Irving-Pease, E. K., et al. (2019). Addendum: Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* 10, 6306. doi:10.1038/s41467-020-20106-2

Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* 21, 584–595. doi:10.1111/1755-0998.13265

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi:10.1016/j.tree.2005.02.010

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, 470–475. doi:10.1038/ng.3775

Megens, H.-J., Crooijmans, R. P. M. A., San Cristobal, M., Hui, X., Li, N., and Groenen, M. A. M. (2007). Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: Differences in microsatellite variation between two areas of domestication. *Genet. Sel. Evol.* 40, 103–128. doi:10.1186/1297-9686-40-1-103

Meng, J. W., He, D. C., Zhu, W., Yang, L. N., Wu, E. J., Xie, J. H., et al. (2018). Human-mediated gene flow contributes to metapopulation genetic structure of the pathogenic fungus *Alternaria alternata* from potato. *Front. Plant Sci.* 9, 198. doi:10.3389/fpls.2018.00198

Onteru, S. K., Fan, B., Du, Z. Q., Garrick, D. J., Stalder, K. J., and Rothschild, M. F. (2012). A whole-genome association study for pig reproductive traits. *Anim. Genet.* 43, 18–26. doi:10.1111/j.1365-2052.2011.02213.x

Ottoni, C., Flink, L. G., Evin, A., Georg, C., De Cupere, B., Van Neer, W., et al. (2013). Pig domestication and human-mediated dispersal in Western Eurasia revealed through ancient DNA and geometric morphometrics. *Mol. Biol. Evol.* 30, 824–832. doi:10.1093/molbev/mss261

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., et al. (2012). Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genet.* 8, e1002752. doi:10.1371/journal.pgen.1002752

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi:10.1534/genetics.112.145037

Peng, Y., Cai, X., Wang, Y., Liu, Z., and Zhao, Y. (2022). Genome-wide analysis suggests multiple domestication events of Chinese local pigs. *Anim. Genet.* 53, 293–306. doi:10.1111/age.13183

Phatsara, C., Jennen, D. G., Ponsuksili, S., Murani, E., Tesfaye, D., Schellander, K., et al. (2007). Molecular genetic analysis of porcine mannose-binding lectin genes, MBL1 and MBL2, and their association with complement activity. *Int. J. Immunogenet* 34, 55–63. doi:10.1111/j.1744-313X.2007.00656.x

Ramirez, O., Burgos-Paz, W., Casas, E., Ballester, M., Bianco, E., Olalde, I., et al. (2015). Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity* 114, 175–184. doi:10.1038/hdy.2014.81

Revilla, M., Puig-Oliveras, A., Castelló, A., Crespo-Piazuelo, D., Paludo, E., Fernández, A. I., et al. (2017). A global analysis of CNVs in swine using whole genome sequence data and association analysis with fatty acid composition and growth traits. *PLoS One* 12, e0177014. doi:10.1371/journal.pone.0177014

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Pääbo, S., et al. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354–357. doi:10.1038/nature12961

Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., and Hermisson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLoS Genet.* 16, e1008867. doi:10.1371/journal.pgen.1008867

Stukenbrock, E. H. (2016). The role of hybridization in the evolution and emergence of new fungal Plant pathogens. *Phytopathology* 106, 104–112. doi:10.1094/PHYTO-08-15-0184-RVW

Suarez-Gonzalez, A., Hefer, C. A., Christe, C., Corea, O., Lexer, C., Cronk, Q. C., et al. (2016). Genomic and functional approaches reveal a case of adaptive introgression from Populus balsamifera (balsam poplar) in P. trichocarpa (black cottonwood). *Mol. Ecol.* 25, 2427–2442. doi:10.1111/mec.13539

Suarez-Gonzalez, A., Hefer, C. A., Lexer, C., Cronk, Q. C. B., and Douglas, C. J. (2018). Scale and direction of adaptive introgression between black cottonwood (Populus trichocarpa) and balsam poplar (P. balsamifera). *Mol. Ecol.* 27, 1667–1680. doi:10.1111/mec.14561

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., et al. (2016). Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nat. Commun.* 7, 12558. doi:10.1038/ncomms12558

Wang, C., Wang, H., Zhang, Y., Tang, Z., Li, K., and Liu, B. (2015). Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. *Mol. Ecol. Resour.* 15, 414–424. doi:10.1111/1755-0998.12311

Wang, J., Liu, C., Chen, J., Bai, Y., Wang, K., Wang, Y., et al. (2020). Genome-wide analysis reveals human-mediated introgression from western Pigs to indigenous Chinese breeds. *Genes. (Basel)* 11, 275. doi:10.3390/genes11030275

Wang, L., Wang, A., Wang, L., Li, K., Yang, G., He, R., et al. (2011). *Animal genetic resources in China: Pigs*. Beijing, China: China Agriculture Press.

Warnes, G. R. (2020). gplots: Various R programming tools for Plotting data. R package version 3.0. Available at: https://CRAN.R-project.org/package=gplots (Accessed December 16, 2022).

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* 9, giaa051. doi:10.1093/gigascience/giaa051

White, S. (2011). From globalized PIG BREEDS TO CAPITALIST PIGS: A study in animal cultures and evolutionary history. *Environ. Hist.* 16, 94–120. doi:10.1093/envhis/emq143

Wichmann, M. C., Alexander, M. J., Soons, M. B., Galsworthy, S., Dunne, L., Gould, R., et al. (2009). Human-mediated dispersal of seeds over long distances. *Proc. Biol. Sci.* 276, 523–532. doi:10.1098/rspb.2008.1131

Wu, P., Wang, K., Yang, Q., Zhou, J., Chen, D., Liu, Y., et al. (2019). Whole-genome re-sequencing association study for direct genetic effects and social genetic effects of six growth traits in Large White pigs. *Sci. Rep.* 9, 9667. doi:10.1038/s41598-019-45919-0

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 F2 resource Population. *Front. Genet.* 9, 401. doi:10.3389/fgene.2018.00401

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

Zeberg, H., and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612. doi:10.1038/s41586-020-2818-3

Zeng, S., Ulbrich, S. E., and Bauersachs, S. (2019). Spatial organization of endometrial gene expression at the onset of embryo attachment in pigs. *BMC Genomics* 20, 895. doi:10.1186/s12864-019-6264-2

Zhang, 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, 1305–1317. doi:10.1534/genetics.118.301143

Zhang, W., Yang, M., Wang, Y., Wu, X., Zhang, X., Ding, Y., et al. (2020a). Genomic analysis reveals selection signatures of the Wannan Black pig during domestication and breeding. *Asian-Australas J. Anim. Sci.* 33, 712–721. doi:10.5713/ajas.19.0289

Zhang, Y., Xue, L., Xu, H., Liang, W., Wu, Q., Zhang, Q., et al. (2020b). Global analysis of alternative splicing difference in peripheral immune organs between tongcheng Pigs and large white Pigs artificially infected with PRRSV *in vivo*. *Biomed. Res. Int.* 2020, 4045204. doi:10.1155/2020/4045204

Zhao, P., Yu, Y., Feng, W., Du, H., Yu, J., Kang, H., et al. (2018). Evidence of evolutionary history and selective sweeps in the genome of Meishan pig reveals its genetic and phenotypic characterization. *Gigascience* 7, giy058. doi:10.1093/gigascience/giy058

Zhao, Q. B., Sun, H., Zhang, Z., Xu, Z., Olasege, B. S., Ma, P. P., et al. (2019). Exploring the structure of haplotype blocks and genetic diversity in Chinese indigenous Pig Populations for conservation purpose. *Evol. Bioinforma. online* 15, 1176934318825082. doi:10.1177/1176934318825082

Zhu, Y., Li, W., Yang, B., Zhang, Z., Ai, H., Ren, J., et al. (2017). Signatures of selection and interspecies introgression in the genome of Chinese domestic Pigs. *Genome Biol. Evol.* 9, 2592–2603. doi:10.1093/gbe/evx186

TYPE Original Research
PUBLISHED 06 February 2023
DOI 10.3389/fgene.2023.1001352

Check for updates

OPEN ACCESS

EDITED BY
Kefei Chen,
Curtin University, Australia

REVIEWED BY
Ottmar Distl,
University of Veterinary Medicine Hannover, Germany
Paolo Zambonelli,
University of Bologna, Italy

*CORRESPONDENCE
Yuchun Pan,
✉ panyuchun1963@aliyun.com
Shaoxiong Lu,
✉ shxlu_ynau@163.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 23 July 2022
ACCEPTED 24 January 2023
PUBLISHED 06 February 2023

CITATION
Wang H, Wang X, Li M, Sun H, Chen Q, Yan D, Dong X, Pan Y and Lu S (2023), Genome-wide association study reveals genetic loci and candidate genes for meat quality traits in a four-way crossbred pig population.
Front. Genet. 14:1001352.
doi: 10.3389/fgene.2023.1001352

Genome-wide association study reveals genetic loci and candidate genes for meat quality traits in a four-way crossbred pig population

Huiyu Wang[1,2†], Xiaoyi Wang[1†], Mingli Li[1], Hao Sun[3], Qiang Chen[1], Dawei Yan[1], Xinxing Dong[1], Yuchun Pan[4]* and Shaoxiong Lu[1]*

[1]Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, China, [2]Faculty of Animal Science, Xichang University, Xichang, Sichuan, China, [3]Faculty of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China, [4]Faculty of Animal Science, Zhejiang University, Hangzhou, Zhejiang, China

Meat quality traits (MQTs) have gained more attention from breeders due to their increasing economic value in the commercial pig industry. In this genome-wide association study (GWAS), 223 four-way intercross pigs were genotyped using the specific-locus amplified fragment sequencing (SLAF-seq) and phenotyped for PH at 45 min *post mortem* (PH45), meat color score (MC), marbling score (MA), water loss rate (WL), drip loss (DL) in the longissimus muscle, and cooking loss (CL) in the psoas major muscle. A total of 227, 921 filtered single nucleotide polymorphisms (SNPs) evenly distributed across the entire genome were detected to perform GWAS. A total of 64 SNPs were identified for six meat quality traits using the mixed linear model (MLM), of which 24 SNPs were located in previously reported QTL regions. The phenotypic variation explained (PVE) by the significant SNPs was from 2.43% to 16.32%. The genomic heritability estimates based on SNP for six meat-quality traits were low to moderate (0.07–0.47) being the lowest for CL and the highest for DL. A total of 30 genes located within 10 kb upstream or downstream of these significant SNPs were found. Furthermore, several candidate genes for MQTs were detected, including pH45 (GRM8), MC (ANKRD6), MA (MACROD2 and ABCG1), WL (TMEM50A), CL (PIP4K2A) and DL (CDYL2, CHL1, ABCA4, ZAG and SLC1A2). This study provided substantial new evidence for several candidate genes to participate in different pork quality traits. The identification of these SNPs and candidate genes provided a basis for molecular marker-assisted breeding and improvement of pork quality traits.

KEYWORDS

GWAS, crossbred pigs, meat quality, SLAF-seq, candidate genes

Introduction

Pork quality is a comprehensive indicator, including meat color, pH, marbling, water-holding capacity, intramuscular fat (IMF), tenderness, etc. (Noidad et al., 2019), which is an important economic factor in the pig industry and has been one of the main objectives in pig breeding programs (Gallardo et al., 2012; Nonneman et al., 2013). In the past, pig breeders have been focused on growth performance but neglected meat quality, resulting in the decline of pork quality. However, due to the fast rise in living standards, consumers favor higher-quality pork. In modern pig breeding, more attention has been paid to improving meat quality traits (MQTs) (Fan et al., 2010). However, it is difficult to genetically improve meat quality using conventional breeding methods because meat quality is measured after slaughter. Previous studies have shown that a lot of pork qualities show low to medium heritability (Lee et al., 2015; Khanal et al., 2019). In the past few years, researchers have been committed to improving meat quality through advanced molecular breeding methods, such as molecular marker assisted selection (MAS) breeding. Recently, many candidate genes affecting MQTs have been reported, including *RYR1*, *PRKAG3*, *PHKG1*, and *IGF2* (Milan et al., 2000; Yu et al., 2008; Škrlep et al., 2010; Ma et al., 2014). To date, a total of 18,011 quantitative trait loci (QTLs) for meat and carcass traits have been accumulated in the pig QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/index, 25 Apr 2022). Among these QTLs, 805, 765, 136, 30, 91, and 1,092 are found to be associated with PH and meat color, marbling score, water holding capacity, cooking loss, and drip loss, respectively. However, most of these QTLs detected by linkage mapping cover large regions of the genome containing hundreds of genes. Furthermore, only a few genes have been successfully applied to improve the MQTs of pigs at present. Consequently, identifying accurate QTL locations and novel candidate genes remains a major challenge.

Genome-wide association study (GWAS) has been increasingly used to identify genomic regions and markers related to quantitative traits more precisely. In recent years, GWAS based on SNP array for MQTs has identified a large number of QTLs and candidate genes (Lee et al., 2012; Luo et al., 2012; Ma et al., 2013; Fabbri et al., 2020; Park et al., 2021). Gao et al. (2021) used the GeneSeek Porcine SNP50K BeadChip for 582 Duroc × (Landrace × Yorkshire) (DLY) commercial pigs to identify genes related to meat-quality traits: thirty-two SNPs and several candidate genes for meat quality were identified. Liu et al. (2015) genotyped 36 Chinese Erhualian pigs and 610 DLY commercial pigs using the Illumina PorcineSNP60K Beadchip, and obtained 35, 985 and 56, 216 high-quality SNPs to perform GWAS for 20 meat quality traits, respectively. Several QTL regions and relevant candidate genes for meat quality traits were detected. However, the SNP array still has disadvantages, for example, that only a small number of known SNPs can be detected, and that marker distribution is biased. Currently, GWAS based on genome-wide sequencing (WGS) is a powerful method to associate genome-wide SNP with meat quality traits (Ji et al., 2018). Wu et al. (2020) used WGS to genotype 30 purebred Qingyu pigs and obtained 18,436,759 filtered SNPs to perform GWAS for meat pH and color. Several SNPs and candidate genes (*CXXC5*, *RYR3*, *BNIP3*, and *MYCT1*) for meat traits were identified. For *Sus Scrofa* with larger genomes, GWAS based on whole-genome sequencing (WGS) is prohibitively expensive. Considering these limitations, specific-locus amplified fragment sequencing (SLAF-seq), a technology based on high-throughput sequencing was developed, which is a cost-effective method for large-scale genotyping (Sun et al., 2013). SLAF-seq technology has the following four significant advantages: the generation of millions of high-density SNP loci covering the whole genome, the ability to detect new SNP loci in unknown mutations, its applicability to any species whether there is a reference genome or not, and the use of representative libraries to reduce sequencing costs. As a consequence, SLAF-seq-based GWAS was successfully applied to detect SNP loci for important quantitative traits in rabbits (Yang et al., 2020), chickens (Wang et al., 2015; Wang et al., 2019; Li et al., 2021), ducks (Xi et al., 2021), and geese (Melak et al., 2021). SLAF-seq has also been successfully used for genotyping of pigs and detected abundant novel mutation sites (Li et al., 2017; Qin et al., 2020). Furthermore, we also identified some genomic regions and several candidate genes for porcine fatness-related and growth-related traits using GWAS based on SLAF-seq technology in our previous studies (Wang et al., 2022a; Wang et al., 2022b).

To produce more genetic variation, A (Duroc×Saba) × [Yorkshire × (Landrace × Saba)] hybrid segregation population was established. As we know, Duroc, Landrace, and Yorkshire pigs are typical lean-type Western commercial breeds widely distributed all over the world and used for commercial production. The shared disadvantage of Western commercial pigs is poor meat quality. However, Chinese native pigs are quite different from Western commercial pigs in meat quality traits. As an invaluable Chinese genetic resource, the fat-type Saba pigs are widely distributed in Yunnan Province, China (Diao et al., 2019), which exhibit high intramuscular fat (IMF) content and superior pork quality. Taking Chinese pig breeds with high meat quality and Western pig breeds with poor meat quality as parents, the hybrid offspring show great differences in meat quality traits and can produce more genetic variation.

Here, we examined 223 four-way crossbred pigs raised under the same environmental conditions for six meat quality traits, including pH at 45 min *post mortem* (pH45), meat color score (MC), marbling score (MA), water loss rate (WL), cooking loss (CL), and drip loss (DL). Subsequently, GWAS based on SLAF-seq was performed, and identified potential loci influencing these traits. The findings served as the foundation for molecular marker-assisted breeding and the improvement for meat quality traits in pigs.

Materials and methods

Ethics statement

All of the animals utilized in this study were handled and used in accordance with the standards established by China's Ministry of Agriculture and Rural Affairs for the care and use of experimental animals. The entire study was given the nod by the Yunnan Agricultural University's (YNAU, Kunming, China) ethics committee.

Animals

A four-way crossbred pig population was established as described previously (Wang et al., 2022a; Wang et al., 2022b). In short, 223 four-way crossbred pigs (115 females and 108 males, DSYLS) investigated were offspring of seven hybrid boars (Duroc × Saba, DS) and 37 hybrid sows (Yorkshire × (Landrace × Saba), YLS) from the pigs and broilers breeding farm in Chuxiong City, Yunnan Province, China (Supplementary Figure S1). These pigs were raised under identical dietary and environmental settings, with automatic water intake and unfettered access to food, which were slaughtered in the same abattoir weighing 105.25 ± 15.75 kg. The ear tissues of 223 pigs were sampled.

Phenotypes

Six meat quality traits were noted after slaughter, including PH45, MC, MA, WL, DL, and CL. The measured muscle samples were from the left side of the carcass. PH45, MC, MA, WL, and DL were measured on the longissimus muscle between the 10th rib and the first lumbar vertebra, and CL was measured on the psoas major muscle. PH45 values were measured at 45 min after slaughter using an automatic pH-STAR. MC (ranging from 1 to 6, 1 presents pale color and 6 presents dark color), and MA (ranging from 1 to 6, 1 presents lack and 6 presents overabundance) were subjectively evaluated according to National Pork Producer Council (NPPC) guidelines. The WL was determined using the filter paper press method as described by Farouk and Wieliczko (2003) with some modifications. Samples were weighed before (Wb) and after (Wa) being subjected to a 35 kg force for 5 min using a pressure instrument (YYW-2, Nanjing Soil Instrument Co., Ltd. Nanjing, China). DL after 24 h storage was measured using a bag method (Honikel, 1987). DL samples were weighed before (Db) and after (Da) being hanged at 4°C for 24 h. Finally, about 20 g cube-like raw meat samples from the psoas major muscle were used to measure CL. The raw was weighed (Cb) and steamed for 30 min. Cooked samples were cooled down to room temperature and re-weighed (Ca). WL, DL, and CL were calculated using the following formula:

$$\mathrm{WL}\ (\%) = \left[(\mathrm{Wb} - \mathrm{Wa}) / \mathrm{Wb}\right] \times 100\%$$

$$\mathrm{DL}\ (\%) = \left[(\mathrm{Db} - \mathrm{Da}) / \mathrm{Db}\right] \times 100\%$$

$$\mathrm{CL}\ (\%) = \left[(\mathrm{Cb} - \mathrm{Ca}) / \mathrm{Cb}\right] \times 100\%$$

Three measurements of PH45, WL, CL, and DL were taken for each sample. Further analyses were conducted using the averages.

The SAS (SAS Institute, Inc., Cary, NC) MEANS procedure was used to create descriptive statistics for meat quality traits under investigation. Using the R package “ggpubr”, the sample distribution was represented as a frequency distribution histogram. The R function “PerformanceAnalytics” carried out the phenotypic correlation analysis. The genetic correlations and genome heritability for six meat quality traits were estimated using the GCTA software (Yang et al., 2011).

SLAF library construction and sequencing

SLAF library construction and sequencing were performed as described previously (Wang et al., 2022a; Wang et al., 2022b). In short, using the phenol-chloroform extraction procedure, genomic DNA was isolated from ear tissue samples. Concentration and purity were then determined using the NanodropTM 2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) and electrophoresis. An electronic digestion prediction experiment used the pig genome (Sscrofa 11.1_102, ftp:/ftp.ensembl.org/pub/release-102/) as the reference genome. *RsaI* and *HaeIII* restriction enzyme combinations were selected to digest eligible genomic DNA according to the selection principle of the enzyme digestion scheme (Sun et al., 2013). The enzyme digested fragment (SLAF tag) was treated by adding single-nucleotide A to the 3′end, and fragments were then ligated to the dual index (Kozich et al., 2013) sequencing adaptors, Adaptor-ligated fragments were then amplified by PCR, purified, pooled, and screened to construct the SLAF library. Meanwhile, to test the validity of the experimental procedure, we also subjected the control genome (*Oryza sativa spp. japonica*; 374. 30 Mb; http://rapdb. dna.affrc.go.jp/) to the identical sequencing procedure. Briefly, SLAF library construction and sequencing for each individual was carried out as previously described (Sun et al., 2013) with a few minor modifications: target DNA fragments of sizes from 314 to 344 base pair (bp) were selected as SLAF tags and used for paired-end sequencing on an Illumina HiSeq 2,500 platform (Illumina, Inc., San Diego, CA, USA) at Beijing Biomarker Technologies Corporation in Beijing, China.

Dual-Index software was used to examine the raw SLAF-seq data in order to acquire the raw sequencing reads for each sample (Kozich et al., 2013). After removing the adapter reads, the guanine-cytosine (GC) content and Q30 ($Q = -10 \times \log_{10} p$) were measured to assess the sequencing accuracy. And then, raw paired-end reads were aligned to the pig reference genome (Sscrofa 11.1_102) using BWA software (Li and Durbin, 2009). Polymorphic SLAFs exhibited sequence polymorphisms between distinct samples.

Identification of SNPs

SNP throughout the entire genome were generated as described previously (Wang et al., 2022a; Wang et al., 2022b). In short, SNP loci were found based on information from polymorphic SLAF tags using predominantly GATK (McKenna et al., 2010). Based on clean reads mapped to the reference genome, local realignments and base recalibration were conducted, and SNPs were detected using GATK software (McKenna et al., 2010). The SAMtools software (Li et al., 2009) was used to detect SNPs in addition to GATK to guarantee the accuracy of the SNPs detected. As the trustworthy set of SNPs to be subjected to the following analysis, we chose the intersection of SNPs found by both GATK and SAMtools. PLINK two software (Purcell et al., 2007) was utilized to filter SNPs according to minor allele frequency (MAF: 0.05) and integrity (int: 0.8). Ultimately, highly consistent population SNPs were detected for GWAS.

Genome-wide association study (GWAS)

A GWAS was carried out to identify the underlying SNP loci or genes linked to meat quality traits in four-way crossbred pigs. Based on the filtered SNPs (227,921 SNPs) and six meat quality phenotypic data, an association analysis was carried out. We used mixed linear model (MLM) of GEMMA software (Zhou and Stephens, 2012) to detect the SNPs associated with meat quality traits. The MLM formula of GEMMA software was as follows:

$$y = W\alpha + x\beta + Z\mu + \varepsilon$$

Where y was an n×1 vector of phenotype in the four-way crossbred pig population; x was an n×1 vector of marker genotypes, W was the matrix of population structure calculated by the ADMIXTURE software (Alexander et al., 2009), and Z was the matrix of the kinship relationship calculated using GCTA software (Yang et al., 2011). α was the vector of fixed effects; β were the marker effects; μ was random effects and ε was the vector of residuals. Finally, for each variant site, an association result could be attained. Bonferroni correction (BC) approach (Zhou and Stephens, 2012) was used for multiple tests in the study. Markers with adjusted $-\log_{10}$ (p) > 5 (control threshold) were regarded to be significant SNPs for meat quality traits (Wang et al., 2022a; Wang et al., 2022b). The threshold p-value for genome-wide 1% and 10% significance were 4.39×10^{-8} (0.01/227,921) and 4.39×10^{-7} (0.1/227,921), respectively, according to the number of filtered SNPs (n = 227,921). A marker was deemed to

TABLE 1 Phenotype and heritability statistics for six meat quality traits in crossbred pigs.

Trait[a]	N[b]	Min[c]	Max[d]	Mean	SD[e]	CV[f]	h^2 (SE)[g]
PH45	223	5.12	7.04	6.16	0.31	4.96	0.34 ± 0.15
MC	223	1.50	4.50	3.26	0.48	14.77	0.20 ± 0.14
MA	223	2.00	5.00	2.91	0.56	19.22	0.23 ± 0.13
WL (%)	221	6.80	25.43	15.95	3.35	21.02	0.19 ± 0.12
CL (%)	223	26.73	46.69	39.30	3.19	8.12	0.07 ± 0.12
DL (%)	213	0.14	11.30	2.27	1.21	53.40	0.47 ± 0.18

[a] CL, cooking loss; DL, Drip loss; MA, marbling score; MC, meat color score; PH45, PH at 45min *post mortem*; WL, water loss rate.
[b] Number of samples.
[c] Minimum.
[d] Maximum.
[e] Standard deviation.
[f] Coefficient of variation.
[g] Heritability (standard error).

be significantly related to the target trait if it passed the threshold score or above the threshold $-\log_{10} p$ given the complexity of the target traits. Finally, the manhattan and Quantile-quantile (Q-Q) plots of GWAS were drawn using the R package "qqman" (Turner, 2014).

Identification, annotation and functional enrichment analysis of candidate genes

Based on the reference (Xie et al., 2017; Xie et al., 2018), the genes in 10 kb upstream or downstream of significant associated SNPs were considered trait-associated potential candidate genes. Using the Ensembl Sscrofa11.1 database (www.ensembl.org), the relevant information of genes within 10 kb upstream or downstream of each significant SNP was obtained. Using Gene Ontology Consortium (http://geneontology.org), GO annotation results of candidate genes were then obtained. GO and KEGG enrichment analyses were performed based on genes located 10 kb upstream and downstream of significant SNPs using the database for annotation, visualization, and integrated discovery (DAVID v6.8, https://david.ncifcrf.gov/). GO terms and KEGG pathways with the threshold p-value ≤ 0.05 were regarded to be significantly enriched.

Haplotype block analysis

Haplotype block analysis was performed with LDBlockShow software (Dong et al., 2021). LD (r^2) value between SNP pairs>0.7 was defined as a LD block.

Results

Phenotype description and genomic heritability for meat quality traits

The statistical data on the six meat quality traits are shown in Table 1. The mean values for PH45, MC, MA, WL, CL, and DL were 6.16%, 3.26%, 2.91%, 15.95%, 39.30%, and 2.27%, respectively. The coefficient of variation (CV) for the six meat quality traits were 4.96, 14.77 19.22, 21.02, 8.12, and 53.40, respectively. The results, therefore, indicated that four-way crossbred pig populations in meat quality traits, especially DL had extraordinary genetic variation. The genomic heritability estimates based on SNP for six meat-quality traits ranged from 0.07 (CL) to 0.47 (DL). The trait distributions are shown in Supplementary Figure S2.

Correlation among meat quality traits

The phenotypic correlation coefficients for PH45, MC, MA, WL, CL, and DL are showed in Table 2. The results showed that WL had the strongest positive correlation with CL ($r = 0.38$, $p < 0.001$). WL had the strongest negatively correlated with PH45 ($r = -0.22$, $p < 0.001$). The six meat quality traits showed low to medium phenotypic correlation ($0.01<|r|<0.38$), indicating that there was no strong phenotypic correlation between the six meat quality traits. The genetic correlations among six meat quality traits are shown in Table 3.

Identification of SLAFs and SNPs

A total of 223 individuals were genotyped and descriptive statistics of the sequence data were presented in our previous study (Wang et al., 2022a; Wang et al., 2022b). In short, a total of 1,190.92 million paired-end reads were obtained. The average value of Q30 and GC content were 90.74% and 44.83%, respectively (Supplementary Table S1), demonstrating that our sequencing results were reliable. Furthermore, a total of 1,552,377 SLAF tags were identified, with 331,608 average SLAFs for accessions. The average sequencing depth of accessions was 11.94 fold (Supplementary Table S2), which guaranteed the accuracy of subsequent analysis. In addition, *Oryza sativa indica* was used as a control during sequencing. The results showed that the enzyme digestion normally efficiency and paired-end comparison efficiency of control data were 90.77% and 95.4%, respectively, indicating that the construction of SLAF libraries was normal.

After genomic mapping and SNP calling, a total of 16,997 polymorphic SLAFs were detected across the accessions.

TABLE 2 Phenotypic correlations for six meat quality traits in crossbred pigs.

Trait[a]	PH45	MC	MA	WL	CL
MC	0.11				
MA	0.14*	0.18**			
WL	−0.22***	0.06	0.09		
CL	−0.19**	−0.05	0.07	0.38***	
DL	−0.11	0.02	0.13*	−0.01	0.14*

[a] CL, cooking loss; DL, Drip loss; MA, marbling score; MC, meat color score; PH45, PH, at 45 min *post mortem*; WL, water loss rate. Negative values represented negative correlation, and positive values represented positive correlation. * significant at p < 0.05, ** significant at p < 0.01, *** significant at p < 0.001.

TABLE 3 Genetic correlations for six meat quality traits in crossbred pigs.

Trait[a]	PH45 T28	MC T29	MA T30	WL T58	CL T32
MC T29	0.49 (0.34)				
MA T30	0.53 (0.26)	**1.00 (0.42)**			
WL T58	−0.48 (0.12)	0.21 (0.23)	0.20 (0.20)		
CL T32	−0.50 (0.17)	−0.09 (0.29)	0.09 (0.27)	**1.00 (0.20)**	
DL T34	**−1.00 (1.66)**	−0.07 (0.49)	0.78 (0.43)	−0.07 (0.33)	0.33 (0.29)

[a] CL, cooking loss; DL, Drip loss; MC, meat color score; MA, marbling score; PH45, PH, at 45 min *post mortem*; WL, water loss rate. Negative values represented negative correlation, and positive values represented positive correlation. The numbers in brackets were standard errors. The extreme values of genetic correlations for meat quality traits were in bold.

Furthermore, 10,784,484 SNPs in all were identified for all individuals. Based on the selection criteria (integrity>0.8; MAF>0.05), a series of quality control filtering of SNPs was carried out to identify 227,921 SNPs used in the subsequent study. Supplementary Figure S3 displayed the density distribution of the filtered and total SNPs across the entire pig genome. SNPs were found in almost all of the non-overlapping 1 Mb regions of the genome. The density distribution of total SNPs and filtered SNPs were calculated on each *Sus Scrofa* autosome and are shown in Table 4. The filtered SNP density across the 18 *Sus Scrofa* chromosomes was one SNP every 10.28 kb on average, demonstrating the data was reliable.

Genome-wide association study and identification of candidate genes

To lessen the impact of population structure and boost the accuracy of GWAS results, the MLM was used to perform GWAS for six meat quality traits. GWAS could be impacted by population stratification, hence quantile-quantile (Q–Q) plots of six meat quality traits were drawn. The Q–Q plot of each trait was shown following the Manhattan plot of the corresponding traits (Figures 1, 2). A total of 64 SNPs were identified as significant ($p < 1.0 \times 10^{-5}$) for the traits studied using MLM (Supplementary Table S3). The genomic inflation factor (λ) at each trait ranged from 1.03 to 1.07.

Among the detected SNPs, three, three, five, three, three, and forty-seven SNPs were significantly associated with PH45, MC, MA, WL, CL and DL, respectively. For pH45, SNPs were distributed in SSC9 (SSC for *Sus scrofa* chromosome), and SSC18. For MC, SNPs were distributed in SSC1, SSC6 and SSC17. For MA, SNPs were distributed in SSC1, SSC3, SSC5, and SSC13. For WL, SNPs were distributed in SSC6, SSC14, and SSC15. For CL, SNPs were distributed in SSC1 and SSC10. For DL, SNPs were distributed in 14 chromosomes except for SSC11, SSC16, SSC17, and SSC18. The phenotypic variation explained (PVE) by the significant SNPs was from 2.43% to 16.32%. Furthermore, 30 genes were thought to be potential candidate genes that were located within 10 kb up- or down-stream of these significant SNPs (Supplementary Table S3).

pH45

GWAS results showed that three SNP loci identified were significantly related to PH45. Among them, the SNP (SSC9: 43364767) was not located in any genes. The significant SNP (rs321002713) on SSC18 explained 11.32% phenotypic variance, which was located within *GRM8*, a protein-coding gene.

MC and MA

A total of three SNPs were significantly associated with MC. The two significant SNPs, rs327814455 on SSC1 and rs690751971 on SSC6, were located within *ANKRD6* and *ENSSSCG00000032113*, respectively. Among, the rs327814455 explained 10.75% phenotypic variance.

For MA, the most significant SNP (rs696643958) on SSC1 was located within *ENSSSCG00000004081*. The significant SNP (rs341748571) on SSC17 explained 10.47% phenotypic variance,

TABLE 4 SNPs distribution on each *Sus Scrofa* chromosome.

Chromosome	Chromosome length (Mb)	Total SNPs	Filtered SNPs	Density of filtered SNPs (kb)
1	274.33	962,754	20,243	13.55
2	151.94	660,827	13,273	11.45
3	132.85	664,042	13,050	10.18
4	130.91	574,489	12,588	10.40
5	104.53	482,531	9709	10.77
6	170.84	824,442	16,685	10.24
7	121.84	591,418	12,173	10.01
8	138.97	558,536	12,296	11.30
9	139.51	634,613	13,944	10.01
10	69.36	418,236	9,101	7.62
11	79.17	386,093	8,224	9.63
12	61.60	390,815	7,156	8.61
13	208.33	729,800	15,140	13.76
14	141.76	662,647	13,709	10.34
15	140.41	546,445	12,131	11.57
16	79.94	363,968	8,639	9.25
17	63.49	364,104	7,783	8.16
18	55.98	302,386	6,760	8.28
Average	125.88	562,119	11,811	**10.28**

SNP, density was presented as the average physical distance between two adjacent SNP loci.
The extreme values of genetic correlations for meat quality traits were in bold.

which was located in the *MACROD2* gene. The SNP rs325690789 on SSC5 was located within *FGD4*, and rs342013877 on SSC13 was located 5 kb upstream of the *ABCG1* gene.

WL and CL

A total of three SNPs (rs1113389876, SSC14:36676133 and SSC15: 19876509) were significantly associated with WL. The most significant SNP (rs1113389876) on SSC6 was located within the *TMEM50A* gene and 7.9 kb upstream of the *RHCE* gene. The significant SNP (rs693644154) on SSC15 was located 2.7 kb upstream of the *RRM2* gene.

For CL, the most significant SNP (SSC1: 271857436) was located within the *MED27* gene. Furthermore, two nearby significant SNPs (rs331296609 and rs344980768) on SSC10 were located in the *PIP4K2A* gene. These two SNPs were mapped to one haplotype block spanning 16 bp affecting CL on SSC10 (Figure 3A), which each explained 2.43% of the CL phenotypic variance.

DL

A total of 47 significant SNPs were identified for DL. Among these SNPs, two SNPs (rs321165533 on SSC6 and rs323693055 on SSC13) exceeded the 1% genome-wide significance level. The SNP rs321165533 explained 14.55% phenotypic variance, which was located within the *CDYL2* gene. Eight SNPs (AEMK02000361.1: 578806, rs337747094 on SSC3, rs320599347 on SSC4, rs333401534 and rs327130062 on SSC8, SSC10:59940478, rs326956966 on SSC12, and rs703586532 on SSC13) exceeded the 10% genome-wide significance level. Among these significant SNPs, two nearby SNPs on SSC13 (rs323693055 and rs703586532) were located in the *CHL1* gene. The rs703586532 and rs323693055 explained 11.58% and 13.46% phenotypic variance, respectively. The SNP rs320599347 was located within the *ABCA4* gene.

On SSC6, two adjacent significant SNPs (rs326829022 and rs1112488011) were located within *FA2H*. On SSC3, four nearby significant SNPs were located in a region from 7863132 to 7863391 bp (0.26 kb interval), which were located within the *ZAG* gene. Two adjacent significant SNPs (SSC2:25635102 and SSC2:25635114) were located within *SLC1A2*. The two significant SNPs were mapped to one haplotype block spanning 12 bp affecting DL on SSC2 (Figure 3B), which each explained more than 9% of the DL phenotypic variance. Additionally, the rs327708082 on SSC2 explained the highest DL phenotypic variance (16.32%), which was located within the *SIL1* gene.

Furthermore, several significant SNPs explained more than 13.35% phenotypic variance, which were not located any known genes, including rs345860122 on SSC4 (13.48% PVE for DL), rs324617714 and rs325613231 on SSC7 (13.35% PVE for DL).

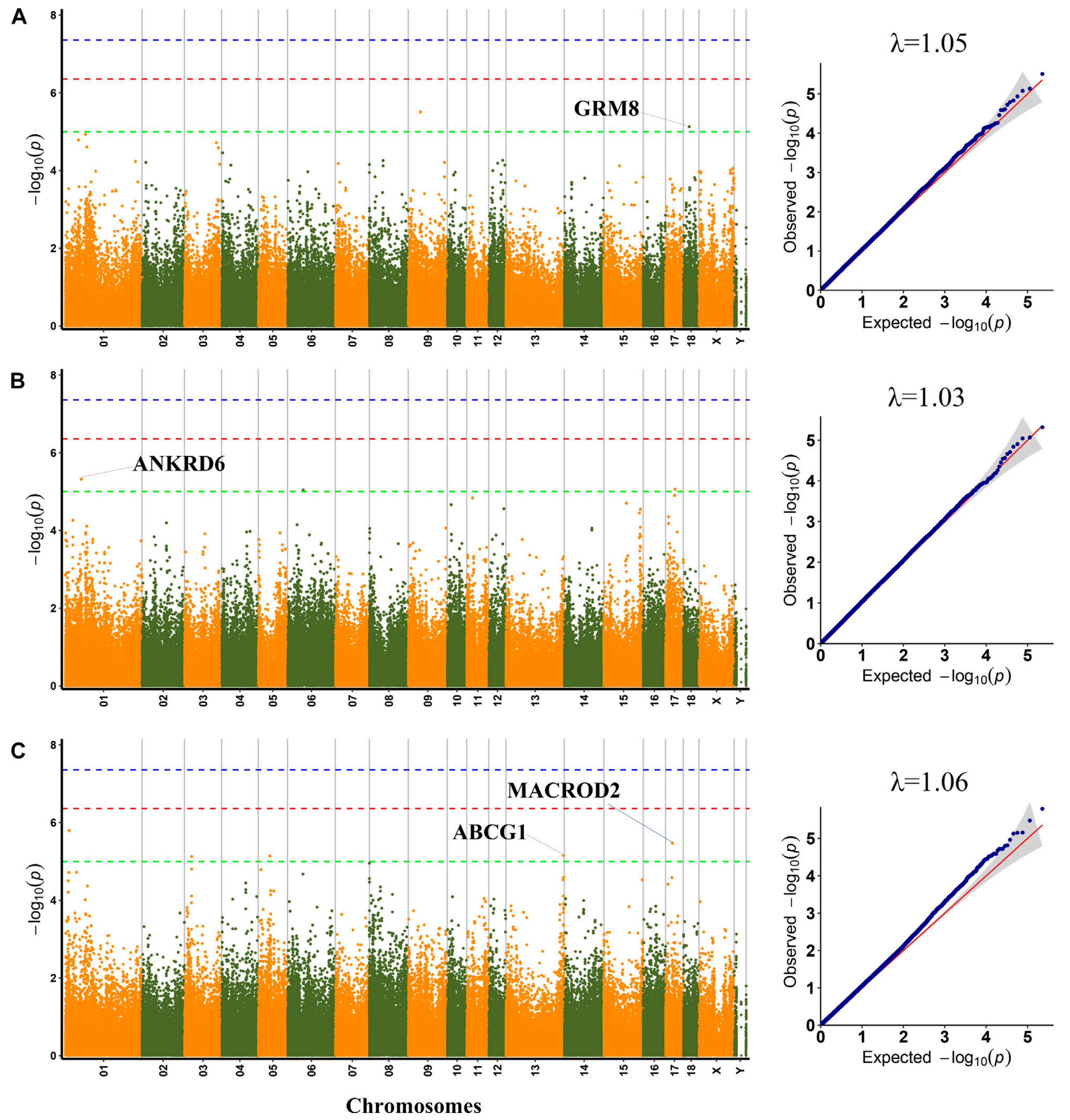


FIGURE 1
Manhattan plots and QQ plots for pH45, MC and MA using MLM. **(A)** pH45 **(B)** MC **(C)** MA. Negative $\log_{10}$ p-value of the filtered high-quality SNPs were plotted against their genomic positions; The dashed lines of green, orange and blue correspond to the Bonferroni-corrected thresholds of $p = 1.00 \times 10^{-5}$ ($-\log_{10} p = 5$), $p = 4.39 \times 10^{-7}$ ($-\log_{10} p = 6.36$), and $p = 4.39 \times 10^{-8}$ ($-\log_{10} p = 7.36$), respectively; λ, Genomic inflation factor.

Comparison with previously reported QTL in pigs

The Pig Quantitative Trait Locus (QTL) Database (Pig QTLdb, https://www.animalgenome.org/cgi-bin/QTLdb/SS/index, 25 Apr 2022) was searched based on SNP and QTL locations to evaluate if QTLs linked to meat quality traits in this study repeat any previously reported QTLs. A total of 64 SNPs significantly associated with meat quality traits in four-way crossbred pigs were identified using the MLM, of which 24 SNPs were located in previously reported QTL regions that were associated with the meat quality traits of pigs (Supplementary Table S4). Three QTLs, including 9.35-Mb (262.87–272.22Mb) on SSC1, 5.29-Mb (7.60–12.89Mb) on SSC6, and 0.09-Mb region (63.38–63.47Mb) on SSC9 for DL were identified.

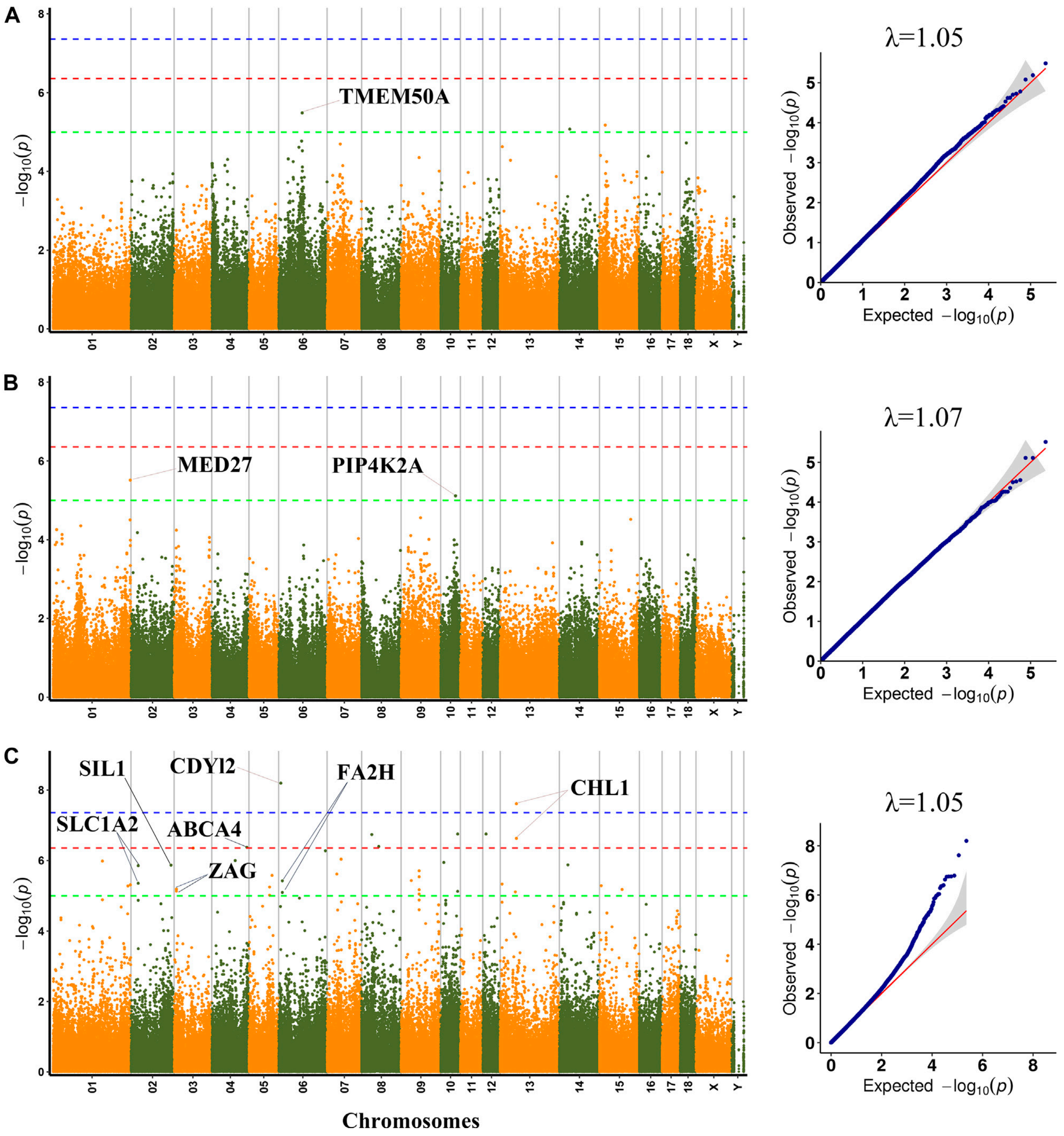


FIGURE 2
Manhattan plots and QQ plots for WL, CL and DL using MLM. **(A)** WL **(B)** CL **(C)** DL. Negative $\log_{10}$ *p*-value of the filtered high-quality SNPs were plotted against their genomic positions; The dashed lines of green, orange and blue correspond to the Bonferroni-corrected thresholds of $p = 1.00 \times 10^{-5}$ ($-\log_{10} p = 5$), $p = 4.39 \times 10^{-7}$ ($-\log_{10} p = 6.36$) and $p = 4.39 \times 10^{-8}$ ($-\log_{10} p = 7.36$), respectively; λ, Genomic inflation factor.

GO annotation and functional enrichment analysis of candidate genes

The result of GO annotation showed that *ENSSSCG00000004081* participated in muscle cell differentiation and actin filament binding. *ABCG1* was involved in negative regulation of lipid storage, response to lipid and phospholipid homeostasis. The GO annotation results of other genes are shown in Supplementary Table S3.

Furthermore, two GO terms (actin binding and photoreceptor outer segment) and one KEGG pathway (ABC transporters) were significantly enriched (*p*-value ≤ 0.05) (Table 5).

Discussion

In this study, we used SLAF-seq technology to obtain 227,921 highly consistent SNPs. Previous studies have proven the

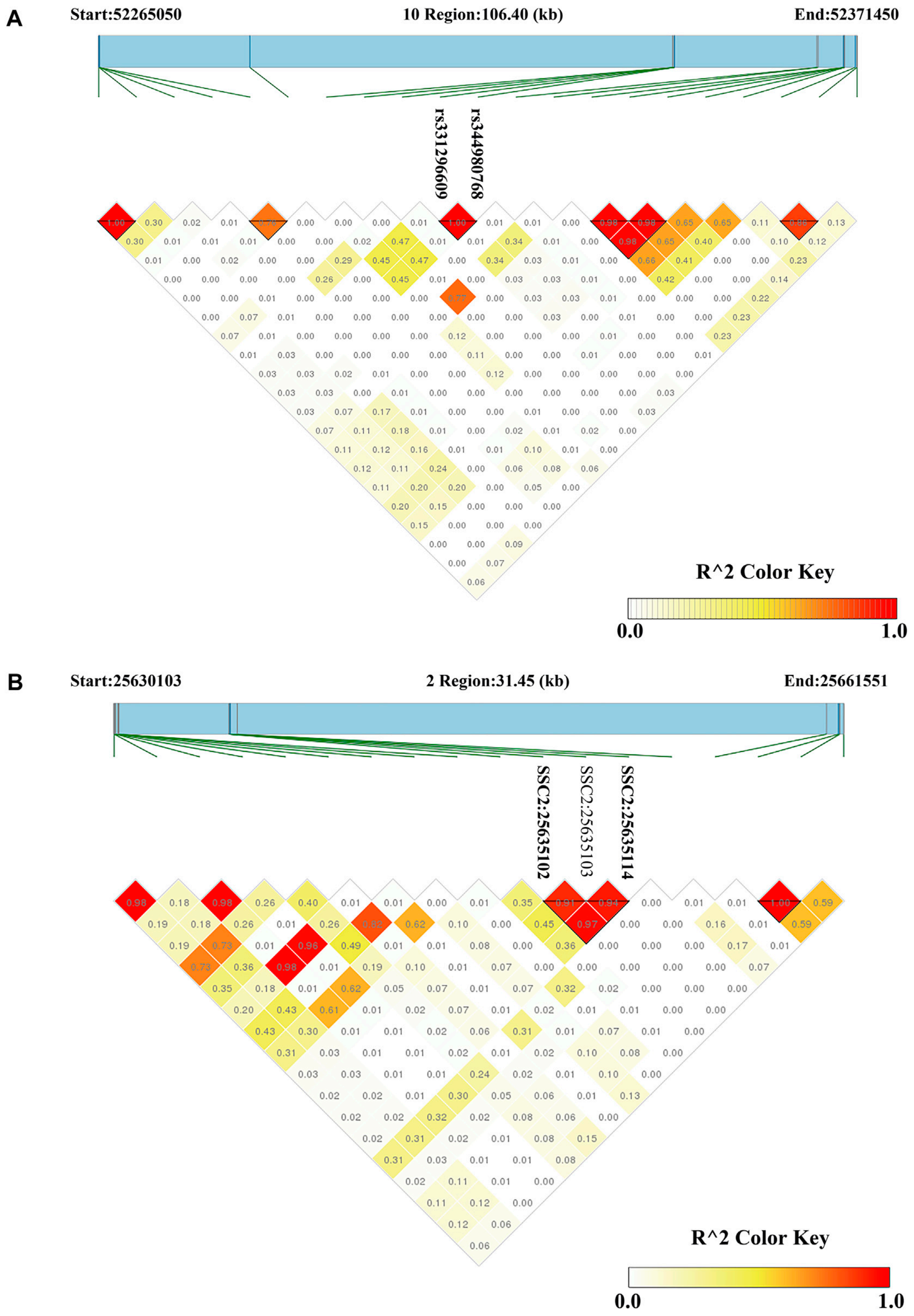


FIGURE 3
Haploview plots of linkage disequilibrium (r2) between SNPs on pig chromosome. **(A)** A region on SSC10 (106.40 kb) contained a haploview block with two significant SNPs related to CL **(B)** A region on SSC2 (31.45 kb) contained a haploview block with two significant SNPs related to DL. Values in the diamond are r^2 values between SNPs. The darker the color, the stronger the correlation between two SNPs.

advantage of the SLAF-seq method in the GWAS, genetic diversities analysis, and construction of genetic map for animals and plants (Qi et al., 2014; Li et al., 2017; Qin et al., 2020; Yang et al., 2020; Li et al., 2021; Mandozai et al., 2021). SLAF-seq technology can obtain more genomic variation sites than SNP chips, detect novel mutation sites and provide high SNP coverage at a low cost. However, SLAF-seq

TABLE 5 Significant GO terms and KEGG pathways associated with meat quality traits in crossbred pigs ($p \leq 0.05$).

Terms[a]	ID	Count	p-value	Genes
KEGG: ABC transporters	ssc02010	2	0.05	*ABCG1, ABCA4*
MF: actin binding	GO:0003779	1	0.0005	*ENSSSCG00000004081*
CC: photoreceptor outer segment	GO:0001750	2	0.05	*ABCA4, PIP4K2A*

[a]CC, cell component; KEGG, kyoto encyclopedia of genes and genomes; MF, molecular function.

technology obtains fewer numbers of molecular markers compared with WGS technology. In further study, we used genome re-sequencing technology to attain genome-wide genetic variation, and provided opportunities for understanding more comprehensively and accurately the genetic architecture of pig meat quality traits. Furthermore, these SNPs were used to calculate genetic parameters for six meat quality traits. The genomic heritability estimates based on SNP for six meat-quality traits were low or moderate (0.07–0.47) (Table 1), which was similar to the results of previous studies (Lo et al., 1992; Miar et al., 2014; Gao et al., 2021). The results showed that these meat-quality traits could be genetically improved. There were a high negative genetic correlation (−1.00 ± 1.66) between DL with pH45 and a positive correlation (0.33 ± 0.29) between DL with CL (Table 3), which was similar to the results of the previous studies (Gjerlaug-Enger et al., 2010; Miar et al., 2014). In addition, there were a high positive genetic correlation between MC and MA (1.00 ± 0.42), which was similar to the results of a previous study (Gjerlaug-Enger et al., 2010). There were a high positive genetic correlation (1.00 ± 0.20) between WL and CL, was similar to results of a previous study (Fernández-Barroso et al., 2020). Besides, they had the highest phenotypic correlation ($r = 0.38$; $p < 0.001$).

The standard deviation (SD) of phenotypic values for PH45, MC, and MA were 0.31, 0.48, and 0.56, respectively, which were similar to the results of Gao et al. (2021). Gao et al. found that SD for PH45, MC, and MA were 0.37, 0.55 and 0.61, respectively, in a three-way crossbred commercial pig population. The SD for WL and CL was 3.35 and 3.19, respectively, which were less than the previous studies, including 5.3 for WL in a Korean Native × Landrace F2 cross population (Lee et al., 2012), and 4.17 for CL in a specially designed heterogeneous F6 pig population (Ji et al., 2018). The SD for DL was 1.21, which was more than the previous studies, including 0.33 for DL in a White Duroc × Erhualian F2 population (Ma et al., 2013), and 2.0 and 2.26 for DL in a Korean Native × Landrace F2 cross population (Choi et al., 2011; Lee et al., 2012). Interestingly, the phenotypic variation explained (PVE) of all significant SNPs detected in this study is greater than 2.43%. Among them, the PVE of 26 SNPs was even greater than 10%. The higher PVE of these molecular markers implies that these markers could be used in molecular marker-assisted selection and genome selection in pigs to increase pork quality. Besides, the genomic inflation factor (λ) at each trait ranged from 1.03 to 1.07 (Table 1), and none of the Q–Q plots showed any sign of inflation, indicating that the MLM effectively controls the false positive result, and effectively lessen the impact of group stratification on GWAS results, which ensure the reliability of GWAS results.

MAF by the significant SNPs was from 0.05 to 0.49 (Supplementary Table S3). Some significant markers had a low MAF (such as rs327708082, MAF = 0.06). The allele with the lowest frequency had large or very small effects on meat quality traits, depending if these allele showed a positive or negative effect. If the allele with the highest frequency has a positive effect, the selection will not work. In view of these problems, further research was needed to carry out.

Comparison of pig populations used in this study with those used in other studies

In previous studies, most of the SNPs and candidate genes for important economic traits of pigs identified based on GWAS mainly used F2 generation populations, which were generated by crossbreeding local pig breeds from different countries with Western lean pig breeds (Liu et al., 2014; Zhang et al., 2014; Cho et al., 2015; Guo et al., 2020; Liu et al., 2020) and purebred pigs (Xiong et al., 2015; Ding et al., 2019; Fabbri et al., 2020; Fu et al., 2020). The F2 generation population is characterized by segregation of traits, large phenotypic variation and more genetic diversity, which is suitable for GWAS. Some studies use white Duroc × Erhualian F2 hybrid pig population to conduct GWAS on growth, fat, meat quality, muscle fiber, body size and body weight traits (Ma et al., 2013; Qiao et al., 2015; Guo et al., 2017; Ji et al., 2017; Guo et al., 2020), and obtained a large number of mutation sites and candidate genes related to the research traits. Besides, two studies used Large White × Minzhu F2 generation population to perform GWAS on meat quality and external traits (Luo et al., 2012; Wang et al., 2014), and identified some SNP loci and candidate genes related to meat quality and external traits. A F2 intercross between Landrace and Korean native pigs was used to perform GWAS for meat quality traits (Lee et al., 2012; Cho et al., 2015). In the present study, three typical Western lean-type pig breeds, Landrace, Yorkshire and Duroc, were hybridized with Saba pig, a Chinese local fat-type pig breed, to establish (Duroc×Saba) × [Yorkshire × (Landrace × Saba)] hybrid segregation population, which was used to perform GWAS for six meat quality traits. The four-way hybrid pig population has greater phenotypic variation and more genetic diversity, is a more ideal population for GWAS than the two-way hybrid population and the purebred pig population.

QTLs identified for meat quality traits

In the present study, 64 SNPs in all were detected using MLM as significant for the meat quality traits studied, of which 24 SNPs were located in previously reported QTL regions for meat quality traits in pigs. Three genomic regions, including 9.35-Mb (262.87–272.22Mb, 3SNPs) on SSC1, 5.29-Mb (7.60–12.89Mb, 3SNPs) on SSC6, and 0.09-Mb region (63.38–63.47Mb, 4SNPs) on SSC9 for DL were located in previously reported QTL regions on SSC1, 6 and 9 for DL (Malek et al., 2001; Thomsen et al., 2004; Liu et al., 2008). Besides, some significant

SNPs overlapped with previously reported QTL regions on SSC9 and SSC18 for pH (Harmegnies et al., 2006; Edwards et al., 2008), on SSC6 for MC (Edwards et al., 2008; Li et al., 2010), on SSC1, 5 and 17 for MA (Rohrer et al., 2005; Cho et al., 2015), on SSC6 for water holding capacity (Su et al., 2004). Among SNPs, 40 SNPs had not been included in any previously reported QTLs for meat quality traits (Supplementary Table S4). Two novel QTLs significantly associated with DL, including a 0.08-Mb region (72.91–72.99Mb) on SSC5, a 3.6-Mb region (53.28–56.88Mb) on SSC13 (Supplementary Table S4). In different studies, depending on the specific genetic backgrounds and sample size, different QTLs may be mapped. Moreover, measuring the phenotype of pork quality is a challenge, and different studies may not be measuring exactly the same location of the muscle for meat quality traits. This could contribute to the differences between studies.

Additionally, a 0.36-Mb region (271.86–272.22Mb) on SSC1 was identified as being significantly associated with CL and DL, containing SSC1:271857436 for CL, and rs710333950 and rs326037487 for DL (Supplementary Table S4). A 9.08-Mb region (24.41–34.49 Mb) on SSC17 was identified as being significantly associated with MC and MA, containing rs341748571 for MA, and rs1112200844 for MC (Supplementary Table S4). The findings suggested that certain chromosomal regions might have varying effects on different meat quality traits. Low phenotypic correlation coefficients (r = 0.14; p < 0.05) and low genetic correlation (0.33 ± 0.29) (Table 3) between CL and DL were founded. Furthermore, low phenotypic correlation coefficients (r = 0.18; p < 0.01) and High genetic correlation (1 ± 0.42) (Table 3) between MC and MA were founded. As a result, the correlation between the two traits might help to partially account for the pleiotropic effects in the region.

Candidate genes for six meat quality traits

Candidate genes for pH45

Pork pH can affect the quality of meat. Abnormal pork pH will lead to the production of PSE (Pale, Soft, Exudative) or DFD (Dark, Firm, Dry) meat. We identified three significant SNPs as being significantly associated with pH45. Among which, the significant SNP (rs321002713) on SSC18 was located within glutamate metabotropic receptor 8 (*GRM8*). The *GRM8* gene encodes a G protein-coupled metabotropic glutamate receptor involved in glutamatergic neurotransmission in the central nervous system (Nakanishi, 1994; Duvoisin et al., 1995). Group III of the eight different metabotropic glutamate receptors, which are connected to the suppression of the cyclic AMP cascade, includes the GRM8 receptor. (Nakanishi, 1992). A study finds that *GRM8* is a porcine candidate gene related to muscling and a SNP in the *GRM8* gene also displayed a strong association with the loin eye area of pigs (Li et al., 2011). *GRM8* was also associated with the relative area of *longissimus dorsi* muscle fiber type I and was considered a plausible candidate gene for this trait (Guo et al., 2020). Perhaps, the *GRM8* gene expressed in *longissimus dorsi* muscle may be a potential candidate gene for porcine pH traits.

Candidate genes for MC

Meat color is a complex trait that depends on the amount of pigment present, the muscle tissue's structural characteristics, and the pace of muscle acidification (Fan et al., 2008; Mármol-Sánchez et al., 2020). The significant SNP (rs327814455) on SSC1 was located within Ankyrin repeat domain-containing protein 6 (*ANKRD6*). *ANKRD6* belongs to the ankyrins gene family. Ankyrins are a family of structural proteins that include binding sites for cytoskeleton proteins and a variety of integral membranes (Gallagher et al., 1997). Ankyrin interactions allow the cytoskeleton to be attached to the plasma membrane (Rubtsov and Lopina, 2000). Van Deveire et al. (2012) have demonstrated that *ANKRD6* is related to the cross-sectional area of human muscle. Particular muscle phenotypes have been linked in certain studies to genetic variations in the Ankyrin genes. A study shows that SNPs in the bovine Ankyrin 1 (*ANK1*) promoter region have been linked to intramuscular fat levels and tenderness of beef (Horodyska et al., 2015). SNPs in pig *ANK1* show relationships with shear force, pH, water-holding capacity, and intramuscular fat (IMF) (Wimmers et al., 2007). In pig muscle with excessive fat, the Ankyrin repeat and sterile alpha motif domain containing 1B (*ANKS1B*) gene was found to be a significantly upregulated expression (Hamill et al., 2012). Additionally, it has been discovered that the expression of Ankyrin repeat domain 1 (*ANKRD1*) in pig muscle correlates with the ultimate pH (Damon et al., 2013). Consequently, the Ankyrin gene *ANKRD6* should be considered a strong candidate gene for the porcine multiple meat quality traits, containing MC.

Candidate genes for MA

The marbling score is closely related to intramuscular fat content (IMF). A low marbling score will affect the pork quality and flavor. The most significant SNP (rs696643958) on SSC1 was located within *ENSSSCG00000004081*. GO annotation result showed that the gene participated in muscle cell differentiation and actin filament binding (Supplementary Table S3). The deposition of fat in muscle is closely related to the growth and development of the muscle (Lai et al., 2004). Thus, the gene may be involved in growth of the muscles and thus affect the fat deposition. On SSC17, one significant SNP (rs341748571) was located within Mono-ADP ribosylhydrolase 2 (*MACROD2*). The *MACROD2* gene encodes the mono-ADP-ribosyltransferase two catalyzing ADP-ribosylation (Feijs et al., 2013). ADP-ribosylation is a post-translational modification participating in a number of biological processes, such as the regulation of transcription, immune cell function, and DNA repair (Kraus and Hottiger, 2013). Some studies find the *MACROD2* gene located at BTA13 which is related to net meat weight in beef cattle (Niu et al., 2021) and may also be affected meat color traits in Nellore cattle (Marin-Garzon et al., 2021). Besides, Ma et al. (2019) find that the *MACROD2* gene may affect porcine backfat thickness traits by affecting fat metabolism. Therefore, the *MACROD2* gene can be considered a candidate gene for the porcine MA.

Another significant SNP (rs342013877) on SSC13 was located 5 kb away from ATP binding cassette subfamily G member 1 (*ABCG1*). In the study, GO annotation results showed that the *ABCG1* gene was involved in negative regulation of lipid storage, response to lipid and phospholipid homeostasis. The *ABCG1* gene has been known to be associated with controlling cellular lipid levels (Kennedy et al., 2005). Adipocyte *ABCG1* can promote lipid accumulation by regulating the lipoprotein lipase (LPL) bioavailability and fat mass growth in a triglyceride (TG)-rich environment (Frisdal et al., 2015). Thus, the *ABCG1* gene also can be considered a strong candidate gene for the pork MA based on its biological functions.

Candidate genes for WL

Pork WL is closely related to the water holding capacity of meat, which is affected by the speed and degree of pH decline, protein hydrolysis and even protein oxidation post-mortem (Huff-Lonergan and Lonergan, 2005). The MLM identified the most significant SNPs on SSC6 for WL and the SNP was located in Transmembrane protein 50A (*TMEM50A*). A study shows that the related gene *TMEM217* is associated with meat color (Ma et al., 2013). Besides, in mice, adipocyte metabolism and differentiation are impacted by the related genes *TMEM120A* and *TMEM120B*, which are significantly expressed in fat (Batrakou et al., 2015). Additionally, *TMEM60 and TMEM236* are two other homologous genes related to marbling fat and fat color in cattle, respectively (Lim et al., 2014). Although no studies have shown that *TMEM50A* played a role in meat quality, it might be regarded as a possible candidate gene for WL. The significant SNP on SSC15 was located 2.7 kb upstream of ribonucleotide reductase regulatory subunit M2 (*RRM2*). The result of GO annotation showed that *RRM2* was involved in deoxyribonucleotide biosynthetic process and oxidation-reduction process (Supplementary Table S3). A study finds that inhibitors of *RRM2* can inhibit cell proliferation (Heidel et al., 2007). At present, there was no direct evidence to prove that *RRM2* was related to WL.

Candidate genes for CL

The CL can affect the juiciness and appearance of the pork (Aaslyng et al., 2003). The two adjacent SNPs on SSC10 for CL were located within phosphatidylinositol-5-phosphate 4-kinase type 2 alpha (*PIP4K2A*). Previous studies have shown that the two SNPs (ASGA0048292 and ASGA0048295) of *PIP4K2A* were associated with meat quality of pigs (Lee et al., 2014). *PIP4K2A* is related to the fatty acid composition of backfat in three crossbred pigs (Crespo-Piazuelo et al., 2020). The *PIP4K2A* gene controls the body responsiveness to insulin, and mutations in the *PIP4K2A* gene can make the skeletal muscle more sensitive to insulin (Carricaburu et al., 2003). This directly leads to an increase insulin-stimulated glucose transport in muscle (Lamia et al., 2004). Perhaps, *PIP4K2A* might influence meat quality-related traits by affecting glucose transport in muscle. Thus, *PIP4K2A* could be considered a candidate gene for CL.

Candidate genes for DL

Drip loss is one of the important indicators to assess pork quality, which is related to ultimate pH, rate of post-mortem pH fall, residual ATP levels, glycolysis rate post-mortem, and activity of several enzymes (Lawrie and Ledward, 2006). The most significant SNP (rs321165533) on SSC6 for DL was located within chromodomain Y-like 2 (*CDYL2*). GO annotation results showed that *CDYL2* was involved in catalytic activity and metabolic processes. A study finds that *CDYL2* is related to porcine teat number (Liu et al., 2022).

Two nearby SNPs (rs703586532 and rs323693055) on SSC13 were located in cell adhesion molecule L1 like (*CHL1*). The study finds that *CHL1* can regulate the cell cycle *via* the p53 pathway and inhibit cell proliferation through the ERK pathway, and was associated with insulin secretion and glucose metabolism (Jiang et al., 2020). Thus, *CHL1* can be considered a strong candidate gene for DL. The SNP rs320599347 on SSC4 was located within ATP binding cassette subfamily A member 4 (*ABCA4*). *ABCA4* is a member of the ABCA subfamily of ATP-binding cassette transporters participating in the transport of phosphatidylethanolamine (Quazi and Molday, 2013). GO annotation result showed that *ABCA4* participated in phospholipid-translocating ATPase activity, phospholipid translocation, and phospholipid transfer to membrane (Supplementary Table S3). On SSC6, two adjacent significant SNPs were located within fatty acid 2-hydroxylase (*FA2H*), which was participated in fatty acid biosynthetic process, lipid modification, and regulation of cell proliferation (Supplementary Table S3). In 3T3-L1 adipocytes, *FA2H* modulates the diffusional mobility of lipids linked with Raft and lipogenesis (Guo et al., 2010).

Furthermore, four nearby significant SNPs on SSC3 were located in a region of 0.26 kb, which were located within zinc-alpha-2-glycoprotein (*ZAG*), which is a glycoprotein included in the class I family of the major histocompatibility complex (MHC). Several studies show that *ZAG* is related to lipid loss (Bao et al., 2005) and lipid metabolism (Garrido-Sanchez et al., 2012) and also stimulates the expression of adiponectin (Gohda et al., 2003). Besides, two adjacent significant SNPs on SSC2 were located in solute carrier family 1 member 2 (*SLC1A2*). Researchers report that the related genes *SLC15A4* c.658AA genotype has better water-holding capacity (D'Astous-Page et al., 2017). Besides, some previous studies find that genes of the solute carrier family (SLC), such as *SLC25A17* and *SLC9A7* are associated with meat color, drip loss, and intramuscular fat, respectively (Ma et al., 2013), and *SLC37A3* and *SLC24A5* are related to meat color (Iqbal et al., 2015; Gao et al., 2021), and *SLC4A8* and *SLC7A10* are associated with purge loss (Nonneman et al., 2013). In addition, it has been reported that *SLC37A4* and *SLC3A2* are promising candidate genes affecting DL (Ponsuksili et al., 2008; Heidt et al., 2013; Zhao et al., 2019). A large of research suggesting genes of the solute carrier family play important role in regulating DL. Thus, it was inferred that the *SLC1A2* gene could be considered a strong candidate gene for pork DL. Finally, the rs327708082 on SSC2 explained the highest DL phenotypic variance (16.32%), which was located in the SIL1 nucleotide exchange factor (*SIL1*) gene. *SIL1* related to stress protection, and moderately increased *SIL1* also ameliorates cellular fitness under stress conditions (Labisch et al., 2018).

However, more pig populations need to be used to verify these SNP loci and candidate genes, and more pig biological experiments need to be conducted to confirm their functions.

Conclusion

We conducted a GWAS based on SLAF-seq for six meat-quality traits in 223 four-way crossbred pigs. A total of 64 SNPs distributed on 16 chromosomes were identified using MLM ($p < 10^{-5}$), of which 24 SNPs were located in previously reported QTL regions. Three QTLs were identified to be related to DL: 0.08-Mb region on SSC5 (72.91–72.99Mb), 3.6-Mb region on SSC13 (53.28–56.88Mb), and 0.09-Mb region on SSC9 (63.38–63.47Mb). Some novel candidate genes for meat quality traits were identified, including pH45 (*GRM8*), MC (*ANKRD6*), MA (*MACROD2* and *ABCG1*), WL (*TMEM50A*), CL (*PIP4K2A*), and DL (*CDYL2*, *CHL1*, *ABCA4*, *ZAG* and *SLC1A2*). Overall, the study presented substantial new evidence for the involvement of several candidate genes in different pork quality traits. These SNPs and candidate genes identified in the study provided a basis for molecular marker-assisted breeding and improvement for meat quality traits in pigs.

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/, SRP376933.

Ethics statement

The animal study was reviewed and approved by the ethics committee of Yunnan Agricultural University (YNAU, Kunming, China). Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

The experiment was conceived and designed by SL and YP. The ear tissues were gathered and the phenotypic information of meat quality traits was determined by ML, XW, DY, and XD. The experiment was carried out by HW, who also processed and analyzed the data. Data processing was aided by ML, HS, and QC The manuscript was written by HW and XW and afterward amended by YP and SL. The final manuscript has been reviewed and approved by all authors.

Funding

This research was funded by the National Natural Science Foundation of China (U1402266), Yunnan Province Important National Science & Technology Specific Projects (202102AE090039), and Yunnan Swine Industry Technology System Program (2020KJTX0016). These funding agencies had no role in the research design, data gathering, data analysis, or writing the manuscript.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1001352/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Establishment flow plot of (Duroc×Saba) × [Yorkshire × (Landrace × Saba)] crossbred pig population.

SUPPLEMENTARY FIGURE S2
Frequency distribution histogram for six meat quality traits. **(A)** PH at 45 min post mortem (PH45). **(B)** Meat color score (MC). **(C)** Marbling score (MA). **(D)** Water loss rate (WL). **(E)** Cooking loss (CL). **(F)** Drip loss (DL).

SUPPLEMENTARY FIGURE S3
The density distribution of total SNPs and filtered SNPs on Sus Scrofa chromosomes. **(A)** The number of total SNPs within 1 Mb window size. **(B)** The number of filtered SNPs within 1 Mb window size. The horizontal axis (X-axis) showed the chromosome length (Mb). The color index indicated the number of labels.

References

Aaslyng, M. D., Bejerholma, C., Ertbjergb, P., Bertramc, H. C., and Andersenc, H. J. (2003). Cooking loss and juiciness of pork in relation to raw meat quality and cooking procedure. *Food Qual. Prefer.* 14, 277–288. doi:10.1016/s0950-3293(02)00086-1

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

Bao, Y., Bing, C., Hunter, L., Jenkins, J. R., Wabitsch, M., and Trayhurn, P. (2005). Zinc-alpha2-glycoprotein, a lipid mobilizing factor, is expressed and secreted by human (SGBS) adipocytes. *FEBS Lett.* 579, 41–47. doi:10.1016/j.febslet.2004.11.042

Batrakou, D. G., de Las Heras, J. I., Czapiewski, R., Mouras, R., and Schirmer, E. C. (2015). TMEM120A and B: Nuclear envelope transmembrane proteins important for adipocyte differentiation. *PLoS One* 10, e0127712. doi:10.1371/journal.pone.0127712

Carricaburu, V., Lamia, K. A., Lo, E., Favereaux, L., Payrastre, B., Cantley, L. C., et al. (2003). The phosphatidylinositol (PI)-5-phosphate 4-kinase type II enzyme controls insulin signaling by regulating PI-3,4,5-trisphosphate degradation. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9867–9872. doi:10.1073/pnas.1734038100

Cho, I. C., Yoo, C. K., Lee, J. B., Jung, E. J., Han, S. H., Lee, S. S., et al. (2015). Genome-wide QTL analysis of meat quality-related traits in a large F2 intercross between Landrace and Korean native pigs. *Genet. Sel. Evol.* 47, 7. doi:10.1186/s12711-014-0080-6

Choi, B. H., Lee, Y. M., Alam, M., Lee, J. H., Kim, T. H., Kim, K. S., et al. (2011). Detection of mendelian and parent-of-origin quantitative trait loci for meat quality in a cross between Korean native pig and Landrace. *Asian-Aust J. Anim. Sci.* 12, 1644–1650. doi:10.5713/ajas.2011.11166

Crespo-Piazuelo, D., Criado-Mesas, L., Revilla, M., Castello, A., Noguera, J. L., Fernandez, 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, 13962. doi:10.1038/s41598-020-70894-2

D'Astous-Page, J., Gariepy, C., Blouin, R., Cliche, S., Methot, S., Sullivan, B., et al. (2017). Identification of single nucleotide polymorphisms in carnosine-related genes and effects of genotypes on pork meat quality attributes. *Meat Sci.* 134, 54–60. doi:10.1016/j.meatsci.2017.07.019

Damon, M., Denieul, K., Vincent, A., Bonhomme, N., Wyszynska-Koko, J., and Lebret, B. (2013). Associations between muscle gene expression pattern and technological and sensory meat traits highlight new biomarkers for pork quality assessment. *Meat Sci.* 95, 744–754. doi:10.1016/j.meatsci.2013.01.016

Diao, S., Huang, S., Chen, Z., Teng, J., Ma, Y., Yuan, X., et al. (2019). Genome-wide signatures of selection detection in three South China indigenous pigs. *Genes (Basel)* 10, 346. doi:10.3390/genes10050346

Ding, R., Yang, M., Quan, J., Li, S., Zhuang, Z., Zhou, S., et al. (2019). Single-locus and multi-locus genome-wide association studies for intramuscular fat in Duroc pigs. *Front. Genet.* 10, 619. doi:10.3389/fgene.2019.00619

Dong, S. S., He, W. M., Ji, J. J., Zhang, C., Guo, Y., and Yang, T. L. (2021). LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief Bioinform.* 22, bbaa227. doi:10.1093/bib/bbaa227

Duvoisin, R. M., Zhang, C., and Ramonell, K. (1995). A novel metabotropic glutamate receptor expressed in the retina and olfactory bulb. *J. Neurosci.* 15, 3075–3083. doi:10.1523/JNEUROSCI.15-04-03075.1995

Edwards, D. B., Ernst, C. W., Raney, N. E., Doumit, M. E., Hoge, M. D., and Bates, R. O. (2008). Quantitative trait locus mapping in an F2 Duroc x pietrain resource population: II. Carcass and meat quality traits. *J. Anim. Sci.* 86, 254–266. doi:10.2527/jas.2006-626

Fabbri, M. C., Zappaterra, M., Davoli, R., and Zambonelli, P. (2020). Genome-wide association study identifies markers associated with carcass and meat quality traits in Italian Large White pigs. *Anim. Genet.* 51, 950–952. doi:10.1111/age.13013

Fan, B., Glenn, K. L., Geiger, B., Mileham, A., and Rothschild, M. F. (2008). Investigation of QTL regions on chromosome 17 for genes associated with meat color in the pig. *J. Anim. Breed. Genet.* 125, 240–247. doi:10.1111/j.1439-0388.2008.00749.x

Fan, B., Lkhagvadorj, S., Cai, W., Young, J., Smith, R. M., Dekkers, J. C., et al. (2010). Identification of genetic markers associated with residual feed intake and meat quality traits in the pig. *Meat Sci.* 84, 645–650. doi:10.1016/j.meatsci.2009.10.025

Farouk, M. M., and Wieliczko, K. J. (2003). Effect of diet and fat content on the functional properties of thawed beef. *Meat Sci.* 64, 451–458. doi:10.1016/S0309-1740(02)00214-0

Feijs, K. L., Forst, A. H., Verheugd, P., and Luscher, B. (2013). Macrodomain-containing proteins: Regulating new intracellular functions of mono(ADP-ribosyl)ation. *Nat. Rev. Mol. Cell Biol.* 14, 443–451. doi:10.1038/nrm3601

Fernández-Barroso, M. Á., Silió, L., Rodríguez, C., Palma-Granados, P., López, A., Caraballo, C., et al. (2020). Genetic parameter estimation and gene association analyses for meat quality traits in open-air free-range Iberian pigs. *J. Anim. Breed. Genet.* 137, 581–598. doi:10.1111/jbg.12498

Frisdal, E., Le Lay, S., Hooton, H., Poupel, L., Olivier, M., Alili, R., et al. (2015). Adipocyte ATP-binding cassette G1 promotes triglyceride storage, fat mass growth, and human obesity. *Diabetes* 64, 840–855. doi:10.2337/db14-0245

Fu, L., Jiang, Y., Wang, C., Mei, M., Zhou, Z., Jiang, Y., et al. (2020). A genome-wide association study on feed efficiency related traits in Landrace pigs. *Front. Genet.* 11, 692. doi:10.3389/fgene.2020.00692

Gallagher, P. G., Tse, W. T., Scarpa, A. L., Lux, S. E., and Forget, B. G. (1997). Structure and organization of the human ankyrin-1 gene. Basis for complexity of pre-mRNA processing. *J. Biol. Chem.* 272, 19220–19228. doi:10.1074/jbc.272.31.19220

Gallardo, D., Pena, R. N., Quintanilla, R., Ramirez, O., Almuzara, D., Noguera, J. L., et al. (2012). Quantitative trait loci analysis of a Duroc commercial population highlights differences in the genetic determination of meat quality traits at two different muscles. *Anim. Genet.* 43, 800–804. doi:10.1111/j.1365-2052.2012.02333.x

Gao, G. X., Gao, N., Li, S. C., Kuang, W. J., Zhu, L., Jiang, W., et al. (2021). Genome-wide association study of meat quality traits in a three-way crossbred commercial pig population. *Front. Genet.* 12, 614087. doi:10.3389/fgene.2021.614087

Garrido-Sanchez, L., García-Fuentes, E., Fernández-García, D., Escote, X., Alcaide, J., Perez-Martinez, P., et al. (2012). Zinc-alpha 2-glycoprotein gene expression in adipose tissue is related with insulin resistance and lipolytic genes in morbidly obese patients. *PLoS One* 7, e33264. doi:10.1371/journal.pone.0033264

Gjerlaug-Enger, E., Aass, L., Odegård, J., and Vangen, O. (2010). Genetic parameters of meat quality traits in two pig breeds measured by rapid methods. *Animal* 4, 1832–1843. doi:10.1017/S175173111000114X

Gohda, T., Makita, Y., Shike, T., Tanimoto, M., Funabiki, K., Horikoshi, S., et al. (2003). Identification of epistatic interaction involved in obesity using the KK/Ta mouse as a type 2 diabetes model: Is Zn-α2 glycoprotein-1 a candidate gene for obesity? *Diabetes* 52, 2175–2181. doi:10.2337/diabetes.52.8.2175

Guo, L., Zhou, D., Pryse, K. M., Okunade, A. L., and Su, X. (2010). Fatty acid 2-hydroxylase mediates diffusional mobility of Raft-associated lipids, GLUT4 level, and lipogenesis in 3T3-L1 adipocytes. *J. Biol. Chem.* 285, 25438–25447. doi:10.1074/jbc.M110.119933

Guo, T., Gao, J., Yang, B., Yan, G., Xiao, S., Zhang, Z., et al. (2020). A whole genome sequence association study of muscle fiber traits in a White Duroc × Erhualian F2 resource population. *Asian-Australas J. Anim. Sci.* 33, 704–711. doi:10.5713/ajas.18.0767

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, 21. doi:10.1186/s12711-017-0295-4

Hamill, R. M., McBryan, J., McGee, C., Mullen, A. M., Sweeney, T., Talbot, A., et al. (2012). Functional analysis of muscle gene expression profiles associated with tenderness and intramuscular fat content in pork. *Meat Sci.* 92, 440–450. doi:10.1016/j.meatsci.2012.05.007

Harmegnies, N., Davin, F., De Smet, S., Buys, N., Georges, M., and Coppieters, W. (2006). Results of a whole-genome quantitative trait locus scan for growth, carcass composition and meat quality in a porcine four-way cross. *Anim. Genet.* 37, 543–553. doi:10.1111/j.1365-2052.2006.01523.x

Heidel, J. D., Liu, J. Y.-C., Yen, Y., Zhou, B., Heale, B. S., Rossi, J. J., et al. (2007). Potent siRNA inhibitors of ribonucleotide reductase subunit RRM2 reduce cell proliferation *in vitro* and *in vivo*. *Clin. Cancer Res.* 13, 2207–2215. doi:10.1158/1078-0432.CCR-06-2218

Heidt, H., Cinar, M. U., Uddin, M. J., Looft, C., Jungst, H., Tesfaye, D., et al. (2013). A genetical genomics approach reveals new candidates and confirms known candidate genes for drip loss in a porcine resource population. *Mamm. Genome* 24, 416–426. doi:10.1007/s00335-013-9473-z

Honikel, K. O. (1987). The water binding of meat. *Fleischwirtzchaft* 67, 1098–1102.

Horodyska, J., Sweeney, T., Ryan, M., and Hamill, R. M. (2015). Novel SNPs in the Ankyrin 1 gene and their association with beef quality traits. *Meat Sci.* 108, 88–96. doi:10.1016/j.meatsci.2015.04.019

Huff-Lonergan, E., and Lonergan, S. M. (2005). Mechanisms of water-holding capacity of meat: The role of postmortem biochemical and structural changes. *Meat Sci.* 71, 194–204. doi:10.1016/j.meatsci.2005.04.022

Iqbal, A., Kim, Y. S., Kang, J. M., Lee, Y. M., Rai, R., Jung, J. H., et al. (2015). Genome-wide association study to identify quantitative trait loci for meat and carcass quality traits in Berkshire. *Asian-Australas J. Anim. Sci.* 28, 1537–1544. doi:10.5713/ajas.15.0752

Ji, J., Zhou, L., Guo, Y., Huang, L., and Ma, J. (2017). Genome-wide association study identifies 22 new loci for body dimension and body weight traits in a White Duroc × Erhualian F2 intercross population. *Asian-Australas J. Anim. Sci.* 30, 1066–1073. doi:10.5713/ajas.16.0679

Ji, J., Zhou, L., Huang, Y., Zheng, M., Liu, X., Zhang, Y. A., et al. (2018). A whole-genome sequence based association study on pork eating quality traits and cooking loss in a specially designed heterogeneous F6 pig population. *Meat Sci.* 146, 160–167. doi:10.1016/j.meatsci.2018.08.013

Jiang, H., Liu, Y., Qian, Y., Shen, Z., He, Y., Gao, R., et al. (2020). CHL1 promotes insulin secretion and negatively regulates the proliferation of pancreatic β cells. *Biochem. Biophys. Res. Commun.* 525, 1095–1102. doi:10.1016/j.bbrc.2020.03.040

Kennedy, M. A., Barrera, G. C., Nakamura, K., Baldán, Á., Tarr, P., Fishbein, M. C., et al. (2005). ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation. *Cell Metab.* 1, 121–131. doi:10.1016/j.cmet.2005.01.002

Khanal, P., Maltecca, C., Schwab, C., Gray, K., and Tiezzi, F. (2019). Genetic parameters of meat quality, carcass composition, and growth traits in commercial swine. *J. Anim. Sci.* 97, 3669–3683. doi:10.1093/jas/skz247

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi:10.1128/AEM.01043-13

Kraus, W. L., and Hottiger, M. O. (2013). PARP-1 and gene regulation: Progress and puzzles. *Mol. Asp. Med.* 34, 1109–1123. doi:10.1016/j.mam.2013.01.005

Labisch, T., Buchkremer, S., Phan, V., Kollipara, L., and Gatz, C., (2018). Tracking effects of SIL1 increase: Taking a closer look beyond the consequences of elevated expression level. *Mol. Neurobiol.* 55, 2524–2546. doi:10.1007/s12035-017-0494-6

Lai, K. M., Gonzalez, M., Poueymirou, W. T., Kline, W. O., Na, E., Zlotchenko, E., et al. (2004). Conditional activation of Akt in adult skeletal muscle induces rapid hypertrophy. *Mol. Cell Biol.* 24, 9295–9304. doi:10.1128/MCB.24.21.9295-9304.2004

Lamia, K. A., Peroni, O. D., Kim, Y. B., Rameh, L. E., Kahn, B. B., and Cantley, L. C. (2004). Increased insulin sensitivity and reduced adiposity in phosphatidylinositol 5-phosphate 4-kinase β−/− mice. *Mol. Cell Biol.* 24, 5080–5087. doi:10.1128/MCB.24.11.5080-5087.2004

Lawrie, R. A., and Ledward, D. (2006). *Lawrie's meat science*. Duxford, UK: Woodhead Publishing in Food Science Technology & Nutrition.

Lee, J. H., Song, K. D., Lee, H. K., Cho, K. H., Park, H. C., and Park, K. D. (2015). Genetic parameters of reproductive and meat quality traits in Korean Berkshire pigs. *Asian-Australas J. Anim. Sci.* 28, 1388–1393. doi:10.5713/ajas.15.0097

Lee, K. T., Lee, Y. M., Alam, M., Choi, B., Park, M., Kim, K. S., et al. (2012). A whole genome association study on meat quality traits using high density SNP chips in a cross between Korean native pig and Landrace. *Asian-Australasian J. Anim. Sci.* 25, 1529–1539. doi:10.5713/ajas.2012.12474

Lee, T., Shin, D. H., Cho, S., Kang, H. S., Kim, S. H., Lee, H. K., et al. (2014). Genome-wide association study of integrated meat quality-related traits of the Duroc pig breed. *Asian-Australas J. Anim. Sci.* 27, 303–309. doi:10.5713/ajas.2013.13385

Li, F., Liu, J., Liu, W., Gao, J., Lei, Q., Han, H., et al. (2021). Genome-wide association study of body size traits in Wenshang Barred chickens based on the specific-locus amplified fragment sequencing technology. *J. Anim. Sci.* 92, e13506. doi:10.1111/asj.13506

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, X., Kim, S. W., Choi, J. S., Lee, Y. M., Lee, C. K., Choi, B. H., et al. (2010). Investigation of porcine FABP3 and LEPR gene polymorphisms and mRNA expression for variation in intramuscular fat content. *Mol. Biol. Rep.* 37, 3931–3939. doi:10.1007/s11033-010-0050-1

Li, X., Kim, S. W., Do, K. T., Ha, Y. K., Lee, Y. M., Yoon, S. H., et al. (2011). Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Mol. Biol. Rep.* 38, 3805–3820. doi:10.1007/s11033-010-0496-1

Li, Z., Wei, S., Li, H., Wu, K., Cai, Z., Li, D., et al. (2017). Genome-wide genetic structure and differentially selected regions among Landrace, Erhualian, and Meishan pigs using specific-locus amplified fragment sequencing. *Sci. Rep.* 7, 10063. doi:10.1038/s41598-017-09969-6

Lim, D., Kim, N. K., Lee, S. H., Park, H. S., Cho, Y. M., Chai, H. H., et al. (2014). Characterization of genes for beef marbling based on applying gene coexpression network. *Int. J. genomics* 2014, 708562. doi:10.1155/2014/708562

Liu, G. S., Kim, J. J., Jonas, E., Wimmers, K., Ponsuksili, S., Murani, E., et al. (2008). Combined line-cross and half-sib QTL analysis in Duroc-Pietrain population. *Mamm. genome* 19, 429–438. doi:10.1007/s00335-008-9132-y

Liu, Q., Yue, J., Niu, N., Liu, X., Yan, H., Zhao, F. P., 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. *Anim. (Basel)* 10, 2186. doi:10.3390/ani10112186

Liu, X., Wang, L. G., Liang, J., Yan, H., Zhao, K. B., Li, N., et al. (2014). Genome-wide association study for certain carcass traits and organ weights in a Large White × Minzhu intercross porcine population. *J. Integr. Agr.* 13, 2721–2730. doi:10.1016/s2095-3119(14)60787-5

Liu, X., Xiong, X., Yang, J., Zhou, L., Yang, B., Ai, H., et al. (2015). Genome-wide association analyses for meat quality traits in Chinese Erhualian pigs and a Western Duroc × (Landrace × Yorkshire) commercial population. *Genet. Sel. Evol.* 47, 44. doi:10.1186/s12711-015-0120-x

Liu, Z., Li, H., Zhong, Z., and Jiang, S. (2022). A whole genome sequencing-based genome-wide association study reveals the potential associations of teat number in Qingping pigs. *Anim. (Basel)* 12, 1057. doi:10.3390/ani12091057

Lo, L. L., McLaren, D. G., McKeith, F. K., Fernando, R. L., and Novakofski, J. (1992). Genetic analyses of growth, real-time ultrasound, carcass, and pork quality traits in Duroc and Landrace pigs: II. Heritabilities and correlations. *J. Anim. Sci.* 70, 2387–2396. doi:10.2527/1992.7082387x

Luo, W., Cheng, D., Chen, S., Wang, L., Li, Y., Ma, X., et al. (2012). Genome-wide association analysis of meat quality traits in a porcine Large White × Minzhu intercross population. *Int. J. Biol. Sci.* 8, 580–595. doi:10.7150/ijbs.3614

Ma, H., Zhang, S., Zhang, K., Zhan, H., Peng, X., Xie, S., et al. (2019). Identifying selection signatures for backfat thickness in Yorkshire pigs highlights new regions affecting fat metabolism. *Genes (Basel).* 10, 254. doi:10.3390/genes10040254

Ma, J., Yang, J., Zhou, L., Ren, J., Liu, X., Zhang, H., et al. (2014). A splice mutation in the PHKG1 gene causes high glycogen content and low meat quality in pig skeletal muscle. *PLoS Genet.* 10, e1004710. doi:10.1371/journal.pgen.1004710

Ma, J., Yang, J., Zhou, L., Zhang, Z., Ma, H., Xie, X., et al. (2013). Genome-wide association study of meat quality traits in a White Duroc × Erhualian F2 intercross and Chinese Sutai pigs. *PLoS One* 8, e64047. doi:10.1371/journal.pone.0064047

Malek, M., Dekkers, J. C., Lee, H. K., Baas, T. J., Prusa, K., Huff-Lonergan, E., et al. (2001). A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. *Mamm. genome* 12, 637–645. doi:10.1007/s003350020019

Mandozai, A., Moussa, A. A., Zhang, Q., Qu, J., Du, Y. Y., Anwari, G., et al. (2021). Genome-wide association study of root and shoot related traits in Spring Soybean (Glycine max L.) at seedling stages using SLAF-Seq. *Front. Plant Sci.* 12, 568995. doi:10.3389/fpls.2021.568995

Marin-Garzon, N. A., Magalhaes, A. F. B., Mota, L. F. M., Fonseca, L. F. S., Chardulo, L. A. L., and Albuquerque, L. G. (2021). Genome-wide association study identified genomic regions and putative candidate genes affecting meat color traits in Nellore cattle. *Meat Sci.* 171, 108288. doi:10.1016/j.meatsci.2020.108288

Mármol-Sánchez, E., Quintanilla, R., Jordana, J., and Amills, M. (2020). An association analysis for 14 candidate genes mapping to meat quality quantitative trait loci in a Duroc pig population reveals that the *ATP1A2* genotype is highly associated with muscle electric conductivity. *Anim. Genet.* 51, 95–100. doi:10.1111/age.12864

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

Melak, S., Wang, Q., Tian, Y., Wei, W., Zhang, L., Elbeltagy, A., et al. (2021). Identification and validation of marketing weight-related SNP markers using SLAF sequencing in male Yangzhou Geese. *Genes (Basel).* 12, 1203. doi:10.3390/genes12081203

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

Milan, D., Jeon, J. T., Looft, C., Amarger, V., Robic, A., Thelander, M., et al. (2000). A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* 288, 1248–1251. doi:10.1126/science.288.5469.1248

Nakanishi, S. (1994). Metabotropic glutamate receptors: Synaptic transmission, modulation, and plasticity. *Neuron* 13, 1031–1037. doi:10.1016/0896-6273(94)90043-4

Nakanishi, S. (1992). Molecular diversity of glutamate receptors and implications for brain function. *Science* 258, 597–603. doi:10.1126/science.1329206

Niu, Q., Zhang, T., Xu, L., Wang, T., Wang, Z., Zhu, B., et al. (2021). Integration of selection signatures and multi-trait GWAS reveals polygenic genetic architecture of carcass traits in beef cattle. *Genomics* 113, 3325–3336. doi:10.1016/j.ygeno.2021.07.025

Noidad, S., Limsupavanich, R., Suwonsichon, S., and Chaosap, C. (2019). Effect of visual marbling levels in pork loins on meat quality and Thai consumer acceptance and purchase intent. *Asian-Australas J. Anim. Sci.* 32, 1923–1932. doi:10.5713/ajas.19.0084

Nonneman, D., Shackelford, S., King, D., Wheeler, T., Wiedmann, R., Snelling, W., et al. (2013). Genome-wide association of meat quality traits and tenderness in swine. *J. Anim. Sci.* 91, 4043–4050. doi:10.2527/jas.2013-6255

Park, J., Lee, S. M., Park, J. Y., and Na, C. S. (2021). A genome-wide association study (GWAS) for pH value in the meat of Berkshire pigs. *J. Anim. Sci. Technol.* 63, 25–35. doi:10.5187/jast.2021.e17

Percie du Sert, N., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., et al. (2020). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLoS Biol.* 18, e3000411. doi:10.1371/journal.pbio.3000411

Ponsuksili, S., Jonas, E., Murani, E., Phatsara, C., Srikanchai, T., Walz, C., et al. (2008). Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle. *BMC Genom* 9, 367. doi:10.1186/1471-2164-9-367

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Qi, Z., Huang, L., Zhu, R., Xin, D., Liu, C., Han, X., et al. (2014). A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS One* 9, e104871. doi:10.1371/journal.pone.0104871

Qiao, R., Gao, J., Zhang, Z., Li, L., Xie, X., Fan, Y., et al. (2015). Genome-wide association analyses reveal significant loci and strong candidate genes for growth and fatness traits in two pig populations. *Genet. Sel. Evol.* 47, 17. doi:10.1186/s12711-015-0089-5

Qin, M., Li, C., Li, Z., Chen, W., and Zeng, Y. (2020). Genetic diversities and differentially selected regions between Shandong indigenous pig breeds and Western pig breeds. *Front. Genet.* 10, 1351. doi:10.3389/fgene.2019.01351

Quazi, F., and Molday, R. S. (2013). Differential phospholipid substrates and directional transport by ATP-binding cassette proteins ABCA1, ABCA7, and ABCA4 and disease-causing mutants. *J. Biol. Chem.* 288, 34414–34426. doi:10.1074/jbc.M113.508812

Rohrer, G. A., Thallman, R. M., Shackelford, S., Wheeler, T., and Koohmaraie, M. (2005). A genome scan for loci affecting pork quality in a Duroc-Landrace F population. *Anim. Genet.* 37, 17–27. doi:10.1111/j.1365-2052.2005.01368.x

Rubtsov, A. M., and Lopina, O. D. (2000). *Ankyrins. FEBS Lett.* 482, 1–5. doi:10.1016/s0014-5793(00)01924-4

Škrlep, M., Kavar, T., and Čandek-Potokar, M. (2010). Comparison of PRKAG3 and RYR1 gene effect on carcass traits and meat quality in Slovenian commercial pigs. *Czech J. Anim. Sci.* 55, 149–159. doi:10.17221/6/2009-cjas

Su, Y. H., Xiong, Y. Z., Jiang, S. W., Zhang, Q., Lei, M. G., Zheng, R., et al. (2004). [Mapping quantitative trait loci for meat quality trait in a Large White × Meishan cross]. *Acta Genet. Sin.* 31, 132–136.

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-Seq: An efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8, e58700. doi:10.1371/journal.pone.0058700

Thomsen, H., Lee, H. K., Rothschild, M. F., Malek, M., and Dekkers, J. C. M. (2004). Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. *J. Anim. Sci.* 82, 2213–2228. doi:10.2527/2004.8282213x

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Biorxiv* 7, e1002043. doi:10.1101/005165

Van Deveire, K. N., Scranton, S. K., Kostek, M. A., Angelopoulos, T. J., Clarkson, P. M., Gordon, P. M., et al. (2012). Variants of the ankyrin repeat domain 6 gene (ANKRD6) and muscle and physical activity phenotypes among European-derived American adults. *J. Strength Cond. Res.* 26, 1740–1748. doi:10.1519/JSC.0b013e31825c2bef

Wang, H., Wang, X., Li, M., Sun, H., Chen, Q., Yan, D., et al. (2022b). Genome-wide association study of growth traits in a four-way crossbred pig population. *Genes (Basel).* 13, 1990. doi:10.3390/genes13111990

Wang, H., Wang, X., Yan, D., Sun, H., Chen, Q., Li, M., et al. (2022a). Genome-wide association study identifying genetic variants associated with carcass backfat thickness, lean percentage and fat percentage in a four-way crossbred pig population using SLAF-Seq technology. *BMC Genom* 23, 594. doi:10.1186/s12864-022-08827-8

Wang, L., Zhang, L., Yan, H., Liu, X., Li, N., Liang, J., et al. (2014). Genome-wide association studies identify the loci for 5 exterior traits in a Large White × Minzhu pig population. *PLoS One* 9, e103766. doi:10.1371/journal.pone.0103766

Wang, W. H., Wang, J. Y., Zhang, T., Wang, Y., Zhang, Y., and Han, K. (2019). Genome-wide association study of growth traits in Jinghai Yellow chicken hens using SLAF-seq technology. *Anim. Genet.* 50, 175–176. doi:10.1111/age.12346

Wang, W., Zhang, T., Zhang, G., Wang, J., Han, K., Wang, Y., et al. (2015). Genome-wide association study of antibody level response to NDV and IBV in Jinghai yellow chicken based on SLAF-seq technology. *J. Appl. Genet.* 56, 365–373. doi:10.1007/s13353-014-0269-y

Wimmers, K., Murani, E., Te Pas, M. F., Chang, K. C., Davoli, R., Merks, J. W., et al. (2007). Associations of functional candidate genes derived from gene-expression profiles of prenatal porcine muscle tissue with meat quality and muscle deposition. *Anim. Genet.* 38, 474–484. doi:10.1111/j.1365-2052.2007.01639.x

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, X., Jiang, A., et al. (2020). Whole-genome sequencing association analysis reveals the genetic architecture of meat quality traits in Chinese Qingyu pigs. *Genome* 63, 503–515. doi:10.1139/gen-2019-0227

Xi, Y., Xu, Q., Huang, Q., Ma, S., Wang, Y., Han, C., et al. (2021). Genome-wide association analysis reveals that EDNRB2 causes a dose-dependent loss of pigmentation in ducks. *BMC Genom* 22, 381. doi:10.1186/s12864-021-07719-7

Xie, D., Dai, Z., Yang, Z., Sun, J., Zhao, D., Yang, X., et al. (2017). Genome-wide association study identifying candidate genes influencing important agronomic traits of Flax (Linum usitatissimum L.) using SLAF-seq. *Front. Plant Sci.* 8, 2232. doi:10.3389/fpls.2017.02232

Xie, D., Dai, Z., Yang, Z., Tang, Q., Sun, J., Yang, X., et al. (2018). Genomic variations and association study of agronomic traits in flax. *BMC Genom* 19, 512. doi:10.1186/s12864-018-4899-z

Xiong, X., Liu, X., Zhou, L., Yang, J., Yang, B., Ma, H., et al. (2015). Genome-wide association analysis reveals genetic loci and candidate genes for meat quality traits in Chinese Laiwu pigs. *Mamm. Genome* 26, 181–190. doi:10.1007/s00335-015-9558-y

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, X., Deng, F., Wu, Z., Chen, S. Y., Shi, Y., Jia, X., et al. (2020). A genome-wide association study identifying genetic variants associated with growth, carcass and meat quality traits in rabbits. *Anim. (Basel)* 10, 1068. doi:10.3390/ani10061068

Yu, D. B., He, Z. L., Zhang, W. F., Jia, X. X., Qiu, X. S., Wang, L. Y., et al. (2008). The genetic effects of IGF2 gene intron3 variance in pigs. *Yi Chuan* 30, 87–93. doi:10.3724/sp.j.1005.2008.00087

Zhang, L. C., Li, N., Liu, X., Liang, J., Yan, H., Zhao, K. B., et al. (2014). A genome-wide association study of limb bone length using a Large White × Minzhu intercross population. *Genet. Sel. Evol.* 46, 56. doi:10.1186/s12711-014-0056-6

Zhao, X., Wang, C., Wang, Y., Lin, H., Wang, H., Hu, H., et al. (2019). Comparative gene expression profiling of muscle reveals potential candidate genes affecting drip loss in pork. *BMC Genet.* 20, 89. doi:10.1186/s12863-019-0794-0

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi:10.1038/ng.2310

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fgene.2023.1141411

OPEN ACCESS

EDITED BY
Li Ma,
University of Maryland, College Park,
United States

REVIEWED BY
Wentao Cai,
Institute of Animal Sciences, Chinese
Academy of Agricultural Sciences, China
Liming Hou,
Nanjing Agricultural University, China

*CORRESPONDENCE
Lusheng Huang,
✉ lushenghuang@htomail.com
Junwu Ma,
✉ ma_junwu@hotmail.com
✉ mjwjxlab@jxau.edu.cn

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

RECEIVED 10 January 2023
ACCEPTED 16 February 2023
PUBLISHED 17 March 2023

CITATION
Xie X, Huang C, Huang Y, Zou X, Zhou R, Ai H, Huang L and Ma J (2023), Genetic architecture for skeletal muscle glycolytic potential in Chinese Erhualian pigs revealed by a genome-wide association study using 1.4M SNP array.
Front. Genet. 14:1141411.
doi: 10.3389/fgene.2023.1141411

Genetic architecture for skeletal muscle glycolytic potential in Chinese Erhualian pigs revealed by a genome-wide association study using 1.4M SNP array

Xinke Xie[1], Cong Huang[1], Yizhong Huang[1], Xiaoxiao Zou[1], Runxin Zhou[1], Huashui Ai[1], Lusheng Huang[1]* and Junwu Ma[1]*

[1]National Key Laboratory for Swine Genetic Improvement and Production Technology, Ministry of Science and Technology of China, Jiangxi Agricultural University, Nanchang, China

Introduction: Muscle glycolytic potential (GP) is a key factor affecting multiple meat quality traits. It is calculated based on the contents of residual glycogen and glucose (RG), glucose-6-phosphate (G6P), and lactate (LAT) contents in muscle. However, the genetic mechanism of glycolytic metabolism in skeletal muscle of pigs remains poorly understood. With a history of more than 400 years and some unique characteristics, the Erhualian pig is called the “giant panda” (very precious) in the world’s pig species by Chinese animal husbandry.

Methods: Here, we performed a genome-wide association study (GWAS) using 1.4M single nucleotide polymorphisms (SNPs) chips for longissimus RG, G6P, LAT, and GP levels in 301 purebred Erhualian pigs.

Results: We found that the average GP value of Erhualian was unusually low (68.09 μmol/g), but the variation was large (10.4–112.7 μmol/g). The SNP-based heritability estimates for the four traits ranged from 0.16–0.32. In total, our GWAS revealed 31 quantitative trait loci (QTLs), including eight for RG, nine for G6P, nine for LAT, five for GP. Of these loci, eight were genome-wide significant ($p < 3.8 \times 10^{-7}$), and six loci were common to two or three traits. Multiple promising candidate genes such as *FTO*, *MINPP1*, *RIPOR2*, *SCL8A3*, *LIFR* and *SRGAP1* were identified. The genotype combinations of the five GP-associated SNPs also showed significant effect on other meat quality traits.

Discussion: These results not only provide insights into the genetic architecture of GP related traits in Erhualian, but also are useful for pig breeding programs involving this breed.

KEYWORDS

Erhualian, pig, glycolysis potential, GWAS, candidate gene

Introduction

Meat quality significantly influences consumers’ preference and purchasing behavior, thus improving meat quality has been being a major goal of pig breeding (Khanal et al., 2019; Zhang et al., 2019). In recent decades, the meat quality of commercial lean pigs has been improved to a certain extent by removing some known unfavorable alleles for meat quality such as PRKAG3 (200Q) (Milan et al., 2000) and RYR1 (615C) alleles (Fujii et al., 1991); however, their meat quality is still cannot meet the demands of discerning consumers (Bonneau and Lebret, 2010). A prominent problem of meat quality of commercial pigs is the high incidence (10%–30%) of inferior meat, including pale, soft, exudative (PSE) meat and dark, firm, dry (DFD) meat, which causes huge economic losses to pig industry and pork processing industry (Lewis et al., 1987; Van der Wal et al., 1988; Shen et al., 2006; Gajana et al., 2013; Gonzalez-Rivas et al., 2020). In contrast, PSE and DFD meat are rarely found in Chinese native pigs, making them valuable genetic resources for cultivating high-quality pig synthetic lines (Chen et al., 2013; Ma et al., 2013; Liu et al., 2015; Wu et al., 2020).

The degree of post-mortem pH decline in porcine skeletal muscle is an important factor accounting for the occurrence of inferior meat, as it significantly affects meat color development, water-holding capacity and tenderness (Bee et al., 2007; Scheffler and Gerrard, 2007; Popp et al., 2015). The ultimate pH value of PSE and DFD meat is usually lower than 5.5 and greater than 6.0, respectively. It is well known that muscle pH value and other pork quality attributes depend on the content of glycogen of muscle at slaughter, as well as the rate and extent of lactate accumulation during 24 h post-slaughter glycolysis metabolism (Sayre et al., 1963; Hamilton et al., 2003). Based on that, muscle glycolytic potential (GP), defined as the potential of lactate production during post-mortem glycolysis, has been proposed as a predictor for the extent of pH decline (Maribo et al., 1999; Scheffler and Gerrard, 2007).

To date, two major genes have been identified to strongly influence GP phenotype and other meat quality traits in commercial pigs. *PRKAG3* is the first causal gene proved to affect GP level, and its R200Q mutation significantly increases muscle glycogen content by up to ~70%, consequently causing acidic meat in Hampshire pigs (Milan et al., 2000). In addition, it was found in Duroc and its hybrid pigs that a splice mutation (g.8283C > A) in the *PHKG1* gene caused a 43% increase in GP and a 20% decrease in drip loss of pork (Ma et al., 2014). These two gene mutations with large effects on GP-related traits are all derived from European commercial pig breeds but are rarely present in Chinese native pigs. However, few studies have systematically investigated the genetic architecture of GP-related traits in Chinese native pigs.

The Chinese Erhualian pig is famous for its high prolificacy (Li et al., 2017). Its meat quality is also good and superior to European commercial pig breeds in color, pH value, water-holding capacity, marbling, intramuscular fat, tenderness, muscle fiber diameter, taste and flavor (Chen et al., 2013). Intriguingly, our recent study demonstrated that compared with Chinese Bama Xiang and Laiwu pigs, Erhualian pigs had a higher muscle ultimate pH value (pH_u), a smaller extent of pH decline (pH_d) from post-mortem 45 min to 24 h, and a greater variation in the two pH index (Huang et al., 2020). The variation in muscle pH value of Erhualian may be at least partially due to the variation in GP level or the concentrations of its components, including residual glycogen and glucose (RG), glucose-6-phosphate (G6P) and lactate (LAT). To verify this speculation, we herein first assessed the variation of GP-related traits in 301 Erhualian pigs, and then evaluated the impact of GP on pH values. Furthermore, to reveal the genetic architecture of the GP phenotypes of Erhualian pigs, we performed a genome-wide association study (GWAS) using high-density 1.4M single nucleotide polymorphisms (SNPs) chip data from the population.

Materials and methods

Animals and sampling

The Erhualian population was established as described previously (Liu et al., 2015). Briefly, the population consists of 160 barrows (castrated males) and 141 gilts, which were produced from 11 boars and 51 dams. These piglets were all born in Jiaoxi and transferred to Nanchang when they were 2–3 months old. Then they were raised under uniform conditions and slaughtered in 11 batches in the same commercial abattoir when they were about 300 days old. At post-mortem 30 min, 2–3 g muscle samples were collected from the *longissimus thoracis* (LT) and then stored at −80°C until assay.

Phenotype measurement and correlation analysis

The contents of residual glycogen and glucose (RG), glucose-6-phosphate (G6P) and lactate (LAT) in LT muscle were determined using the glycogen assay kits (E2GN-100) BioAssay Systems, the glucose-6-phosphate assay kits (EG6P-100) BioAssay Systems and the lactic acid assay kits (A019-2) from the Nanjing Jiancheng Bioengineering Institute, respectively. GP were calculated as the sum of: 2 × (RG + G6P) + Lactate (Monin and Sellier, 1985) and expressed as μmol of lactic acid equivalent per g of fresh muscle. In addition, pH values of LT muscle was measured twice on each sample at 45 min (pH_i) and 24 h (pH_u) after slaughter using a Delta 320 pH meter, and their mean values were calculated separately. The difference between pH_i and pH_u was taken as pH decline (pH_d). Drip loss was assayed using an EZ-Drip Loss method. Three color parameters L*, a* and b* on the surface cuts of LT were objectively evaluated with a CM-2600d/2500d Minolta Chroma meter (Liu et al., 2015). The correlations between the GP related traits and the pH traits were evaluated using Pearson correlation analysis.

Genotyping and quality control

Genomic DNA was extracted from pig ear tissue using standard phenol/chloroform extraction. In order to meet the genotyping requirements, all DNA samples were standardized to a final concentration of 50 ng/μL and quality controlled. Totally, 301 Erhualian pigs were genotyped using the 1.4M SNP Beadchips, which is a customized Affymetrix Axiom chip. The SNPs on the chip were found from whole genome sequence data of 150 Chinese indigenous pigs and 38 International commercial pigs, and were evenly distributed across the pig genome (Gong et al., 2019). PLINK v1.9 was used for quality control of the SNP chip data (Purcell et al., 2007). SNPs with a call rate ≥ 0.90 or a minor allele frequency (MAF) ≥ 0.05, and samples with a call rate greater than

TABLE 1 Descriptive statistics of four muscle GP-related traits in 301 Erhualian pigs.

Trait	Mean ± SD[a] (μmol/g)	Max. (μmol/g)	Min. (μmol/g)	CV[b]	h^2 (se)[c]
RG	6.88 ± 3.78	16.99	0.16	0.55	0.20 (0.11)
G6P	5.46 ± 2.59	15.44	0.32	0.47	0.32 (0.13)
LAT	43.41 ± 13.15	76.40	6.27	0.30	0.16 (0.09)
GP	68.09 ± 21.30	112.71	10.42	0.31	0.24 (0.10)

[a] Standard deviation.
[b] Coefficient of variation.
[c] Heritability estimates (standard errors).

95%, were retained. Consequently, 732,609 SNPs and all individuals were kept for GWAS analysis.

Genome-wide association analyses for GP related traits

We used linear regression models to adjust GP related traits for sex and slaughter batch, and then applied Genome-wide Efficient Mixed-Model Association (GEMMA) method for genetic association analyses and SNP-based heritability estimation, which were conducted using the "–lmm 1" and "–gk 1" commands of GEMMA software (Zhou and Stephens, 2012). GEMMA examined the associations of SNPs with phenotypic values under the following linear mixed model:

$$\mathbf{Y} = \mathbf{Sa} + \mathbf{u} + \mathbf{e}$$

where **Y** is a vector of residual phenotypic values that were corrected for fixed effects (sex and slaughter batch) using lm function in R program. **S** is the incidence vector for **a**, and **a** is the additive genetic effect of the SNP under test; **u** is a vector of random polygenic effects that is assumed to follow a multivariate normal distribution MVN(0, $\mathbf{G}\,\sigma_a^2$), where **G** is genomic relationship matrix that was constructed based on qualified SNPs and σ_a^2 is the polygenetic additive variance; **e** is a vector of residual errors with a distribution of N (0, $\mathbf{I}\,\sigma_e^2$), where **I** is identity matrix and σ_e^2 is residual variance.

Bonferroni correction method was used to set the genome-wide significant (0.05/N) and suggestive (1/N) association thresholds, where N is the number of independent association tests or SNPs (Lander and Kruglyak, 1995; Yang et al., 2005). Considering that SNP clusters in high linkage disequilibrium (LD) may cause overestimate of N and significance thresholds, we first pruned the full SNP set (732,609 SNPs that passed quality control) to 130,404 independent SNPs ($r^2 < 0.3$) by the command "indep-pairwise 50 5 0.3" in PLINK v1.9. Therefore, a SNP was considered to be genome-wide significance at $p < 3.8 \times 10^{-7}$ (0.05/130,404), and to be suggestive significance at $p < 7.6 \times 10^{-6}$ (1/130,404). The impact of population stratification was estimated by the quantile-quantile (Q-Q) plot and genotype data PCA analysis (Pearson and Manolio, 2008). The phenotypic variances explained by the significant SNPs were estimated by (V_{reduce} - V_{full})/V_{reduce}, where V_{full} and V_{reduce} are residual variances of models for association analysis with and without the SNP term, respectively. Haplotype block or LD analysis was performed for the chromosomal regions with multiple significant SNPs clustered around the peak SNP. Haplotype blocks were identified using the HAPLOVIEW v4.2 software with default settings (Barrett et al., 2005).

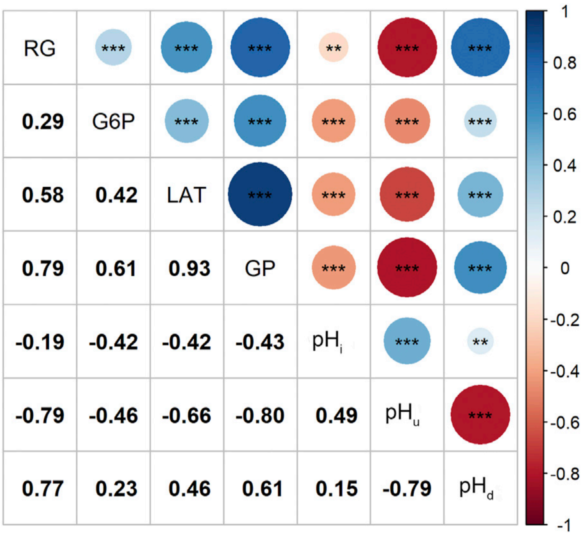


FIGURE 1
Magnitudes of correlations between the four GP related traits and the 3 pH related traits, **: $p < 0.01$, ***: $p < 0.001$.

Results

Descriptive and correlation statistics of GP related traits

The descriptive statistical results of four GP-related traits (RG, G6P, LAT and GP) in the longissimus muscle of 301 Erhualian pigs were given in Table 1. Our data showed that the average value of GP in Erhualian was 68.09 μmol/g (Table 1). The coefficient of variation of GP related traits was relatively large (0.30–0.55). Moreover, these traits had low to moderate heritability (0.16–0.32).

Correlation analysis results showed that the contents of RG, G6P and LAT were not only positively correlated with each other, but also had a strong and positive correlation with the GP content (r ≥ 0.61, $p < 0.001$; Figure 1). Since changes in muscle glucose metabolism may lead to

TABLE 2 The summary of top SNPs significantly associated with muscle GP related traits.

Trait	Peak SNP[1]	Pleiotropic[2]	Chr[3]	Pos (bp)[4]	Freq[5]	Beta (s.e)[6]	*p*-value[7]	Nearest gene	Location	Var(%)[8]
RG	rs338058884		6	31,271,210	0.102	−2.02 (0.42)	2.15E-06	*FTO*	intron	7.8
RG	rs318410870	2	7	19,677,069	0.246	−1.43 (0.29)	1.25E-06	*RIPOR2*	intron	5.3
RG	rs80841859		7	44,653,377	0.463	−1.24 (0.27)	7.49E-06	*TFAP2D*	intron	9.2
RG	rs321246758	3	7	93,785,804	0.423	−1.45 (0.27)	9.86E-08	*SLC8A3*	intron	8.7
RG	rs331183308	6	14	99,636,228	0.322	1.48 (0.29)	6.67E-07	*MINPP1*	intergenic	3.6
RG	15_135271601		15	135,271,601	0.417	1.43 (0.28)	5.68E-07	*AGAP1*	intergenic	5.2
RG	rs318442172	5	16	23,986,799	0.331	1.34 (0.27)	7.58E-07	*LIFR*	intergenic	6.7
RG	rs322341359		17	29,911,510	0.083	2.4 (0.48)	8.02E-07	*FOXA2*	intergenic	7.1
G6P	rs323123457		1	25,796,643	0.355	0.93 (0.19)	9.15E-07	*NHSL1*	intron	13.1
G6P	rs319985393		7	48,002,599	0.457	0.92 (0.18)	8.63E-07	*ADAMTS7*	intergenic	9
G6P	rs326947848		7	85,861,162	0.5	1.04 (0.22)	5.56E-06	*RGMA*	intergenic	6.3
G6P	rs320783325		8	32,614,482	0.257	1.14 (0.21)	9.81E-08	*LIMCH1*	intron	10.2
G6P	rs330779127		9	11,154,160	0.434	−0.93 (0.18)	3.43E-07	*ACER3*	intron	12.4
G6P	rs325508440		12	55,011,608	0.231	−1.08 (0.23)	5.63E-06	*MYH13*	intergenic	2.4
G6P	rs333704759		13	7,126,930	0.47	−0.96 (0.18)	1.56E-07	*SGO1*	intergenic	11
G6P	rs340666100		14	23,234,922	0.338	1.08 (0.23)	3.63E-06	*GALNT9*	Downstream gene	8.6
G6P	rs337801210	6	14	99,543,552	0.114	1.49 (0.32)	3.70E-06	*MINPP1*	intergenic	8.6
LAT	rs697205060	1	5	28,321,339	0.127	6.78 (1.38)	1.49E-06	*SRGAP1*	intron	8.8
LAT	rs80827576		5	31,833,661	0.15	6.2 (1.33)	4.44E-06	*CAND1*	intergenic	8.5
LAT	rs318410870	2	7	19,677,069	0.246	−6.12 (1.09)	4.86E-08	*RIPOR2*	intron	8.6
LAT	rs332736034		7	109,115,167	0.279	−5.84 (1.17)	9.16E-07	*ENSSSCG00000042684*	intergenic	9
LAT	rs327466581		9	133,730,734	0.111	7.59 (1.57)	2.03E-06	*ENSSSCG00000015620*	intergenic	6.6
LAT	rs345209200		11	67,518,507	0.491	−4.84 (1.02)	3.14E-06	*STK24*	intron	9.7
LAT	rs345106152	4	12	28,646,375	0.434	−4.83 (0.92)	3.27E-07	*ENSSSCG00000043336*	upstream	10
LAT	rs318442172	5	16	23,986,799	0.331	5.06 (0.99)	5.64E-07	*LIFR*	intergenic	8.6
LAT	rs330527025		18	34,087,184	0.459	4.61 (0.97)	2.83E-06	*IMMP2L*	intron	5.3
GP	rs697205060	1	5	28,321,339	0.127	11.23 (2.14)	3.06E-07	*SRGAP1*	intron	5.6
GP	rs318410870	2	7	19,677,069	0.246	−9.40 (1.81)	3.77E-07	*RIPOR2*	intron	8.8
GP	rs332409349	3	7	93,944,227	0.326	−8.95 (1.82)	1.40E-06	*SYNJ2BP*, COX16	intron	5.6
GP	rs345106152	4	12	28,646,375	0.434	−7.58 (1.53)	1.14E-06	*ENSSSCG00000043336*	upstream gene	12.2
GP	rs318442172	5	16	23,986,799	0.331	7.93 (1.66)	2.85E-06	*LIFR*	intergenic	9.7

[1] One SNP, wthout feature ID (rs) in NCBI, was named according to their physical positions on the Sscrofa11.1 assembly.

[2] The numbes indicated pleiotropic loci associated with more than one trait.

[4] 3The locations of the associated SNPs, on the *Sus Scrofa* Build 11.1 assembly.

[5] Minor allele frequeny.

[6] Beta estimates (stanard errors for beta).

[7] The *p* values that ar lower than the genome-wide significance threshold (3.8×10^{-7}) are underlined.

[8] Phenotypic variance hat the peak SNP, explain.

changes in its pH values, we estimated the correlations between the four GP-related traits and pH index. As expected, our data indicates that higher RG and GP contents tended to cause a greater post-mortem pH decline (pH_d) and a lower ultimate pH (pH_u) value ($r \le -0.79$, $p < 0.001$; Figure 1), which is consistent with other studies (Hamilton et al., 2003; Luo et al., 2017).

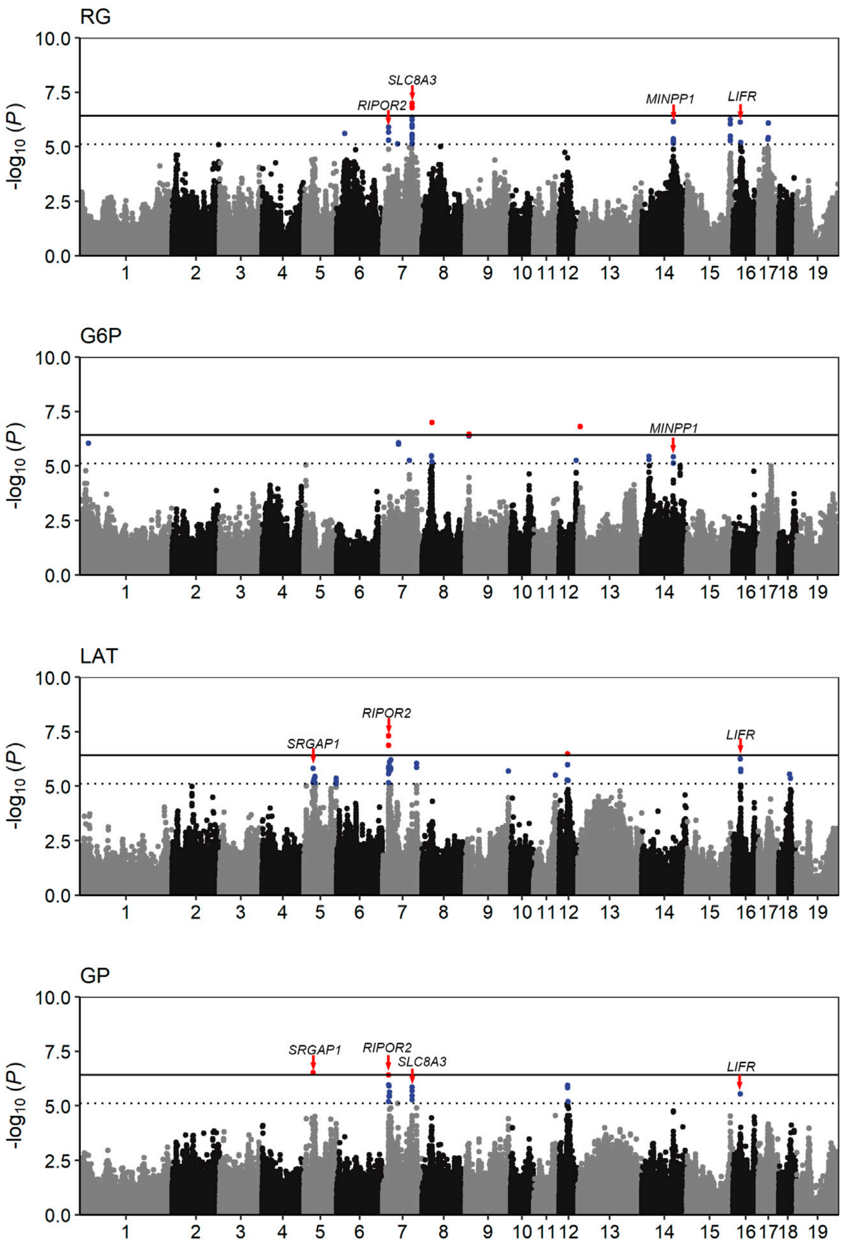


FIGURE 2
Manhattan plots of the GWAS result for the four GP related traits. The solid and dotted lines represent the genome-wide significance threshold and the suggestive significance threshold, respectively. The candidate genes close to five pleiotropic loci were identified.

Summary of GWAS results

The population stratification was evaluated using Q-Q plots and PCA plots (Supplementary Figure S1, S2). The inflation factors (λ) for four traits were between 1.018 and 1.119 and PCA analysis showed that the population was relatively clustered, which suggests that our population had no obvious stratification. Through GWAS analysis, we totally identified 31 quantitative trait loci (QTLs) for the four GP related traits (Table 2; Figure 2), including eight for RG, nine for G6P, nine for LAT, five for GP. Notably, six out of the 31 QTLs showed significant association with more than one trait, including the QTLs with lead SNPs rs331183308 for RG and G6P, rs321246758 for RG and GP, rs318410870 and rs318442172 for RG, LAT and GP, and rs697205060 and rs345106152 for LAT and GP. Perhaps because LAT content accounted for the major fraction of GP at 45 min post-mortem (43.41/68.09 ≈ 63.8%; Table 1), they shared four common GWAS peak SNPs (Table 2).

QTLs for RG trait

The RG level in muscle at post-mortem 45 min is weakly associated with pH_i (r = −0.19, p < 0.01) but strongly associated with pH_d (r = 0.77, p < 0.01; Figure 1), suggesting that it plays an important role in promoting the subsequent decline of pH in meat. In the Erhualian pigs, a total of eight QTLs associated with RG were identified, which individually explained 3.6%–9.2% of the phenotypic variance. Among them, only the QTL with peak SNP rs321246758 ($p = 9.86 \times 10^{-8}$), which is located in the intron of *SLC8A3* gene on *Sus scrofa* chromosome 7 (SSC7), exceeded the genome-wide significance threthold (Table 2). This QTL region for RG overlapped with the QTL region for GP, in which the peak SNPs rs321246758 for RG and the peak SNP rs332409349 for GP was in moderate LD ($r^2 = 0.63$) (Supplementary Figure S3A), and a common SNP rs322439157 significantly associated with both traits was found, suggesting the existence of a pleiotropic locus in the *SLC8A3–SYNJ2BP* gene region (93.78–93.95 Mb) on SSC7 (Supplementary Figure S3C). The minor allele of rs322439157 were correlated with decreased levels of both RG and GP (Supplementary Figure S3B). In addition to rs321246758, three other GWAS signals appeared in the genes' introns, including rs338058884, rs80841859, and rs318410870 located in the introns of *FTO*, *TFAP2D* and *RIPOR2*, respectively.

QTLs for G6P

The conversion of glucose to G6P is the first rate-limiting step in the glycolysis pathway. We detected nine QTLs significantly associated with G6P, including three genome-wide significant QTLs ($p < 3.8 \times 10^{-7}$) with the peak SNPs rs320783325, rs330779127 and rs333704759 on SSC8, 9 and 13, respectively. Among them, the rs320783325, an intron variant in *LIMCH1* gene, had the most significant association with G6P ($p = 9.81 \times 10^{-8}$). In addition, we found that on SSC14, the GWAS peak SNP rs337801210 for G6P was located 92.7 kb proximal to the GWAS peak SNP rs331183308 for RG (Table 2). Although a weak LD ($r^2 = 0.21$) were observed between rs337801210 and rs331183308, they and nine other significant SNPs (2 for G6P and seven for RG) linked to them were concentrated in the 99.52–99.64 Mb region, which only contains the *MINPP1* gene. Moreover, the minor alleles of rs337801210 and rs331183308 were both associated with increased phenotypic values (Table 2). Thus, there may be a pleiotropic QTL in the *MINPP1* region that affects both G6P and RG.

QTLs for LAT

Under anaerobic conditions, the end product of glycolysis in muscle cells is lactic acid. Our GWAS also revealed nine QTLs significantly associated with LAT (Table 2; Figure 2). Of them, two reached genome-wide significance: one on SSC7 with peak SNP rs318410870 ($p = 4.86 \times 10^{-8}$), and another on SSC12 with the peak SNP rs345106152 ($p = 3.27 \times 10^{-7}$). Moreover, these two peak SNPs were also significantly with GP. Notably, the pleiotropic locus rs345106152 explained the largest portion of phenotypic variance for LAT (10%) and GP (12.2%) (Table 2).

QTLs for GP

In this study, two genome-wide significant QTLs and three suggestive QTLs were found to affect the GP trait. These five QTLs for GP were also significantly associated with at least one of the three GP components (RG, G6P and LAT). The GWAS peak SNPs rs697205060 ($p = 3.06 \times 10^{-7}$) and rs318410870 ($p = 3.77 \times 10^{-7}$) represented the genome-wide significant QTLs on SSC5 and SSC7, respectively. For rs697205060, an intron variant of *SRGAP1*, its minor allele was associated with increased concentrations of GP and LAT, while the minor allele at rs318410870, an intron variant of the *RIPOR2* gene, was associated with reduced concentrations of GP, RG and LAT (Table 2; Figures 3A–C). Furthermore, the LD analysis showed that the peak SNP rs318410870 and other three SNPs (including the second significant SNP) were all located in a 3-kb LD block in the *RIPOR2* gene (Figure 3D). In addition, on SSC16, the SNP rs318442172 near the *LIFR* gene was also detected to be significantly associated with GP, RG and LAT.

Effects of multiple pleiotropic loci combination on meat quality traits

Next, we asked what would happen to other meat quality traits if we selected the five loci that affect GP and other components (including rs697205060, rs332409349, rs318410870, rs345106152, and rs318442172) in breeding. To this end, we first identified the genome combination types of these five SNPs in the Erhualian population. 73 genotype combinations were found in this population (Supplementary Table S1). According to the number of alleles (called $allele_2$) that increase GP values at the five peak SNPs in each individual's genotype combination, we divided all individuals into three groups: group A with 0–2 $alleles_2$, group B with three to seven $alleles_2$, and group C with 8–10 $alleles_2$. Analysis of variance showed that there were significant differences in GP, pH_i, pH_u, pH_d, drip loss, redness (a*) and yellowness (b*) of meat between the three groups, but no significant differences in lightness (L*). Moreover, the GP, pH_d, drip loss, a* and b* values tended to increase in the genotype combination groups with more $allele_2$, while their corresponding pH_i and pH_u showed a significant downward trend (Figure 4).

Discussion

Many studies have investigated the variation of muscle GP level and its correlation with meat quality in Western commercial pig purebreds, crossbreds and Western Chinese hybrid pigs (van Laack and Kauffman, 1999; Bee et al., 2006; Duan et al., 2009; Conte et al., 2021), but such studies have been rarely conducted in Chinese indigenous pig breeds. In this study, the post-mortem muscle glycolytic phenotype of Erhualian pigs was characterized for the first time. We found that the average GP value of Erhualian was 68.09 μmol/g, which is only nearly half of the values reported by D'Astous-page et al. (D'Astous-Pagé et al., 2017) for Duroc (161.06 μmol/g), Landrace (164.27 μmol/g) and Yorkshire (159.27 μmol/g) and the value reported by Duan et al. (Duan et al., 2009) for a White Duroc × Erhualian F2 pigs (136.66 μmol/g). In addition, we showed that the RG and GP levels were highly negatively ($r \leq -0.79$, $p < 0.001$) correlated with pH_u, which is consistent with the notion that higher glycogen or GP at slaughter leads to greater pH_u

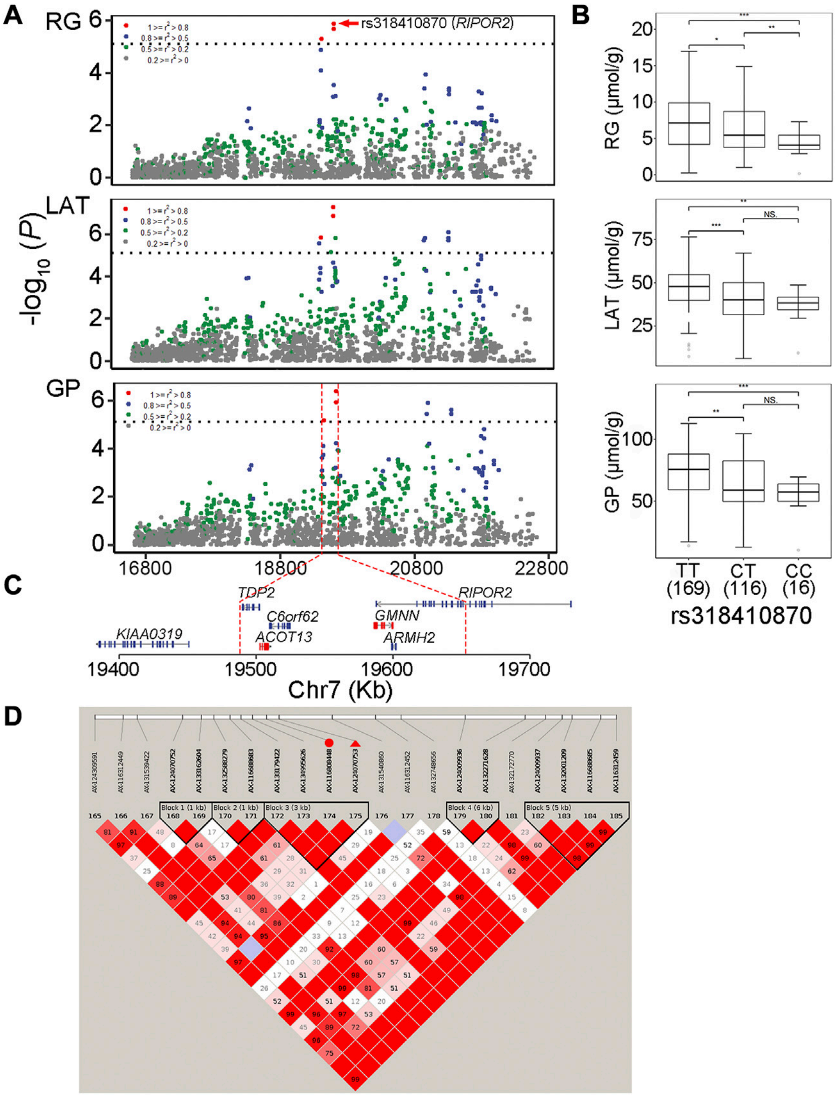


FIGURE 3
The GWAS peak SNP rs318410870 in RIPOR2 gene with pleiotropic effect on residual glycogen and glucose (RG), lactate (LAT) and glycolytic potential (GP) levels in muscle **(A)** Regional association plots of the QTL region centered by rs318410870 (in red dots). Different colors stand for degrees of linkage disequilibrium (r2) between corresponding SNPs and the peak SNP **(B)** Boxplots showing the difference in phenotypic values between the three genotypes of the SNP rs318410870 **(C)** The physical locations of genes surrounding the significant SNPs **(D)** Haplotype blocks on the QTL region containing rs318410870 indicated with red triangle and its one neighboring significant SNP indicated with red dot associated with the three traits.

decline (Scheffler and Gerrard, 2007). Given the significant effect of GP on pH decline in muscle, it can be inferred that the lower glycogen and glucose contents in the muscle of Erhualian before slaughter may be the main reason why its meat pH$_u$ value was higher than that of Western commercial pig breeds (6.01 vs. ~5.5) (Huang et al., 2020). In addition, the four GP-related traits of the Erhualian pigs showed considerable variation (CV = 30%–55%; Table 1), which may largely explain the wide variation of some meat quality traits related to GP in this population, such as the CV (63.4%) of pH$_d$ value (Huang et al., 2020).

Numerous QTLs for meat quality traits have been identified in various F2 pig crosses between Chinese and Western outbred lines or in Western commercial hybrid pigs, but few QTLs have been detected in founder lines, which to some extent hinder the development of pure-breeding and cross-breeding. To our best knowledge, this is the first study to explore genetic loci for GP-related traits in a Chinese local pig breed (Erhualian). Unlike most pig GWAS using 50 K or 60 K SNP chips, this study used 1.4 M SNP chips, which improved the statistical power and QTL mapping resolution. In fact, we not only detected 31 QTLs (8 loci per trait on average), but also found QTLs in some regions with small LD.

By comparing with the QTLs for GP deposited in pig QTL database (pigQTLdb), we found that all QTLs detected in this study, except for two, had not been previously reported. On SSC7, the GWAS peak locus (rs332736034) we identified for LAT at 109.12 Mb is close to the QTL for the same trait detected

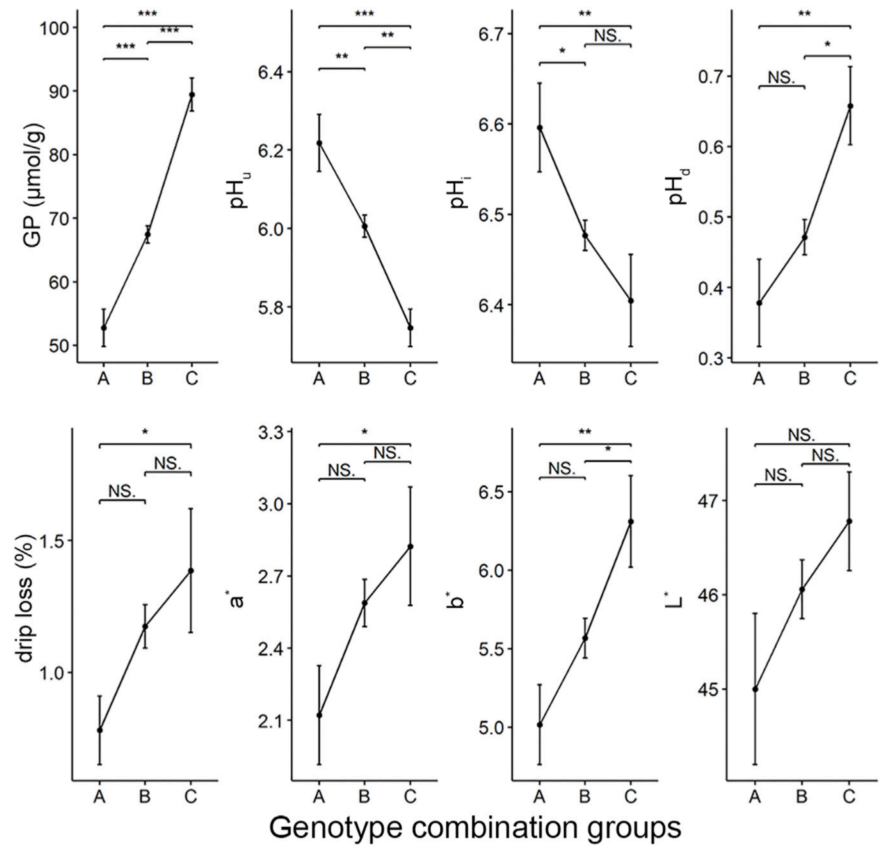


FIGURE 4
The differences in meat quality traits between three groups of genotype combinations of five GWAS peak SNPs for GP. The color parameters L*, a*, b* indicate the lightness, redness and yellowness of the meat, respectively. There were 31, 234 and 30 individuals in group A, B and C respectively.

in a Meishan × Pietrain F2 family (Reiner et al., 2002). Similarly, on SSC15, the significant region associated with RG of Erhualian overlapped with the QTL interval found in a Berkshire × Yorkshire F2 intercross (Ciobanu et al., 2001). Thus, the alleles at the two loci may be segregating rather than fixed in Erhualian pigs.

In the regions around the GWAS peaks, we screened a number of candidate genes, providing new clues for understanding the genetic mechanism of glycolytic phenotypes. For the identified pleiotropic loci, five strong candidate genes were proposed based on their functional annotations. The GWAS signal (rs331183308) for RG and G6P found on SSC14 was adjacent to the *MINPP1* gene. The protein encoded by *MINPP1* is multiple inositol polyphosphate phosphatase, which can remove 3-phosphate from inositol phosphate substrates and convert 2,3 bisphosphoglycerate (2,3-BPG) to 2-phosphoglycerate that is part of the glycolytic pathway (Cho et al., 2008). The related pathways of *MINPP1* include Rapoport-Luebering glycolytic shunt and inositol phosphate metabolism (Cho et al., 2008). Thus, *MINPP1* could be regarded as strong candidate genes for RG and G6P. In addition, the peak locus rs321246758 in the intron of *SLC8A3* on SSC7 accounted 8.7% of the phenotypic variance in RG. The *SLC8A3* gene encodes a member of the sodium/calcium exchanger integral membrane protein family. Calcium plays a crucial role in muscle contraction and energy metabolism (Berridge et al., 2003), and SLC8A3 contributes to Ca^{2+} transport during excitation-contraction coupling in muscle. Its loss leads to muscle necrosis and abnormal Ca^{2+} homeostasis (Michel et al., 2014). Therefore, *SLC8A3* is a good candidate for RG.

The two GWAS peaks rs318410870 on SSC7 and rs318442172 on SSC16 were common to RG, LAT and GP. The SNP rs318410870 is an intronic variant of the *RIPOR2* gene. RIPOR2 can inhibit the activity of small GTPase RhoA and regulate myoblast differentiation (Yoon et al., 2007). RhoA plays an important role in actin cytoskeleton reorganization through RhoA-ROCKS signaling pathway (Ridley and Hall, 1992). Khue Ha Minh Duong et al. indicated that RhoA mediates glucose transport by regulating the vesicle trafficking machinery in an Akt-independent manner (Duong and Chun, 2019), and David Wu et al. also showed that activation of RhoA induces celluar glycolysis through translocation of glucose transporter GLUT3, which provides energy for cell contraction (Wu et al., 2021). Therefore, RIPOR2 may influence the process of glucose metabolism by inhibiting RhoA. Another GWAS locus affecting three traits, rs318442172, is an intronic variant in *LIFR* gene. LIFR protein, a member of the type I cytokine receptor family, binds to a converter subunit gp130 to form a receptor complex which mediates the role of the leukemia inhibitory factor (LIF) in cellular differentiation and proliferation (Hunt and White, 2016). Jessica C Hogan et al. showed that LIF regulates the expression of genes involved in lipid synthesis and has an impact on insulin-stimulated glucose uptake (Hogan and

Stephens, 2005). In addition, Suhu Liu et al. found that LIFR can regulates the expression level of glucose transporter GLUT1, and the reduced expression level of LIFR leads to an increase in GLUT1 expression level, glycolysis and mitochondrial respiration (Liu et al., 2021). Thus, *LIFR* and *RIPOR2* could be regarded as candidate genes for the GP-related traits.

Two QTLs with the peak SNPs rs697205060 and rs345106152 were detected to affect both GP and LAT. The SNP rs697205060 is located in the intron of the *SRGAP1* gene, which encodes a GTPase activator for RhoA and Cdc42. As mentioned above, RhoA plays an important role in glucose transport (Ridley and Hall, 1992; Duong and Chun, 2019; Wu et al., 2021). Therefore, *SRGAP1* may influence the process of glucose metabolism by activating RhoA, and could be regarded as a promising candidate gene for the GP-related traits.

In addition to the pleiotropic loci, several possible candidate genes for the QTL associated with a single trait were also recognized. For example, the *FTO* gene containing the significant SNP rs338058884 for RG is a promising candidate. Melina Claussnitzer et al. found that the SNP rs1421085 in *FTO* gene disrupts a conserved motif for the *ARID5B* repressor, which leads to derepression of a potent preadipocyte enhancer and an increase in lipid storage with the increasing expression of IRX3 and IRX5 (Claussnitzer et al., 2015). Moreover, *FTO*'s related pathways include Glucose/Energy Metabolism and many studies have shown that *FTO* is closely related to glucose metabolism (Maia-Landim et al., 2018; Yang et al., 2019; Krüger et al., 2020; De Soysa et al., 2021). So, *FTO* is likely involved in regulating muscle glycogen content. On SSC17, the coding gene closest to the peak SNP rs322341359 for RG is *FOXA2* that encodes Forkhead Box A2. As a transcription activator, FOXA2 has a role in regulation of the expression of genes responsible for glucose homeostasis. Ping Wang et al. showed that FOXA2 restrains the proliferation of liver progenitor cells by decreasing PI3K/Akt/HK2-mediated glycolysis (Wang et al., 2020). Jennifer N Dines et al. reported that aberrant glucose homeostasis occurs in an individual with a missense variant in *FOXA2* (Dines et al., 2019). Therefore, FOXA2 are likely related to the SSC17 QTL effect on RG.

The most significant GWAS signal (rs320783325 with $p = 9.81 \times 10^{-8}$) for G6P was present in the *LIMCH1* gene. This gene encodes calponin homology domains-containing protein 1 which enables myosin II head/neck binding activity. The absence of LIMCH1 can affects the formation of actin stress fibers as well as the stability of focal adhesions, which are basic structures to ensure benign contraction and expansion in skeletal muscle (Lin et al., 2017). Fiuza-luces et al. showed that there was a significant relationship between LIMCH1 expression and adaptation of skeletal muscle to endurance training, independent of muscle glycogen availability (Fiuza-Luces et al., 2018). However, it is not ruled out that LIMCH1 alters the rate of glycolysis when it participates in stress pathways regulated by exercise, which warrants further investigation.

It is very important to identify genetic loci affecting GP phenotypes for genetic improvement of meat quality traits, as demonstrated by the application of mutations in *PRKAG3* and *PHKG1*, two known major genes in pig breeding (Galve et al., 2013). In this study, we identified five GP-associated peak SNPs, with minor allele frequencies of 0.127–0.434 (Table 2), indicating that there is sufficient room for selection of these loci in Erhualian pigs. More importantly, we observed that the genotype combinations of these five loci significantly affected not only GP, but also pH, color and water-holding capacity of meat (Figure 4). Therefore, the results provide a reference for improving meat quality uniformity and breeding efficiency of Erhualian pigs by multi-marker assistant selection.

Conclusion

We found that the GP content in the skeletal muscle of Erhualian pigs was low and had a wide range of phenotypic variation, which may play a role in the formation of specific meat quality traits of this breed. Further, our GWAS analysisin the Erhualian population identified 31 loci significantly associated with the GP related traits, including 29 novel loci. Among them, six loci exhibited pleiotropic effects on these GP related traits. *SCL8A3*, *MINPP1*, *SRGAP1*, *RIPOR2*, and *LIFR* near the pleiotropic QTL peaks are highlighted as novel candidate genes related to glucose metabolism in muscle, which are worthy of further study. We also demonstrate the potential application of GP-associated SNPs in improving a variety of meat quality traits. Therefore, this study not only advances our understanding of the genetic architecture of GP related traits but also provide genetic markers for improving the pork quality of Erhualian.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Datasets belong to Jiangxi Agricultural University. Requests to access these datasets should be directed to ma_junwu@hotmail.com.

Ethics statement

The animal study was reviewed and approved by the ethics committee of Jiangxi Agricultural University. Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

XX performed the experiments, analyzed the data and wrote the article. CH, YH, XZ, and RZ helped with the experimental process and data analysis. HA contributed materials. LH and JM conceived and designed the experiments. All authors reviewed and approved the article.

Funding

This study was supported by the National Natural Science Foundation of China (No 31790413 and 31872339).

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.2023.1141411/full#supplementary-material

SUPPLEMENTARY FIGURE S1
The Q-Q plots for 4 GP-related traits. The x axis represents the expected -log10 (P) and the y axis represents the observed -log10 (P)

SUPPLEMENTARY FIGURE S2
The PCA plots for the genotype data of Erhuanlian population

SUPPLEMENTARY FIGURE S3
The pleiotropic QTL for residual glycogen and glucose (RG) and glycolytic potential (GP) in muscle at about 93.78 Mb on SSC7. **(A)** Regional association plots of the QTL region of 4 Mb centered by the peak SNP (in red dots). Different colors stand for degrees of linkage disequilibrium (r2) between corresponding SNPs and the peak SNP. **(B)** Boxplots showing the difference in RG and GP values between the three genotypes of the common significant SNP rs322439157. **(C)** The physical locations of genes surrounding the significant SNPs

References

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

Bee, G., Anderson, A. L., Lonergan, S. M., and Huff-Lonergan, E. (2007). Rate and extent of pH decline affect proteolysis of cytoskeletal proteins and water-holding capacity in pork. *Meat Sci.* 76 (2), 359–365. doi:10.1016/j.meatsci.2006.12.004

Bee, G., Biolley, C., Guex, G., Herzog, W., Lonergan, S. M., and Huff-Lonergan, E. (2006). Effects of available dietary carbohydrate and preslaughter treatment on glycolytic potential, protein degradation, and quality traits of pig muscles. *J. Anim. Sci.* 84 (1), 191–203. doi:10.2527/2006.841191x

Berridge, M. J., Bootman, M. D., and Roderick, H. L. (2003). Calcium signalling: Dynamics, homeostasis and remodelling. *Nat. Rev. Mol. Cell Biol.* 4 (7), 517–529. doi:10.1038/nrm1155

Bonneau, M., and Lebret, B. (2010). Production systems and influence on eating quality of pork. *Meat Sci.* 84 (2), 293–300. doi:10.1016/j.meatsci.2009.03.013

Chen, J., Huang, R., Li, Q., and Gu, Y. (2013). Research progress of germplasm resources of Erhuface pig and experience in development and utilization. *CHINA SWINE IND.* 8 (S1), 72–75.

Cho, J., King, J. S., Qian, X., Harwood, A. J., and Shears, S. B. (2008). Dephosphorylation of 2,3-bisphosphoglycerate by MIPP expands the regulatory capacity of the Rapoport-Luebering glycolytic shunt. *Proc. Natl. Acad. Sci. U. S. A.* 105 (16), 5998–6003. doi:10.1073/pnas.0710980105

Ciobanu, D., Bastiaansen, J., Malek, M., Helm, J., Woollard, J., Plastow, G., et al. (2001). Evidence for new alleles in the protein kinase adenosine monophosphate-activated gamma(3)-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. *Genetics* 159 (3), 1151–1162. doi:10.1093/genetics/159.3.1151

Claussnitzer, M., Dankel, S. N., Kim, K-H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373 (10), 895–907. doi:10.1056/NEJMoa1502214

Conte, S., Pomar, C., Paiano, D., Duan, Y., Zhang, P., Lévesque, J., et al. (2021). The effects of feeding finishing pigs of two genders with a high fiber and high fat diet on muscle glycolytic potential at slaughter and meat quality. *Meat Sci.* 177, 108484. doi:10.1016/j.meatsci.2021.108484

D'Astous-Pagé, J., Gariépy, C., Blouin, R., Cliche, S., Méthot, S., Sullivan, B., et al. (2017). Identification of single nucleotide polymorphisms in carnosine-related genes and effects of genotypes on pork meat quality attributes. *Meat Sci.* 134, 54–60. doi:10.1016/j.meatsci.2017.07.019

De Soysa, A. K. H., Langaas, M., Jakic, A., Shojaee-Moradie, F., Umpleby, A. M., Grill, V., et al. (2021). The fat mass and obesity-associated (FTO) gene allele rs9939609 and glucose tolerance, hepatic and total insulin sensitivity, in adults with obesity. *PloS One* 16 (3), e0248247. doi:10.1371/journal.pone.0248247

Dines, J. N., Liu, Y. J., Neufeld-Kaiser, W., Sawyer, T., Ishak, G. E., Tully, H. M., et al. (2019). Expanding phenotype with severe midline brain anomalies and missense variant supports a causal role for FOXA2 in 20p11.2 deletion syndrome. *Am. J. Med. Genet. A* 179 (9), 1783–1790. doi:10.1002/ajmg.a.61281

Duan, Y. Y., Ma, J. W., Yuan, F., Huang, L. B., Yang, K. X., Xie, J. P., et al. (2009). Genome-wide identification of quantitative trait loci for pork temperature, pH decline, and glycolytic potential in a large-scale White Duroc x Chinese Erhualian resource population. *J. Anim. Sci.* 87 (1), 9–16. doi:10.2527/jas.2008-1128

Duong, K. H. M., and Chun, K-H. (2019). Regulation of glucose transport by RhoA in 3T3-L1 adipocytes and L6 myoblasts. *Biochem. Biophys. Res. Commun.* 519 (4), 880–886. doi:10.1016/j.bbrc.2019.09.083

Fiuza-Luces, C., Santos-Lozano, A., Llavero, F., Campo, R., Nogales-Gadea, G., Díez-Bermejo, J., et al. (2018). Muscle molecular adaptations to endurance exercise training are conditioned by glycogen availability: A proteomics-based analysis in the McArdle mouse model. *J. Physiol.* 596 (6), 1035–1061. doi:10.1113/JP275292

Fujii, J., Otsu, K., Zorzato, F., de Leon, S., Khanna, V. K., Weiler, J. E., et al. (1991). Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* 253 (5018), 448–451. doi:10.1126/science.1862346

Gajana, C. S., Nkukwana, T. T., Marume, U., and Muchenje, V. (2013). Effects of transportation time, distance, stocking density, temperature and lairage time on incidences of pale soft exudative (PSE) and the physico-chemical characteristics of pork. *Meat Sci.* 95 (3), 520–525. doi:10.1016/j.meatsci.2013.05.028

Galve, A., Burgos, C., Varona, L., Carrodeguas, J. A., Cánovas, Á., and López-Buesa, P. (2013). Allelic frequencies of PRKAG3 in several pig breeds and its technological consequences on a Duroc × Landrace-Large White cross. *J. Animal Breed. Genet. = Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie.* 130 (5), 382–393. doi:10.1111/jbg.12042

Gong, H., Xiao, S., Li, W., Huang, T., Huang, X., Yan, G., et al. (2019). Unravelling the genetic loci for growth and carcass traits in Chinese Bamaxiang pigs based on a 1.4 million SNP array. *J. Anim. Breed. Genet.* 136 (1), 3–14. doi:10.1111/jbg.12365

Gonzalez-Rivas, P. A., Chauhan, S. S., Ha, M., Fegan, N., Dunshea, F. R., and Warner, R. D. (2020). Effects of heat stress on animal physiology, metabolism, and meat quality: A review. *Meat Sci.* 162, 108025. doi:10.1016/j.meatsci.2019.108025

Hamilton, D. N., Miller, K. D., Ellis, M., McKeith, F. K., and Wilson, E. R. (2003). Relationships between longissimus glycolytic potential and swine growth performance, carcass traits, and pork quality. *J. Anim. Sci.* 81 (9), 2206–2212. doi:10.2527/2003.8192206x

Hogan, J. C., and Stephens, J. M. (2005). Effects of leukemia inhibitory factor on 3T3-L1 adipocytes. *J. Endocrinol.* 185 (3), 485–496. doi:10.1677/joe.1.05980

Huang, Y., Zhou, L., Zhang, J., Liu, X., Zhang, Y., Cai, L., et al. (2020). A large-scale comparison of meat quality and intramuscular fatty acid composition among three Chinese indigenous pig breeds. *Meat Sci.* 168, 108182. doi:10.1016/j.meatsci.2020.108182

Hunt, L. C., and White, J. (2016). The role of leukemia inhibitory factor receptor signaling in skeletal muscle growth, injury and disease. *Adv. Exp. Med. Biol.* 900, 45–59. doi:10.1007/978-3-319-27511-6_3

Khanal, P., Maltecca, C., Schwab, C., Gray, K., and Tiezzi, F. (2019). Genetic parameters of meat quality, carcass composition, and growth traits in commercial swine. *J. Anim. Sci.* 97 (9), 3669–3683. doi:10.1093/jas/skz247

Krüger, N., Biwer, L. A., Good, M. E., Ruddiman, C. A., Wolpe, A. G., DeLalio, L. J., et al. (2020). Loss of endothelial FTO antagonizes obesity-induced metabolic and vascular dysfunction. *Circ. Res.* 126 (2), 232–242. doi:10.1161/CIRCRESAHA.119.315531

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11 (3), 241–247. doi:10.1038/ng1195-241

Lewis, P. K., Yakes, L. Y., Noland, P. R., and Brown, C. J. (1987). The effect of DFD classification and internal cooking temperature on certain pork muscle characteristics. *Meat Sci.* 21 (2), 137–144. doi:10.1016/0309-1740(87)90026-X

Li, P-H., Ma, X., Zhang, Y-Q., Zhang, Q., and Huang, R-H. (2017). Progress in the physiological and genetic mechanisms underlying the high prolificacy of the Erhualian pig. *Yi Chuan* 39 (11), 1016–1024. doi:10.16288/j.yczz.17-119

Lin, Y-H., Zhen, Y-Y., Chien, K-Y., Lee, I. C., Lin, W-C., Chen, M-Y., et al. (2017). LIMCH1 regulates nonmuscle myosin-II activity and suppresses cell migration. *Mol. Biol. Cell* 28 (8), 1054–1065. doi:10.1091/mbc.E15-04-0218

Liu, S., Gandler, H. I., Tošić, I., Ye, D. Q., Giaccone, Z. T., and Frank, D. A. (2021). Mutant KRAS downregulates the receptor for leukemia inhibitory factor (LIF) to enhance a signature of glycolysis in pancreatic cancer and lung cancer. *Mol. Cancer Res.* 19, 1283–1295. doi:10.1158/1541-7786.MCR-20-0633

Liu, X., Xiong, X., Yang, J., Zhou, L., Yang, B., Ai, H., et al. (2015). Genome-wide association analyses for meat quality traits in Chinese Erhualian pigs and a Western Duroc × (Landrace × Yorkshire) commercial population. *Genet. Sel. Evol.* 47, 44. doi:10.1186/s12711-015-0120-x

Luo, J., Shen, Y. L., Lei, G. H., Zhu, P. K., Jiang, Z. Y., Bai, L., et al. (2017). Correlation between three glycometabolic-related hormones and muscle glycolysis, as well as meat quality, in three pig breeds. *J. Sci. Food Agric.* 97 (9), 2706–2713. doi:10.1002/jsfa.8094

Ma, J., Yang, J., Zhou, L., Ren, J., Liu, X., Zhang, H., et al. (2014). A splice mutation in the PHKG1 gene causes high glycogen content and low meat quality in pig skeletal muscle. *PLoS Genet.* 10 (10), e1004710. doi:10.1371/journal.pgen.1004710

Ma, J., Yang, J., Zhou, L., Zhang, Z., Ma, H., Xie, X., et al. (2013). Genome-wide association study of meat quality traits in a White Duroc×Erhualian F2 intercross and Chinese Sutai pigs. *PloS One* 8 (5), e64047. doi:10.1371/journal.pone.0064047

Maia-Landim, A., Ramírez, J. M., Lancho, C., Poblador, M. S., and Lancho, J. L. (2018). Long-term effects of Garcinia cambogia/Glucomannan on weight loss in people with obesity, PLIN4, FTO and Trp64Arg polymorphisms. *BMC Complement. Altern. Med.* 18 (1), 26. doi:10.1186/s12906-018-2099-7

Maribo, H., Støier, S., and Jørgensen, P. F. (1999). Procedure for determination of glycolytic potential in porcine m. longissimus dorsi. *Meat Sci.* 51 (2), 191–193. doi:10.1016/s0309-1740(98)00130-2

Michel, L. Y. M., Verkaart, S., Koopman, W. J. H., Willems, P. H. G. M., Hoenderop, J. G. J., and Bindels, R. J. M. (2014). Function and regulation of the Na^+-Ca^{2+} exchanger NCX3 splice variants in brain and skeletal muscle. *J. Biol. Chem.* 289 (16), 11293–11303. doi:10.1074/jbc.M113.529388

Milan, D., Jeon, J. T., Looft, C., Amarger, V., Robic, A., Thelander, M., et al. (2000). A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* 288 (5469), 1248–1251. doi:10.1126/science.288.5469.1248

Monin, G., and Sellier, P. (1985). Pork of low technological quality with a normal rate of muscle pH fall in the immediate post-mortem period: The case of the Hampshire breed. *Meat Sci.* 13 (1), 49–63. doi:10.1016/S0309-1740(85)80004-8

Pearson, T. A., and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA* 299 (11), 1335–1344. doi:10.1001/jama.299.11.1335

Popp, J., Wicke, M., Klein, G., and Krischek, C. (2015). The relationship of pork longissimus muscle pH to mitochondrial respiratory activities, meat quality and muscle structure. *Animal* 9 (2), 356–361. doi:10.1017/S1751731114002365

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Reiner, G., Heinricy, L., Muller, E., Geldermann, H., and Dzapo, V. (2002). Indications of associations of the porcine FOS proto-oncogene with skeletal muscle fibre traits. *Anim. Genet.* 33 (1), 49–55. doi:10.1046/j.1365-2052.2002.00805.x

Ridley, A. J., and Hall, A. (1992). The small GTP-binding protein rho regulates the assembly of focal adhesions and actin stress fibers in response to growth factors. *Cell* 70 (3), 389–399. doi:10.1016/0092-8674(92)90163-7

Sayre, R. N., Briskey, E. J., and Hoekstra, W. G. (1963). Porcine muscle glycogen structure and its association with other muscle properties. *Proc. Soc. Exp. Biol. Med.* 112, 223–225. doi:10.3181/00379727-112-27999

Scheffler, T. L., and Gerrard, D. E. (2007). Mechanisms controlling pork quality development: The biochemistry controlling postmortem energy metabolism. *Meat Sci.* 77 (1), 7–16. doi:10.1016/j.meatsci.2007.04.024

Shen, Q. W., Means, W. J., Thompson, S. A., Underwood, K. R., Zhu, M. J., McCormick, R. J., et al. (2006). Pre-slaughter transport, AMP-activated protein kinase, glycolysis, and quality of pork loin. *Meat Sci.* 74 (2), 388–395. doi:10.1016/j.meatsci.2006.04.007

Van der Wal, P. G., Bolink, A. H., and Merkus, G. S. (1988). Differences in quality characteristics of normal, PSE and DFD pork. *Meat Sci.* 24 (1), 79–84. doi:10.1016/0309-1740(89)90009-0

van Laack, R. L., and Kauffman, R. G. (1999). Glycolytic potential of red, soft, exudative pork longissimus muscle. *J. Anim. Sci.* 77 (11), 2971–2973. doi:10.2527/1999.77112971x

Wang, P., Cong, M., Liu, T., Li, Y., Liu, L., Sun, S., et al. (2020). FoxA2 inhibits the proliferation of hepatic progenitor cells by reducing PI3K/Akt/HK2-mediated glycolysis. *J. Cell Physiol.* 235 (12), 9524–9537. doi:10.1002/jcp.29759

Wu, D., Harrison, D. L., Szasz, T., Yeh, C-F., Shentu, T-P., Meliton, A., et al. (2021). Single-cell metabolic imaging reveals a SLC2A3-dependent glycolytic burst in motile endothelial cells. *Nat. Metab.* 3 (5), 714–727. doi:10.1038/s42255-021-00390-y

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, X., Jiang, A., et al. (2020). Whole-genome sequencing association analysis reveals the genetic architecture of meat quality traits in Chinese Qingyu pigs. *Genome* 63 (10), 503–515. doi:10.1139/gen-2019-0227

Yang, Q., Cui, J., Chazaro, I., Cupples, L. A., and Demissie, S. (2005). Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet.* 6 (1), S134. doi:10.1186/1471-2156-6-S1-S134

Yang, Y., Shen, F., Huang, W., Qin, S., Huang, J-T., Sergi, C., et al. (2019). Glucose is involved in the dynamic regulation of m6A in patients with type 2 diabetes. *J. Clin. Endocrinol. Metab.* 104 (3), 665–673. doi:10.1210/jc.2018-00619

Yoon, S., Molloy, M. J., Wu, M. P., Cowan, D. B., and Gussoni, E. (2007). C6ORF32 is upregulated during muscle cell differentiation and induces the formation of cellular filopodia. *Dev. Biol.* 301 (1), 70–81. doi:10.1016/j.ydbio.2006.11.002

Zhang, Y., Zhang, J., Gong, H., Cui, L., Zhang, W., Ma, J., et al. (2019). Genetic correlation of fatty acid composition with growth, carcass, fat deposition and meat quality traits based on GWAS data in six pig populations. *Meat Sci.* 150, 47–55. doi:10.1016/j.meatsci.2018.12.008

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

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/fgene.2023.1080279

Check for updates

OPEN ACCESS

EDITED BY
Kefei Chen,
Curtin University, Australia

REVIEWED BY
Peipei Ma,
Shanghai Jiao Tong University, China
Lingyang Xu,
Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (CAAS), China

*CORRESPONDENCE
Jun He,
✉ hejun@hunau.edu.cn
Xiao-Lin Wu,
✉ xuw8@wisc.edu

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

RECEIVED 26 October 2022
ACCEPTED 14 March 2023
PUBLISHED 28 March 2023

CITATION
Li Z, He J, Yang F, Yin S, Gao Z, Chen W, Sun C, Tait RG, Bauck S, Guo W and Wu X-L (2023), A look under the hood of genomic-estimated breed compositions for brangus cattle: What have we learned?
Front. Genet. 14:1080279.
doi: 10.3389/fgene.2023.1080279

A look under the hood of genomic-estimated breed compositions for brangus cattle: What have we learned?

Zhi Li[1], Jun He[1]*, Fang Yang[1], Shishu Yin[1], Zhendong Gao[1], Wenwu Chen[1], Chuanyu Sun[2], Richard G. Tait[2], Stewart Bauck[2], Wei Guo[3] and Xiao-Lin Wu[3,4]*

[1]College of Animal Science and Technology, Hunan Agricultural University, Changsha, Hunan, China, [2]Biostatistics and Bioinformatics, Neogen GeneSeek, Lincoln, NE, United States, [3]Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, United States, [4]Council on Dairy Cattle Breeding, Bowie, MD, United States

The Brangus cattle were developed to utilize the superior traits of Angus and Brahman cattle. Their genetic compositions are expected to be stabilized at 3/8 Brahman and 5/8 Angus. Previous studies have shown more than expected Angus lineage with Brangus cattle, and the reasons are yet to be investigated. In this study, we revisited the breed compositions for 3,605 Brangus cattle from three perspectives: genome-wise (GBC), per chromosomes (CBC), and per chromosome segments (SBC). The former (GBC) depicted an overall picture of the “mosaic” genome of the Brangus attributable to their ancestors, whereas the latter two criteria (CBC and SBC) corresponded to local ancestral contributions. The average GBC for the 3,605 Brangus cattle were 70.2% Angus and 29.8% Brahman. The K-means clustering supported the postulation of the mixture of 1/2 Ultrablack (UB) animals in Brangus. For the non-UB Brangus animals, the average GBC were estimated to be 67.4% Angus and 32.6% Brahman. The 95% confidence intervals of their overall GBC were 60.4%–73.5% Angus and 26.5%–39.6% Brahman. Possibly, genetic selection and drifting have resulted in an approximately 5% average deviation toward Angus lineage. The estimated ancestral contributions by chromosomes were heavily distributed toward Angus, with 27 chromosomes having an average Angus CBC greater than 62.5% but only two chromosomes (5 and 20) having Brahman CBC greater than 37.5%. The chromosomal regions with high Angus breed proportions were prevalent, tending to form larger blocks on most chromosomes. In contrast, chromosome segments with high Brahman breed proportion were relatively few and isolated, presenting only on seven chromosomes. Hence, genomic hitchhiking effects were strong where Angus favorable alleles resided but weak where Brahman favorable alleles were present. The functions of genes identified in the chromosomal regions with high (≥75%) Angus compositions were diverse yet may were related to growth and body development. In contrast, the genes identified in the regions with high (≥37.5%) Brahman compositions were primarily responsible for disease resistance. In conclusion, we have addressed the questions concerning the Brangus genetic make-ups. The results can help form a dynamic picture of the Brangus breed formation and the genomic reshaping.

KEYWORDS

admixture model, ancestry, composite animals, genomic hitchhiking, linkage disequiblibrium, SNP

Introduction

Brangus beef cattle were developed to combine the desirable traits of Angus and Brahman cattle. Angus cattle are well known for their superior carcass qualities, and Angus cows have excellent fertility and milking capability. The Brahman cattle have developed disease resistance, overall hardiness, and outstanding maternal instincts thanks to rigorous natural selection. The crossbreeding to create the Brangus breed dated to 1932, according to the USDA 1935 Yearbook in Agriculture. Yet, Brangus registration by the International Brangus Breeders Association (IBBA) started in 1949. For official registration, a Brangus animal needs to be genetically stabilized at 3/8 Brahman and 5/8 Angus by pedigree, be solid black or red, and be polled (Briggs and Briggs, 1980). Both sire and dam must be recorded with IBBA (San Antonio, TX). Hence, knowing the breed compositions of individual animals is a requisite to official animal registrations. Such information also allows for utilizing the "stable" heterosis to explore methods for predicting heterosis (Akanno et al., 2017), and it permits the implementation of precise animal farming management decisions (Berry, 2019).

After the breed formation, subsequent *inter-se* mating and selection have been conducted with Brangus over time. For example, the United States IBBA has developed expected progeny differences (EPD) for quantitative traits, such as birth weight, weaning weight, yearling weight, milk production, total maternal calving ease, and intramuscular fat. Likely, artificial selection pressure employed at varying levels on these traits of interest in the past decades could have resulted, to some extent, in the deviation of the ancestral genomic proportions in Brangus from the previously targeted ratios. Previous studies showed significantly elevated Angus genomic breed composition (GBC) in Brangus cattle (He et al., 2018; Paim T. D. P. et al., 2020; Paim T. D. P. et al., 2020; Li et al., 2020; Wang et al., 2020; Wu et al., 2020), and the reasons are yet to be revealed. There were several plausible assumptions. For example, it was postulated that selecting Brangus for Angus favorable traits (e.g., carcass, growth, feed efficiency) could increase the genomic breed compositions of Brangus cattle toward Angus (Paim T. D. P. et al., 2020; Wu et al., 2020). In theory, the proportion of actual genotypes passed from one generation to the next can vary between individuals owing to Mendelian sampling, genetic recombination rate and linkage disequilibrium (LD) (Falconer and Mackay, 1996). Hence, selecting Brangus for traits more prevalent in Angus (e.g., carcass, growth, feed efficiency) can favor Angus alleles, sweeping more 'Angus' haplotypes to further generations. Another possible reason could be the mixture of 1/2 Ultrablack (UB) animals (i.e., the first-generation progenies derived from Brangus × Angus), which are expected to have 81.25% Angus lineage. In October 2005, the International Brangus Breeders Association (IBBA) board of directors approved the creation of the Ultrablack and Ultrared (UR) program to take advantage of the strengths of the Brangus and Angus or Red Angus breeds, which combine environmental adaptability and maternal excellence of Brangus with the exceptional marbling, calving ease and name recognition of Angus or Red Angus. The UB and UR animals are registered composite animals with a validated and documented lineage between 12.5% and 87.5% Brangus breeding. The remaining 87.5%–12.5% must be a registered Angus to be a UB or a Red Angus to be a UR. The second assumption is likely, yet scientifically supporting evidence is needed. Apart from the possible mixture, the actual genomic breed compositions of Brangus cattle are not known precisely after decades of crossbreeding and selection.

Selection may have left signatures on the genome after the breed formation (Goszczynski et al., 2017; Paim T. D. P. et al., 2020). These genomic regions with selection sweeps can have different breed compositions than expected due to the selective advantages of genes from one of the founders. Paim T. D. P. et al. (2020) evaluated the overall ancestral breed compositions and local ancestral contributions by chromosomes in Brangus cattle. They also related haplotypes to ancestral traits under selection. Such information could lead to a better understanding of how hybridization and crossbreeding systems have shaped the genetic architecture of these composite animals. However, their conclusions were built on small samples of the two founder breeds with genotypes (i.e., 68 Brahman and 95 Angus). Statistically, the inference of allelic frequencies and haplotypes, and, therefore, ancestry origins, are subject to large errors in small samples.

In this study, we revisited the estimation of breed compositions for Brangus cattle from three perspectives, genomic-wise, per chromosome, and per chromosome segment. The genotyped animals in the two ancestral breeds included 20,359 Angus and 509 Brahman cattle. Hence, our sample sizes for the two ancestral breeds were significantly larger than those used by Paim T. D. P. et al. (2020). We took a consistent approach to estimate ancestral breed compositions from the three perspectives; all were assessed with an admixture model. Three measures of breed compositions were defined: 1) genomic-estimated breed compositions (GBC), 2) chromosomal-estimated breed compositions (CBC), and 3) segmental-estimated breed compositions (SBC). Note that the latter two quantities corresponded to local ancestral genomic contributions, measured per chromosome and chromosomal regions, respectively. Possible population stratification was inferred based on global ancestral genomic proportions, whereas genomics dynamics due to crossbreeding and selection were visualized through local ancestral contributions.

Materials and methods

Animals and genotype data

The experimental data consisted of 3,605 Brangus cattle, 20,359 Angus cattle, and 483 Brahman cattle. The latter two ancestral breeds were used as the reference populations for estimating GBC for Brangus cattle. All the animals were genotyped with a GeneSeek Genomic Profiling (GGP) bovine 50 K V1 (version 1) chip, except 349 Brahman cattle were genotyped with an Illumina 777 K bovine SNP chip (Table 1).

TABLE 1 **Descriptive statistics of genotype data for Brangus and their ancestral breeds (Angus and Brahman)[a].**

Type	Breed	Number of animals	Number of SNPs	Allele A frequency	
				Mean	SD
Composite	Brangus	3,605	49,463	0.477	0.231
Ancestry	Angus	20,359 (20,322)	49,463	0.492	0.247
	Brahman	349 (349)	777,962	0.439	0.343
		160 (134)	49,463	0.431	0.363

[a]The numbers in the brackets are genotyped animals that remained after data cleaning.

The genotypes were extracted from the Neogen GeneSeek genotyping databases representing samples shared between the Neogen global laboratories. The SNP map positions were based on the UMD 3.1 reference bovine genome assembly (Merchant et al., 2014).

Reference SNPs were selected from the common set between the GGP bovine 50 K V1 chip and the Illumina 777 K bovine SNP Beadchip. The data cleaning removed SNPs with a call rate of less than 95%, SNPs on the two sex chromosomes, and SNPs without map position. SNPs violating the Hardy-Weinberg equilibrium ($p <$ 1.0E-8) were also excluded. For SNPs with greater than 0.99 correlations on each chromosome, only the one with higher or the highest minor allelic frequency was kept. When there were ties in allelic frequencies, a random one was taken. The final SNP set retained 41,672 common SNPs for the subsequent analyses. In each ancestral population, outlier individuals were excluded as those with (−2)log (likelihood) exceeding a given cutoff value (i.e., 2.0 by default) (He et al., 2018). This test excluded 37 Angus and 26 Brahman animals from the reference populations. The means and standard deviations of allele A frequencies of the SNPs after data cleaning in the three populations are shown in Table 1. On average, Angus cattle had a higher allele A frequency (0.492) than Brahman cattle (0.431–0.439). The average allele A frequency for Brangus cattle was 0.477, which fell between the two ancestral populations yet closer to Angus. The Angus population had a smaller standard deviation of allele A frequency than the Brahman population. The Brangus cattle had a smaller standard deviation of allele A frequencies than the two ancestral populations (Table 1).

Estimation of breed compositions

Breed compositions were estimated for individual Brangus animals using the admixture model by Bansal and Libiger (2015) (BL-Admixture). This method utilizes the same form of likelihood model as in the STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009) software packages. However, the BL-Admixture model runs faster. In particular, STRUCTURE takes a Bayesian approach and relies on a Markov Chain Monte Carlo (MCMC) algorithm to sample the posterior distribution, which can extremely computationally intensive with large data. STRUCTURE and Admixture represent unsupervised analysis of the ancestry of multiple individuals and jointly estimate allele frequencies for the ancestral populations and the relative contribution of each ancestral population to each individual's genome. The BL-Admixture estimates the ancestry for a single individual using information about allele frequencies at a large number of loci for multiple reference populations. The latter allele frequencies are obtained from previous unsupervised admixture analysis, or simply from the reference population assuming no genetic drift and selection after these breeds were formed. We took the latter approach.

Consider one biallelic locus (say j) genotyped on an animal (say i). Assume that allele frequencies on this locus are known for each ancestry breed. The admixture model postulates that a progeny's genotype at this locus is determined according to an allelic frequency as a weighted average of T reference (ancestry) breeds: $f_{ij} = \sum_{t=1}^{T} w_{it} q_{tj}$, where q_{tj} is the frequency of allele B for the jth SNP in the tth reference population,; w_{it} is the corresponding weight. Then, under the assumption of Hardy-Weinberg equilibrium, the probabilities of observing each genotype (g_{ij}) on this animal are the following:

$$Pr\left(g_{ij} \middle| f_{ij}\right) = \begin{cases} \left(1 - f_{ij}\right)^2 & g_{ij} = 0 \\ 2f_{ij}\left(1 - f_{ij}\right) & g_{ij} = 1 \\ f_{ij}^2 \quad g_{ij} = 2 \end{cases} \tag{1}$$

Now, consider $j = 1, \ldots, M$ SNPs genotyped on this animal, assuming their mutual independence. The log-likelihood computed for all the M SNPs measured on the ith animal, denoted by l_i, is as follows.

$$\begin{aligned} l_i &= \sum_{j=1}^{M} \ln\left(Pr\left(g_{ij} \middle| f_{ij}\right)\right) \\ &= \left[\sum_{j=1}^{M} g_{ij} \ln\left(f_{ij}\right) + \left(2 - g_{ij}\right) \ln\left(1 - f_{ij}\right)\right] + C \end{aligned} \tag{2}$$

where $C = \sum_{j=1}^{M} ln \begin{pmatrix} 2 \\ g_{ij} \end{pmatrix}$. Note that the assumption of mutual independence between SNPs does not hold precisely due to linkage or/and random associations between them. Nevertheless, the admixture model is often robust to this assumption violation because the percentage of SNPs in high LD are rare when using moderate to high-density SNP panels (He et al., 2018; Li et al., 2020).

The admixture coefficients, $\boldsymbol{W_i} = (\, w_{i1} \;\ldots\; w_{iT} \,)'$, taken to the GBC of animal i attributable to the reference (ancestral) breeds, are obtained using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method under the restrictions that $w_{ik} \geq 0$; $\sum_{k=1}^{T} w_{iT} = 1$ (Nocedal and Wright, 2006). BFGS is a powerful, Quasi-Newton second derivative line search family method to solve non-linear optimization problems. Optimizing the likelihood function of BFGS iteratively removed the non-zero mixing coefficient, which

TABLE 2 Mean and standard deviation (SD) of R^2 linkage disequilibrium (LD) among SNPs evaluated by varying window sizes (1 Mb, 5 Mb, and 10 Mb) on each chromosome[a].

Chromosome	Window = 1 Mb			Window = 5 Mb			Window = 10 Mb		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
1	159	0.144	0.044	32	0.089	0.016	16	0.068	0.014
2	137	0.137	0.042	28	0.087	0.015	14	0.066	0.010
3	122	0.155	0.052	25	0.095	0.029	13	0.073	0.025
4	121	0.159	0.054	25	0.101	0.028	13	0.081	0.026
5	122	0.135	0.049	25	0.088	0.025	13	0.068	0.019
6	120	0.127	0.051	24	0.073	0.020	12	0.055	0.016
7	113	0.150	0.070	23	0.090	0.028	12	0.069	0.021
8	114	0.153	0.062	23	0.093	0.028	12	0.072	0.021
9	106	0.129	0.049	22	0.083	0.023	11	0.064	0.014
10	105	0.131	0.052	21	0.080	0.015	11	0.063	0.014
11	108	0.148	0.049	22	0.086	0.020	11	0.064	0.015
12	91	0.136	0.050	19	0.088	0.024	10	0.070	0.019
13	85	0.156	0.057	17	0.097	0.031	9	0.071	0.017
14	84	0.164	0.063	17	0.099	0.029	9	0.070	0.013
15	86	0.157	0.055	18	0.104	0.038	9	0.073	0.016
16	82	0.150	0.078	17	0.091	0.039	9	0.069	0.021
17	75	0.120	0.042	15	0.066	0.007	8	0.051	0.010
18	66	0.145	0.047	14	0.082	0.017	7	0.058	0.005
19	64	0.126	0.038	13	0.077	0.013	7	0.059	0.009
20	72	0.125	0.045	15	0.083	0.030	8	0.073	0.043
21	72	0.146	0.090	15	0.086	0.026	8	0.066	0.018
22	62	0.129	0.033	13	0.078	0.016	7	0.060	0.016
23	53	0.130	0.083	11	0.077	0.015	6	0.061	0.021
24	63	0.138	0.051	13	0.083	0.012	7	0.065	0.019
25	43	0.107	0.038	9	0.062	0.013	5	0.050	0.018
26	52	0.146	0.043	11	0.091	0.016	6	0.072	0.022
27	46	0.129	0.041	10	0.084	0.043	5	0.056	0.017
28	47	0.113	0.036	10	0.074	0.021	5	0.053	0.010
29	52	0.122	0.046	11	0.079	0.026	6	0.069	0.039

[a]N = number of segments with valid SNP, genotypes on a chromosome.

did not significantly improve the model fitting, thus obtaining a concise set of individual mixture coefficients. In the admixture model, the value of each admixture coefficient is bounded between 0 and 1, and the sum of admixture coefficients (GBC) computed for each animal is one under the assumption of 100% genetic contributions by the T ancestral breeds to each animal.

Breed compositions for the Brangus animals were estimated genome-wide, per chromosome, and per chromosomal segment, respectively. The number of reference SNPs per chromosome varied from 712 (chromosome 25) to 2,568 (chromosome 1), and the average distance between SNPs ranged from 0.05 to 0.07 Mb. SBC were obtained on three window sizes: 1 Mb, 5 Mb, and 10 Mb, respectively. The widow sizes were taken arbitrarily yet still based on two factors, the average length of gene in the bovine genome and the minimum number of SNPs to give stable estimates. To minimize the errors in the estimated SBC due to insufficient SNP coverage, chromosomal segments with less than five SNPs were excluded from computing SBC. There were 2,522 1-Mb segments, 518 5-Mb

TABLE 3 Means and standard deviations (SD) of genomic-estimated breed composition (GBC) for the 3,605 Brangus cattle as a single population (K = 1) and as subpopulations according to K-means clustering (K = 2 and 3, respectively)[1,2,3].

K	Group	Number of animals	Angus-GBC %		Brahman %	
			Mean	SD	Mean	SD
1	A	3,605	70.22	6.75	29.78	6.75
2	B-1	713	81.63	4.66	18.37	4.66
	B-2	2,892	67.40	3.37	32.60	3.37
3	C-1	640	82.08	3.62	17.92	3.62
	C-2	381	70.93	4.03	29.07	4.03
	C-3	2,584	67.17	3.40	32.83	3.40

segments, and 269 10-Mb segments, respectively. The average number of SNPs per segment ranged from 15.2 to 18.8 (1 Mb), 72.0 to 91.1 (5 Mb), and 132.0 to 169.1 (10 Mb), respectively (Table 2). Overall, the average R^2 LD per segment decreased as the window size increased, ranging from 0.107 to 0.164 on 1 Mb chromosomal regions, from 0.062 to 0.104 on 5-Mb chromosomal regions, and from 0.050 to 0.081 on 10 Mb chromosomal regions (Table 2). The R^2 LD on 1-Mb regions also had a larger average standard deviation (0.05) than those on 5-Mb and 10-Mb regions (0.02).

Clustering of brangus cattle

K-means clustering (Jain, 2010) was conducted on the GBC for the 3,605 Brangus cattle to reveal possible population stratifications, where $K = 2$ and 3, respectively. Initially, the number of clusters K was specified *a priori*, and randomly assigned all the animals to each of K distinct, non-overlapping clusters as their initial clusters. Then, the K-means algorithm computed the cluster centroid for each cluster, which is a vector of genotype means for the M reference SNPs, and it re-assigned each animal to the cluster whose centroid was the closest. The last two steps proceeded iteratively till the total within-clustering variation, defined by the sum of all the pairwise squared Euclidean distance in each cluster and summed over all the K clusters, was minimized as much as possible (Hartigan and Wong, 1979):

$$minimize\left\{\sum\nolimits_{k=1}^{K}\frac{1}{n_k}\sum\nolimits_{i,i'\in C_k}\sum\nolimits_{j=1}^{M}\left(g_{ij}-g_{i'j}\right)^2\right\}$$

where C_k stands of cluster k, and n_k is the number of animals in the kth cluster. The K-mean clustering analysis was implemented by the "stats" R package.

Gene set enrichment analysis

Gene ontology (GO) term enrichment analysis, or gene set enrichment analysis, was conducted using the "gprofiler" R package with genes identified in 1-Mb chromosomal regions featuring either ancestral breed. This R package performed functional gene enrichment analysis on the input gene lists, mapped genes to known functional resources, and detected statistically significantly enriched terms. Briefly, chromosomal segments satisfying Angus SBC ≥0.75 or Brahman SBC ≥0.50 were extracted. Both cutoff thresholds represented equal upward GBC deviations (i.e., 12.5%) from their expected values. Then, the gene information was extracted by querying in the Ensembl database according to their chromosomal locations on the selected chromosomal segments. Finally, GO clustering analysis was performed on the gene list generated with the different filters. The annotation databases included Gene Ontology–biological processes, cellular components, molecular function (http://geneontollogy.org/) (Ashburner et al., 2000), and Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome. jp/kegg/) (Kanehisa et al., 2016). Other relevant R packages used in this study included "org. Bt.e.g., db", "biomaRt" (Smedley et al., 2015), and "clusterProfiler" (Yu et al., 2012). Gene lists were extracted using the "getBM" function in the "biomaRt" R package. Briefly, given a set of filters (e.g., the

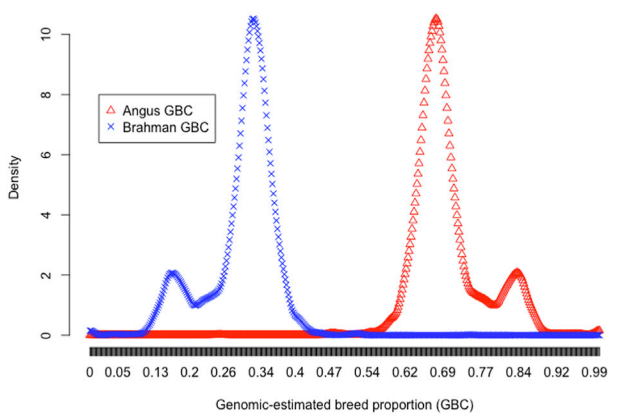


FIGURE 1
Density plots of genomic-estimated Angus (red triangle) and Brahman (blue cross) breed compositions for 3,605 Brangus cattle.

chromosome number, the start and end positions), it retrieved attributes of all genes in this interval from the BioMart database. The attributes information includes "ensembl gene id", "chromosome number", "gene start position", "gene end position", and "gene description", etc. Finally, gene extraction was performed on all eligible SBC fragments, and all extracted genes were merged to form a gene list after removing duplication. Enrichment analysis was then performed. All the QTLs were queried and aggregated in the QTLdb database (Release 49, https://www.animalgenome.org/cgi-bin/QTLdb/BT/index) (Hu et al., 2022).

Results and discussion

Genomic-estimated breed compositions

The estimated GBC for the 3,605 Brangus cattle, on average, were 70.2% Angus and 29.8% Brahman (Table 3), which significantly deviated from the officially expected values (i.e., 62.5% Angus and 37.5% Brahman) ($p < 2.2e\text{-}16$). The 95% confidence intervals were 61.1%–85.8% Angus and 14.2%–38.9% Brahman. There were 459 (12.7%) Brangus cattle with Angus GBC ≥80.0% and 19 (0.5%) Brangus cattle with Angus GBC ≥90.0%. Similarly, elevated Angus GBC for the Brangus animals were documented in some previous studies. For example, Paim T. D. P. et al. (2020) estimated that Brangus were 70.4% Angus and 29.6% Brahman based on high-density SNP genotypes (777,962 SNP, BovineHD Beadchip, Illumina, San Diego, CA, United States). Li et al. (2020) showed that Brangus cattle was 69.8%–70.5% Angus and 29.5%–30.2% Brahman based on multiple models, including an admixture model, linear regression, and ridge-regression BLUP, each with a selectively uniform 20 K SNP panel. Wang et al. (2020) proposed using regularized admixture models to estimate GBC for purebred animals to deal with the so-called "Impure Purebred Paradox", a phenomenon suggesting a higher-than-expected false-negative rate in the identification of purebred animals. They showed that Brangus were, on average, 71.1%–77.1% Angus and 22.9%–28.9% Braham based on various regularized admixture models. Using path analysis model, Wu et al. (2020) showed that Brangus cattle were 68.2%–71.8% Angus and 28.2%–31.8% Brahman. Hence, regardless of the population sizes and the statistical methods used, these studies have consistently pinpointed that significantly higher-than-expected Angus breed proportions in Brangus cattle.

Multiple reasons are likely responsible for the elevated Angus lineage in Brangus. Firstly, population stratification could exist with Brangus, given the fact that the IBBA approved the creation of UB and UR animals in 2005. The density plots of the estimated Angus or Brahman breed proportions in the 3,605 animals were bimodal, which served as preliminary evidence for the Brangus population stratification (Figure 1). Then, *K*-means clustering was conducted to partition the 3,605 animals into two and three clusters, respectively (Table 3). With K = 2 (i.e., two clusters), the B-1 cluster with a higher average Angus GBC (81.6%) consisted of 713 (19.8%) Brangus cattle, whereas the B-2 cluster with a lower average Angus GBC (67.4%) included 2,892 (80.2%) animals. With $K = 3$ (i.e., three clusters), the cluster (C-1) with the highest average Angus GBC (82.1%) had 640 (17.8%) Brangus animals. The cluster with the lowest average Angus GBC (67.2%) included 2,584 (71.7%) Brangus animals. In both sets of clustering results ($K = 2$ *versus* $K = 3$), the average Angus GBC for animals in the top Angus composition clusters (B-1 *versus* C-1) agreed approximately with each other (81.6% *versus* 82.1%), and they corresponded roughly to the expected Angus breed proportion (81.25%) for the ½ UB animals. We thus suspected that these animals could be the ½ UB animals. Using a path analysis approach, Wu et al. (2020) confirmed that the ½ UB animals were, on average, 81.25% Angus and 18.75% Brahman. The path analysis decomposed the relationships between the ancestors and the composite animals into direct and indirect path effects. The above percentage only accounted for direct effects by the path-analysis interpretation, assuming a zero correlation between the two ancestral breeds (Wu et al., 2020). In reality, however, Angus and Brahman cattle are connected due to sharing common remote ancestors, though the correlation can be low. For example, the correlation of allele A frequencies ranged from 0.05 to 0.10 between the two ancestral breeds, subject to the SNP panel sizes. If considering the indirect path effects, the actual Angus breed proportions for the ½ UB animals could be slightly higher. On the other hand, the two majority clusters, B-2 and C-3, had roughly comparable averages of Angus GBC (67.4% *versus* 67.2%). These were likely Brangus cattle without the mixture of ½ UB animals. The average Angus breed proportion for non-UB Brangus was 67.4% based on the clustering analysis with $K = 2$. The 95% confidence interval of Angus breed proportions in Brangus was between 60.4% and 73.5%. Hence, there was, on average, an approximately 5% deviation of GBC toward Angus breed proportions in non-UB Brangus cattle, possibly resulting from selecting Brangus cattle for phenotypes where Angus has advantages. Without selection, genomic-estimated breed compositions would agree approximately with the expected ratios (e.g., Funkhouser et al., 2017; Gobena et al., 2018). Note that ½ UB animals are not precisely registered Brangus. They are officially given a "UB" prefix for registration purposes and are shown on their registered IDs. It is also worth mentioning that the B-2 (or C-3) cluster could include some advanced UB crosses because they had comparable Angus breed proportions as those in the non-UB Brangus cattle in clusters B-2 and C-3. In early 2013, the IBBA further approved the breeding-up of UB (or UR) cattle to registered Brangus. The rule states that an IBBA-

TABLE 4 Means and standard deviations (SD) of chromosomal-estimated breed compositions (CBC) for the 3,605 Brangus cattle.

Chromosome	Angus-CBC %		Brahman-CBC %		Number of SNPs	Average spacing, Mb
	Mean	SD	Mean	SD		
1	72.31	12.28	27.69	12.28	2,568	0.06
2	70.19	14.10	29.81	14.10	2,218	0.06
3	74.67	12.51	25.33	12.51	2,081	0.06
4	75.67	12.24	24.33	12.24	1,889	0.06
5	56.86	15.91	43.14	15.91	2,120	0.06
6	69.7	14.18	30.3	14.18	1,988	0.06
7	67.66	15.17	32.34	15.17	1,815	0.06
8	70.08	14.26	29.92	14.26	1,774	0.06
9	65.94	15.57	34.06	15.57	1,830	0.06
10	73.77	12.89	26.23	12.89	1,697	0.06
11	72.27	13.50	27.73	13.50	1,709	0.06
12	67.47	16.30	32.53	16.30	1,419	0.06
13	65.13	16.82	34.87	16.82	1,446	0.06
14	73.58	15.22	26.42	15.22	1,405	0.06
15	79.84	12.89	20.16	12.89	1,371	0.06
16	64.74	18.31	35.26	18.31	1,334	0.06
17	79.21	13.09	20.79	13.09	1,214	0.06
18	73.86	13.07	26.14	13.07	1,152	0.06
19	71.51	15.47	28.49	15.47	1,184	0.05
20	59.76	16.22	40.24	16.22	1,352	0.05
21	62.58	16.53	37.42	16.53	1,215	0.06
22	78.17	13.51	21.83	13.51	1,002	0.06
23	72.09	14.81	27.91	14.81	953	0.06
24	72.97	14.72	27.03	14.72	1,045	0.06
25	73.33	17.64	26.67	17.64	712	0.06
26	78.30	15.38	21.70	15.38	862	0.06
27	78.66	15.64	21.34	15.64	736	0.06
28	67.99	17.40	32.01	17.40	789	0.06
29	68.77	18.53	31.23	18.53	792	0.07
Average	70.93	14.97	29.07	14.97	1,437	0.06

registered Brangus sire or dam mated to an IBBA-registered UB or UR sire or dam that results in at least 7/8th (87.5%) Brangus also qualifies as a registered Brangus. For example, mating a ¾ UB animal (i.e., progenies derived from crossing ½ UB animals with Brangus) to a registered Brangus resulted in 7/8 UB animals, which meets the 87.5% Brangus makeup. On average, a 7/8 UB animal was 67.2% Angus. Finally, the average Angus GBC for the animals in cluster C2 was 70.9%, which roughly corresponded to the expected Angus lineage (71.9%) for ¾ UB animals.

Chromosomal-estimated breed compositions

The ancestral breed proportions were estimated by chromosomes in the 3,605 Brangus (Table 4). Overall, the average estimated CBC per chromosome varied substantially, from 56.9% (chromosome 5) to 79.8% (chromosome 15). Converse to the Angus CBC, the average Brahman CBC was the lowest on chromosome 15 (20.2%) and the highest on chromosome

TABLE 5 Minimum (min), maximum (max), and mean of segmental-estimated breed compositions (SBC) per chromosome for Brangus.

Chrom	Window = 1 Mb						Window = 5 Mb						Window = 10 Mb					
	SBC-Angus (%)			SBC-Brahman (%)			SBC-Angus (%)			SBC-Brahman (%)			SBC-Angus (%)			SBC-Brahman (%)		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
1	39.8	95.6	71.5	4.4	60.2	28.5	52.2	95.4	75.3	4.6	47.8	24.7	53.4	92.6	74.5	7.4	46.6	25.5
2	25.6	89.1	68.1	10.9	74.4	31.9	47.0	87.4	72.2	12.6	53.0	27.8	53.3	82.8	71.9	17.2	46.7	28.1
3	49.9	92.0	73.4	8.0	50.1	26.6	51.6	90.9	75.3	9.1	48.4	24.8	51.6	88.9	74.0	11.1	48.4	26.0
4	46.5	95.3	73.9	4.7	53.5	26.1	62.7	88.5	76.1	11.5	37.3	23.9	65.1	86.1	75.5	13.9	34.9	24.5
5	31.1	79.4	57.6	20.6	68.9	42.4	35.6	80.9	59.2	19.1	64.4	40.8	37.9	80.9	59.7	19.1	62.2	40.3
6	43.5	91.6	67.0	8.4	56.5	33.0	52.4	87.7	69.5	12.3	47.6	30.5	55.5	82.4	69.4	17.6	44.5	30.6
7	38.7	88.1	65.9	11.9	61.3	34.1	51.0	86.0	69.0	14.1	49.0	31.0	58.2	84.4	68.7	15.6	41.8	31.3
8	43.3	91.6	68.6	8.4	56.7	31.4	57.9	83.2	70.6	16.8	42.2	29.4	60.7	79.9	70.0	20.1	39.3	30.0
9	37.0	90.9	66.3	9.2	63.0	33.7	49.8	89.2	69.6	10.8	50.3	30.4	49.5	86.9	69.1	13.1	50.5	30.9
10	38.2	92.7	71.9	7.3	61.8	28.1	51.4	92.3	75.5	7.7	48.6	24.5	59.2	90.1	75.0	9.9	40.8	25.0
11	44.6	88.7	70.7	11.3	55.4	29.3	58.8	87.6	73.1	12.4	41.2	26.9	61.5	87.4	72.6	12.6	38.5	27.4
12	41.3	90.0	65.5	10.0	58.7	34.5	48.8	86.1	69.1	13.9	51.2	31.0	50.3	86.1	70.1	13.9	49.7	29.9
13	35.7	84.5	64.0	15.5	64.3	36.0	54.0	81.4	65.2	18.6	46.1	34.8	53.9	74.0	65.6	26.0	46.1	34.5
14	35.7	89.0	71.2	11.0	64.3	28.8	53.5	84.6	74.7	15.4	46.5	25.3	54.1	83.4	74.9	16.6	45.9	25.1
15	48.6	91.1	76.9	8.9	51.5	23.1	68.6	89.7	81.4	10.3	31.4	18.6	70.5	86.6	81.4	13.4	29.5	18.7
16	40.6	88.8	63.1	11.3	59.4	37.0	54.7	81.3	65.2	18.7	45.3	34.8	54.9	72.8	64.7	27.3	45.2	35.3
17	56.3	91.3	75.6	8.7	43.7	24.4	70.8	88.4	80.9	11.6	29.2	19.1	75.3	88.1	80.7	12.0	24.8	19.3
18	45.4	95.9	72.5	4.1	54.7	27.5	47.0	95.6	74.4	4.4	53.1	25.6	51.7	94.0	74.4	6.0	48.3	25.6
19	48.5	89.0	70.8	11.0	51.5	29.2	62.7	82.8	72.1	17.2	37.3	27.9	65.0	82.5	71.4	17.5	35.0	28.6
20	33.6	86.5	59.8	13.5	66.5	40.2	41.4	76.9	61.8	23.1	58.7	38.2	48.2	75.5	61.4	24.6	51.9	38.6
21	39.9	84.5	61.5	15.5	60.1	38.5	42.1	84.7	64.6	15.3	57.9	35.4	51.5	75.4	63.5	24.6	48.5	36.5
22	55.4	93.6	76.0	6.4	44.6	24.0	69.8	87.8	79.4	12.2	30.2	20.6	72.5	85.2	79.3	14.8	27.5	20.7
23	43.3	91.3	69.8	8.7	56.7	30.2	61.3	89.3	73.5	10.7	38.7	26.6	63.8	85.4	74.7	14.6	36.2	25.3
24	48.5	90.5	72.4	9.5	51.6	27.6	60.5	87.4	74.0	12.6	39.5	26.1	60.8	81.4	73.1	18.7	39.2	27.0

(Continued on following page)

TABLE 5 (*Continued*) Minimum (min), maximum (max), and mean of segmental-estimated breed compositions (SBC) per chromosome for Brangus.

Chrom	Window = 1 Mb						Window = 5 Mb						Window = 10 Mb					
	SBC-Angus (%)			SBC-Brahman (%)			SBC-Angus (%)			SBC-Brahman (%)			SBC-Angus (%)			SBC-Brahman (%)		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
25	46.8	83.8	69.9	16.2	53.2	30.1	56.6	85.7	74.0	14.3	43.4	26.0	56.6	80.8	71.9	19.2	43.4	28.1
26	27.7	89.6	73.2	10.4	72.3	26.8	69.4	89.6	81.0	10.4	30.6	19.0	73.4	87.7	80.8	12.3	26.6	19.2
27	57.9	90.9	75.9	9.1	42.1	24.1	71.7	88.9	81.2	11.2	28.3	18.8	72.7	87.2	81.0	12.8	27.3	19.0
28	51.3	89.4	67.1	10.6	48.7	32.9	60.0	81.6	69.7	18.4	40.0	30.3	63.5	75.0	69.5	25.0	36.5	30.5
29	39.2	84.3	67.3	15.7	60.8	32.7	57.6	84.0	69.6	16.0	42.4	30.4	57.1	78.6	68.6	21.4	42.9	31.4
Overall	25.6	95.9	69.2	4.1	74.4	30.8	35.6	95.6	72.3	4.4	64.4	27.7	37.9	94.0	72.0	6.0	62.2	28.0

5 (CBC 43.1%). We noted that the estimated ancestral breed proportions were heavily distributed toward Angus, which agreed with the breeding target for greater Angus "blood" than Brahman. There were 27 chromosomes with Angus CBC greater than 62.5%, but only two chromosomes (5 and 20) had Brahman CBC greater than 37.5%. Our results coincided with a previous study by Paim T. D. P. et al. (2020). They showed that chromosome 15 had the highest Angus proportion (84.7%), and chromosome five had the largest Brahman proportion (43.7%). Still, there were some differences. Paim T. D. P. et al. (2020) used principal component analysis to describe the population relationships. They showed that the first principal components (PC1), which accounted for ancestral breeds, were uniformly distributed between the two ancestral breeds when evaluated on chromosomes 16, 25, and 29. Because the average Brahman breed proportion is expected to be 5/8 (not 5/5), a uniform distribution of PC1 for Brangus between the two ancestral breeds would suggest equal ancestral genomic proportions. Hence, their results were an indication of significant deviates in breed compositions of Brangus toward Brahman on these three chromosomes. A principal analysis is a popular feature-reducing technique for analyzing large, high-dimension data, yet ignoring the detailed information of individual features, which were local ancestral contributions on specific chromosomal regions. In the present study, we directly evaluated breed compositions by chromosomes. Our results showed that only chromosomes 5 and 20 had less than 62.5% Angus breed proportions on average, which agreed with Paim T. D. P. et al. (2020). But all the other chromosomes (including 16, 25, and 29) had an average Angus CBC greater than 62.5%, meaning they deviated toward Angus instead. These differences could also arise from sampling biases, because the two ancestral breeds used by Paim T. D. P. et al. (2020) were small. There was a high chance that the sampled allelic frequencies may deviate substantially from the actual allelic frequencies. Furthermore, because the variance of an allelic frequency is inversely proportional to the population size (Falconer and Mackay, 1996), the estimated allelic frequencies for ancestral breeds were also subject to large deviations in small samples.

Despite the large chromosome-by-chromosome variability for ancestral breed proportions, GBC estimated by the mean of the average CBC across the 29 chromosomes per animal (denoted by C_GBC) agreed roughly with the average GBC in the 3,605 Brangus cattle. The mean (standard deviation) of Angus C_GBC for the 3,605 Brangus cattle was 70.9% (6.7%). The mean (standard deviation) of Brahman C-GBC was 29.1% (6.7%). The correlation between C_GBC and GBC for the 3,605 Brangus animals was 0.99 (See Supplementary Figure S1A). Within each subpopulation, the average CBC approximately agreed to (or slightly larger than) the corresponding GBC on average. For example, in the two clusters obtained by the K-means clustering with K = 2, the cluster for the mixed UB animals had, on average, 82.5% Angus and 17.5% Brahman breed compositions, and the cluster for non-UB animals had, on average, 68.2% Angus and 31.8% Brahman breed compositions.

Segmental-estimated breed compositions

Ancestral breed compositions were evaluated on 1-Mb, 5-Mb, and 10-Mb windows on each chromosome (Table 5). Like CBC, the

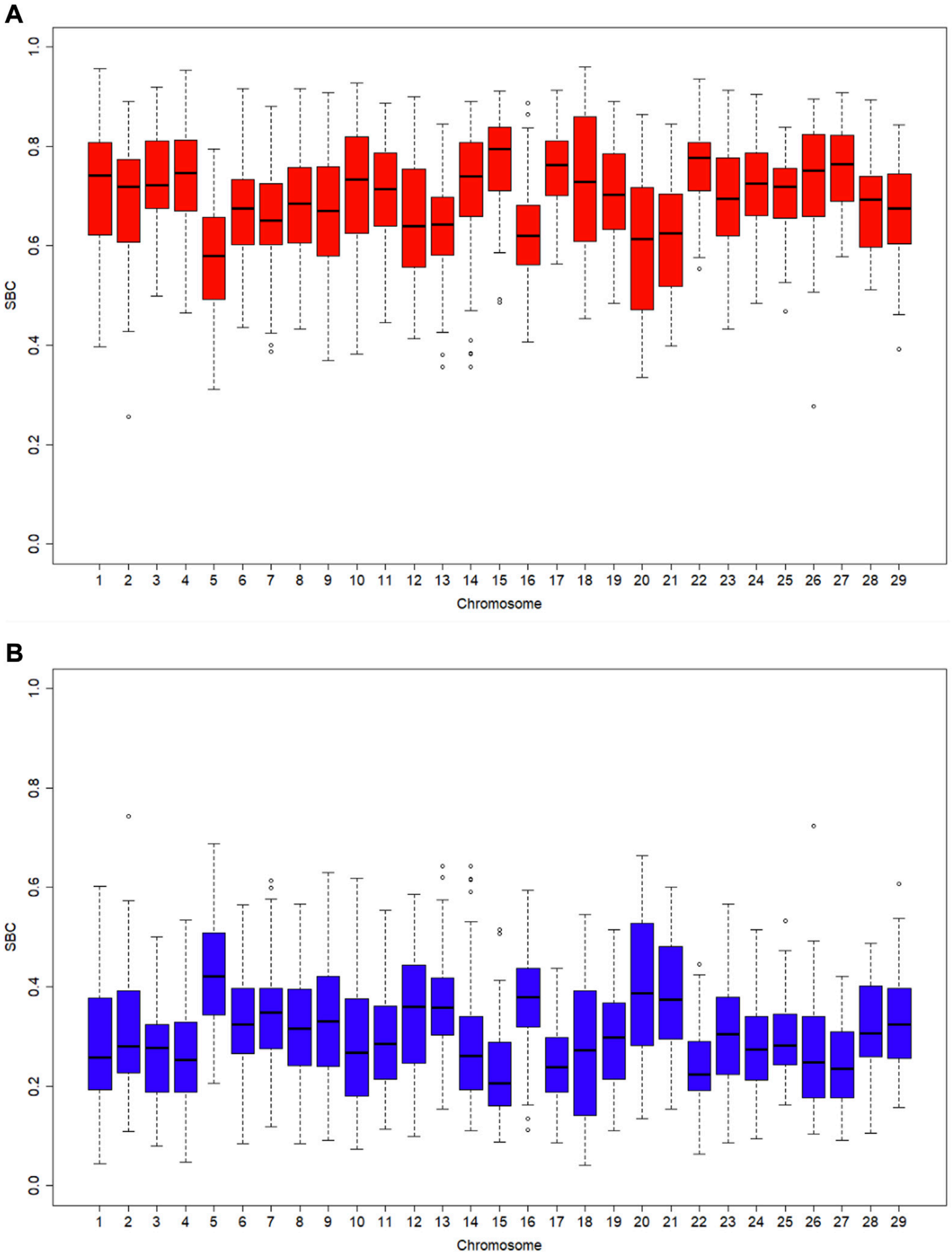


FIGURE 2
Boxplots of segmental-estimated breed compositions (SBC) for 3,605 Brangus cattle with the window size being 1 Mb on each chromosome: **(A)** Angus breed proportion; **(B)** Brahman breed proportion.

mean and range of the estimated SBC per segment varied substantially with chromosomes (Figure 2). Regardless of the window sizes, all the chromosomes, except 5 and 20, had the mean Angus SBC per chromosome exceeding 62.5%, and only chromosomes 5 and 20 had greater than 37.5% Brahman SBC. The average Angus SBC was the largest on chromosome 15 (76.9%–81.4%) and the smallest on chromosome 5 (57.6%–59.7%) (Table 5). Complementary to the Angus SBC, the Brahman SBC was the largest on chromosome 5 (40.3%–42.4%) and the smallest on chromosome 15 (18.6%–23.1%) (Table 5). Nevertheless, the segment with the maximum Angus proportion appeared on chromosome 18 (94.0%–95.9%), although chromosome 15 had the largest average Angus proportion. The maximum Brahman proportion segment was found on chromosome 2 (74.4%) when evaluated on 1 M windows and chromosome 5 (62.2%–64.4%) on 5-Mb and 10-Mb windows. The overall average of Angus SBC across the 29 chromosomes in the 3,605 Brangus cattle were 69.2% (1-Mb windows), 72.3% (5-Mb windows), and 72.0% (10-Mb windows), respectively. The overall

10.3389/fgene.2023.1080279

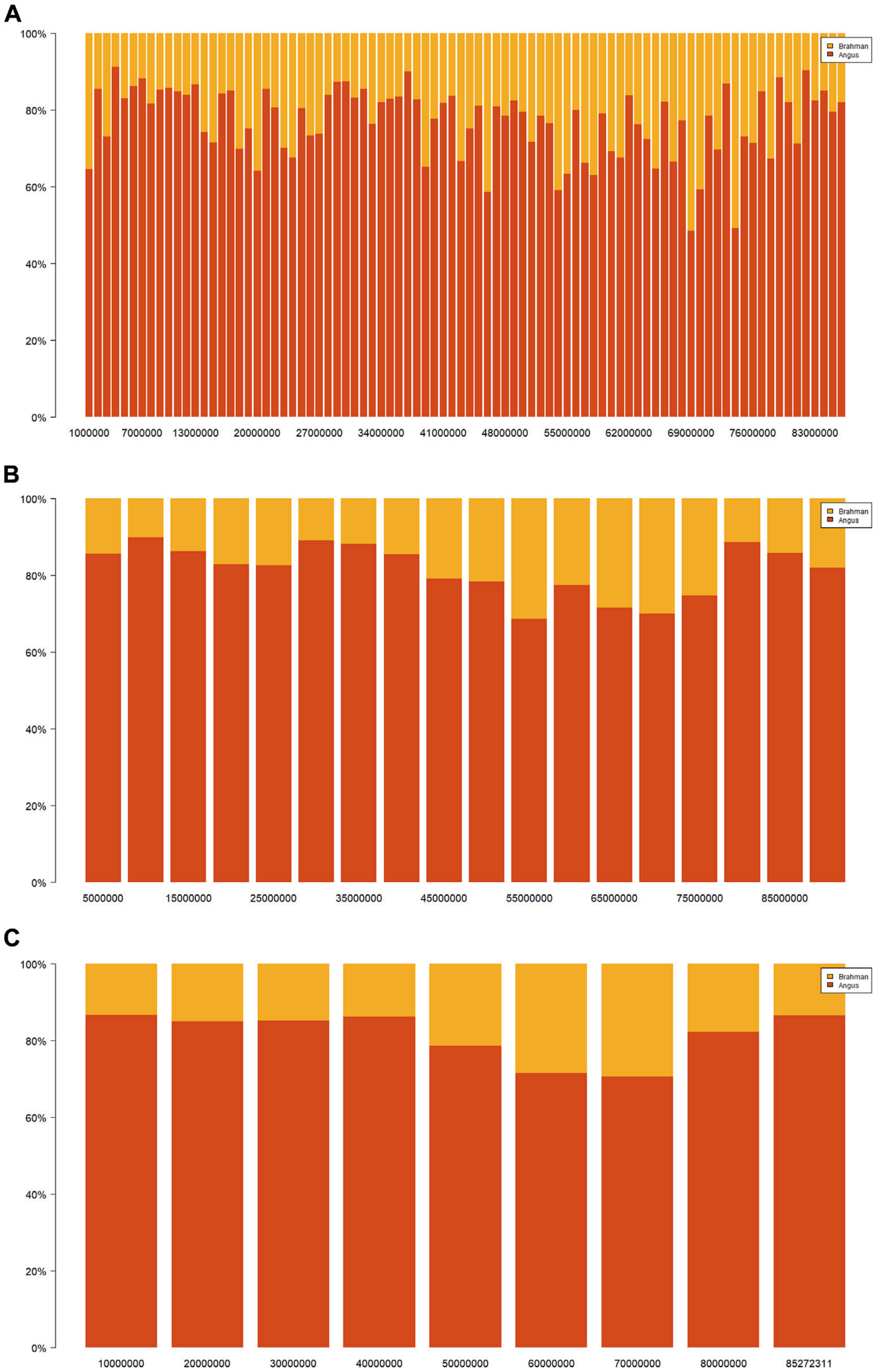


FIGURE 3
Distributions of Angus (red) or Brahman (yellow) breed proportions by varying window sizes on chromosome 15 in 3,605 Brangus cattle: **(A)** 1 Mb, **(B)** 5 Mb, and **(C)** 10 Mb. The black dashed line represents the population mean of the estimated genomic-estimated breed compositions (GBC), and the blue dashed line represents the population mean of the chromosomal-estimated breed compositions (CBC).

average of Brahman SBC across the 29 chromosomes was 30.8% (1-Mb windows), 27.7% (5-Mb windows), and 28.0% (10-Mb windows), respectively. These overall averages of genomic breed compositions by averaging Angus and Brahman SBC across all 29 chromosomes per animal (denoted by S_GBC) roughly agreed with the average Angus and Brahman CBC (70.9% Angus *versus* 29.1% Brahman; Table 4) and the average GBC (70.2% Angus *versus* 29.8% Brahman; Table 3). The correlation between Angus S_GBC and GBC for the 3,605 Brangus animals was greater than 0.99, regardless of the window sizes (See Supplementary Figures S1B–D). Overall, the variation was larger when evaluated on a smaller interval than a larger interval (Figure 3). The Angus SBC on chromosome 15 varied between 68.6% and 89.7% when assessed with a 5-Mb window size (Table 5). The range became smaller (70.5%–86.6%) when evaluated with a larger window size (i.e., 10-Mb), and the range became larger (48.6%–91.1%) when assessed with a smaller window size (e.g., 1-Mb). Similar trends generally held on all the chromosomes (Table 5). The Brahman SBC showed precisely opposite trends.

Ancestral breed proportions per segment varied considerably on each chromosome (Supplementary Figure S2). For example, the Brahman SBC on chromosome one were 4.4%–60.2% (1-Mb windows), 4.6%–47.8% (5-Mb windows), and 7.4%–46.6% (10-Mb windows), respectively (Table 5). There were segments with high Angus proportions and high Brahman proportions, respectively, on the chromosomes. For example, the average Angus SBC in 1–5 Mb and 6–10 Mb segments on chromosome 15 were 85.5% and 89.7%, respectively, which were substantially higher than the CBC for that chromosome. The average Angus SBC computed in the 40–75 Mb segment was between 68.6% and 79.1%, still higher than the officially expected Angus proportion (62.5%). Approximately 95% of the 2,522 1-Mb chromosomal segments had, on average, between 45.0% and 90.0% Angus breed proportions and between 10.0% and 65.0% Brahman breed proportions (Table 6). Likewise, approximately 90% of the total 5-Mb (or 10-Mb) chromosomal segments had between 50.0% and 90.0% Angus breed proportions and between 10.0% and 65.0% Brahman breed proportions (Table 6).

The variability of ancestral breed compositions per chromosome or segment could be a direct effect of selection. For example, Goszczynski et al. (2017) showed increased indicine haplotypes in the bovine leucocyte antigen region of Brangus cattle raised in Argentina, potentially due to selection for adaptation to the environments. In a brief search of bovine QTL in the QTLdb database, we found two postnatal growth traits and one body mass QTL reported on these two chromosomes, which were located on chromosome 15 at 61.6 Mb (body weight gain) and 17.0 Mb (birth body weight) (Snelling et al., 2010), and chromosome 17 at 12.0 Mb (average daily gain) (Rolf et al., 2012). These two chromosomes had the highest Angus CBC (chromosome 15: 79.84%; chromosome 17: 79.21%). These QTL were also located in regions with high Angus proportions. There are QTLs associated with disease resistance on chromosome 5 (Machado et al., 2010), which had the highest Braham CBC (43.14%). There were also health-related QTLs on chromosome 5, which included infectious bovine keratoconjunctivitis susceptibility (Kizilkaya et al., 2013), cold tolerance (Howard et al., 2014), and immune capacity (Leach et al., 2010), to list a few of them. All these QTLs were located in regions with high Brahman proportions in chromosome 5. There was a pleiotropic QTL affecting birth, yearling, and mature weights on chromosome 20 at 7–8 Mb (Weng et al., 2016), which was a region with a high Angus proportion. However, this chromosome had a high Brahman proportion (40.2%) in our study and Paim T. D. P. et al. (2020). The presence of Brahman favorable alleles in chromosomal regions with high Angus breed proportions, or *vice versa*, exemplifies the successful complementary of favorable alleles for the traits of interest from both ancestral breeds.

Gene set enrichment analysis

Gene set enrichment analysis was conducted with functional genes and QTL extracted on chromosome segments with high (≥75%) Angus and high (≥50%) Brahman breed proportions, respectively, evaluated on 1-Mb chromosomal intervals (Figure 4). We adopted these two cutoff threshold values because each represented an equal (12.5%) upward deviation from their expected values. We computed SBC on 1-Mb intervals for the subsequent gene set enrichment analysis because we rendered it an appropriate window size to capture gene regions of interest. Based on the *Bos taurus* UMD 3.1.1 assembly, the genome size of domestic cattle is approximately 2,670 Mb, which contains 26,815 genes (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bos_taurus/105/#FeatureCountsStats). Hence, the average spacing between neighboring genes is around 0.1 Mb. In contrast, the chromosome regions defined on 5-Mb or 10-Mb intervals could be too broad. Nevertheless, SNPs on 1 Mb intervals had higher LD than those on 5-Mb or 10-Mb intervals. The average R^2 LD per segment on each chromosome ranged from 0.11 to 0.16 in 1-Mb intervals, from 0.06 to 0.10 in 5-Mb intervals, and from 0.05 to 0.08 in 10-Mb intervals (Table 2). When computing the likelihood, we noted that SNPs with higher LD tended to give more weight to these highly linked SNPs or their regions.

There were 852 segments with Angus SBC ≥75% and 169 segments with Brahman SBC ≥50%, which accounted for 33.8% and 6.7%, respectively, of the 2,522 1-Mb chromosomal segments on the genome (Table 7.). The total length of chromosomal segments with ≥ 75% Angus proportions was the longest on chromosome one and the shortest on chromosomes 5 and 13. Relatively speaking, chromosome 15 had the largest percentage (64.5%) of high Angus breed proportion segments, and Chromosome five had the least percentage (5.8%) of total high Angus breed proportion segments. Twenty-five chromosomes had more than 20% of the chromosome length as high (≥75%) Angus proportion regions. In contrast, only two chromosomes (5 and 20) had more than 20% of their total length as high (≥50%) Brahman breed proportions. Adjacent high Angus proportion segments often joined together to form larger blocks, presenting almost in all 29 chromosomes (Figure 4A). In contrast, high Brahman proportion segments were relatively sparse and isolated, though some formed small blocks, and they were present only on seven (2, 5, 6, 9, 10, 20, and 21) chromosomes (Figure 4B). These results agreed with a previous report yet with a different approach. Using a chromosome painting approach based

TABLE 6 Distributions of segmental-estimated Angus and Brahman breed compositions (Angus-SBC and Brahman-SBC) with the segments defined by 1 MB, 5 Mb, and 10 Mb intervals, respectively.

Range	Angus-SBC						Brahman-SBC					
	1 Mb		5 Mb		10 Mb		1 Mb		5 Mb		10 Mb	
	n	n%	n	n%	n	n%	n	n%	n	n%	n	n%
1	0	0	0	0	0	0	0	0	0	0	0	0
(1, 0.95]	5	0.2	2	0.4	0	0	0	0	0	0	0	0
(0.95, 0.9]	38	1.5	7	1.3	4	1.5	0	0	0	0	0	0
(0.9, 0.85]	139	5.5	62	11.9	23	8.6	0	0	0	0	0	0
(0.85, 0.8]	301	11.9	69	13.3	39	14.5	0	0	0	0	0	0
(0.8, 0.75]	369	14.6	78	15.0	41	15.2	0	0	0	0	0	0
(0.75, 0.7]	395	15.7	76	14.6	45	16.7	2	0.1	0	0	0	0
(0.7, 0.65]	377	14.9	89	17.1	47	17.5	4	0.2	0	0	0	0
(0.65, 0.6]	331	13.1	67	12.9	33	12.3	18	0.7	1	0.2	1	0.4
(0.6, 0.55]	238	9.4	32	6.2	18	6.7	54	2.1	5	1.0	0	0
(0.55, 0.5]	160	6.3	25	4.8	14	5.2	91	3.6	6	1.2	4	1.5
(0.5, 0.45]	91	3.6	6	1.2	4	1.5	160	6.3	25	4.8	14	5.2
(0.45, 0.4]	54	2.1	5	1.0	0	0	238	9.4	32	6.2	18	6.7
(0.4, 0.35]	18	0.7	1	0.2	1	0.4	331	13.1	67	12.9	33	12.3
(0.35, 0.3]	4	0.2	0	0	0	0	377	14.9	89	17.1	47	17.5
(0.3, 0.25]	2	0.1	0	0	0	0	395	15.7	76	14.6	45	16.7
(0.25, 0.2]	0	0	0	0	0	0	369	14.6	78	15.0	41	15.2
(0.2, 0.15]	0	0	0	0	0	0	301	11.9	69	13.3	39	14.5
(0.15, 0.1]	0	0	0	0	0	0	139	5.5	62	11.9	23	8.6
(0.1, 0.05]	0	0	0	0	0	0	38	1.5	7	1.3	4	1.5
(0.05, 0]	0	0	0	0	0	0	5	0.2	2	0.4	0	0
SUM	2,522	100	519	100	269	100	2,522	100	519	100	269	100

on a copying model (Lawson et al., 2012), Paim T. D. P. et al. (2020) showed that chromosomal regions with high Angus breed proportions were prevalent on chromosomes, often large blocks, yet chromosome segments with high Brahman breed proportion were fewer and isolated. A follow-up study on selection signatures in Brangus revealed that the majority of selection signatures in Brangus cattle came from Angus (Paim T. D. P. et al., 2020). The copying model related the patterns of LD across chromosomes to the underlying recombination process and used a hidden Markov method to reconstruct a sampled haplotype. In the present study, we directly estimated breed compositions on each chromosome flanked by varying window sizes based on a mixture model. The admixture coefficients inferred from the admixture model can be probabilistically interpreted as reflecting that as identity-in-state (Jannink and Wu, 2003). Yet, when confined to long chromosomal chunks, it approximated the probability of identity by descent (Browning, 2008). Hence, though using a different approach, we came to similar findings.

In theory, crossing breaks and shuffles chromosomes randomly over time when directional selection is absent, which reflects genetic drift. However, with selection, it increases and even fixes favorable alleles. Meanwhile, it also changes the allelic frequencies of genes in LD with the genes under selection due to genomic hitchhiking, also known as genetic draft (Smith and Haigh, 1974; Ma et al., 2019). In other words, genomic hitchhiking occurs when a polymorphic locus is in LD with a second locus undergoing a selective sweep. As a result, the linked allele will also increase in frequency, in some cases, until it becomes fixed in the population. Overall, genomic hitchhiking reduces genetic variation and leaves footprints across the genome known as the signatures of selection (Sabeti et al., 2007; Singh et al., 2020). The many high Angus breed proportion regions reflected the presence of multiple favorable alleles scattered across the chromosomes. We observed the prevalence of large blocks with high Angus breed proportions (Figure 4A), possibly

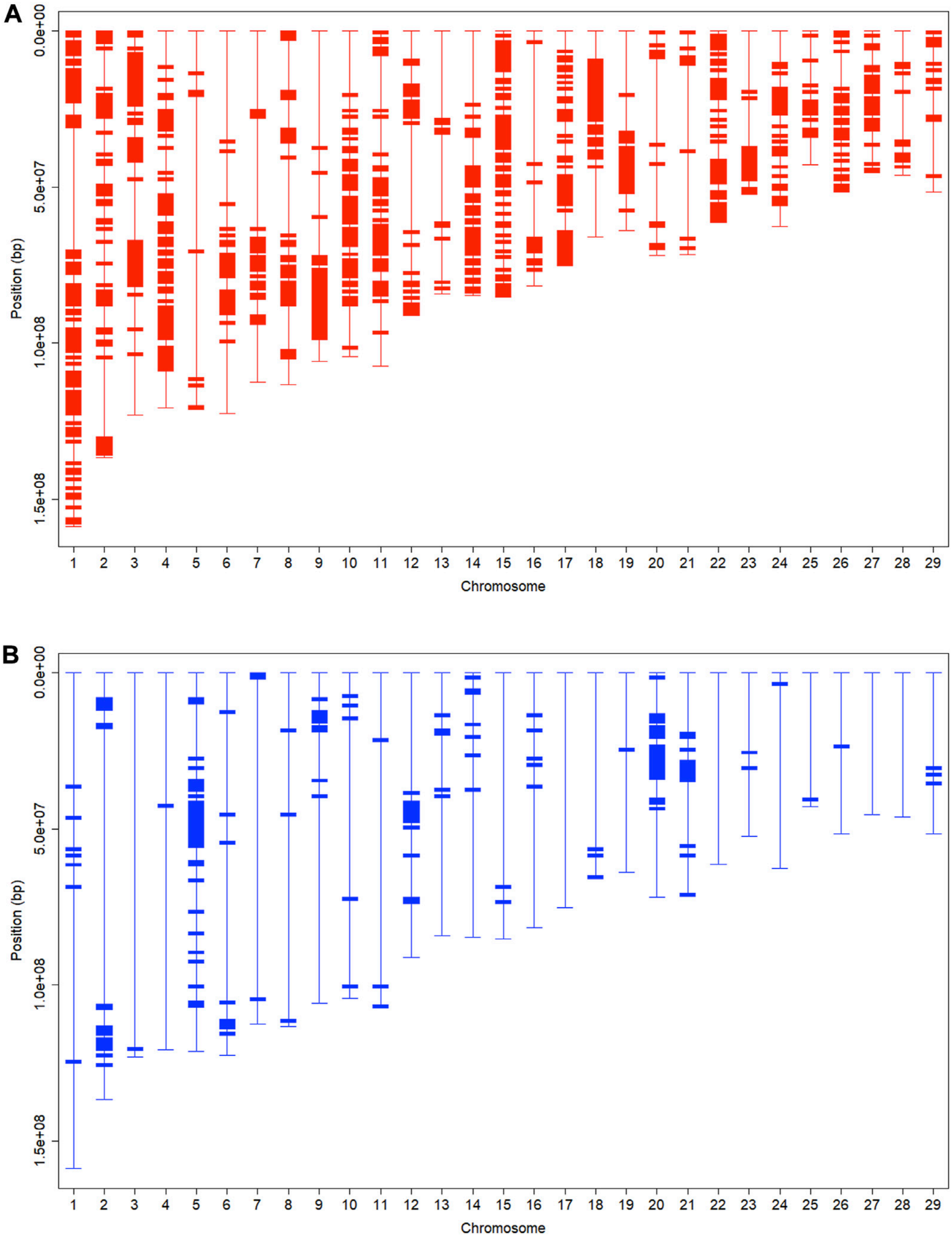


FIGURE 4
Chromosomal fragments with: **(A)** high (≥75%) Angus breed proportion (upper), and **(B)** high (≥50%) Brahman breed proportions (bottom), evaluated with a 1 Mb window size on 29 chromosomes in 3,605 Brangus cattle.

resulting from the genomic hitchhiking effects when selecting Brangus for Angus favorable traits. This is equivalent to saying that genomic hitchhiking effects were strong around the genomic regions with Angus favorable alleles, leading to the presence of large blocks with high Angus breed proportions. In contrast, genomic hitchhiking effects were weak around Brahman favorable alleles because chromosomal blocks with high Brahman breed proportions were relatively few, isolated, and small in size.

There were 9,025 genes on the chromosomal segments with Angus SBC ≥75% (Supplementary Table S1) and 1,877 genes on the chromosomal segments with Brahman SBC ≥50% (Supplementary Table S2). Many genes in high (≥75%) Angus regions are responsible for biological processes related to animal development, such as regulation of biological processes, anatomical structure development, anatomical structural morphogenesis, animal organ development, and skeletal system development, and, in KEGG, related to

TABLE 7 Summary of chromosomal segments with high Angus breed proportions (Angus SBC ≥ 75%) and high Brahman breed proportions (Brahman SBC ≥ 50%), respectively.

Chromosome	Total length	Angus SBC ≥ 75%			Brahman SBC ≥ 50%		
		N	Length	Percent (%)	N	Length	Percent (%)
1	158855123	79	79,000,000	49.7	7	7,000,000	4.4
2	136769635	46	46,000,000	33.6	17	17,000,000	12.4
3	123148964	50	50,000,000	40.6	1	1,000,000	0.8
4	120635950	60	60,000,000	49.7	1	1,000,000	0.8
5	121183174	7	7,000,000	5.8	34	34,000,000	28.1
6	122509741	25	25,000,000	20.4	8	8,000,000	6.5
7	112628884	23	23,000,000	20.4	3	3,000,000	2.7
8	113367096	32	32,000,000	28.2	3	3,000,000	2.6
9	105695468	29	29,000,000	27.4	9	9,000,000	8.5
10	104301732	46	46,000,000	44.1	5	5,000,000	4.8
11	107274061	42	42,000,000	39.2	3	3,000,000	2.8
12	91131021	25	25,000,000	27.4	12	12,000,000	13.2
13	84229982	9	9,000,000	10.7	5	5,000,000	5.9
14	84628243	37	37,000,000	43.7	7	7,000,000	8.3
15	85272311	55	55,000,000	64.5	2	2,000,000	2.3
16	81701834	11	11,000,000	13.5	5	5,000,000	6.1
17	75132928	41	41,000,000	54.6	0	0	0.0
18	65999195	30	30,000,000	45.5	3	3,000,000	4.5
19	64007021	21	21,000,000	32.8	1	1,000,000	1.6
20	71992748	11	11,000,000	15.3	22	22,000,000	30.6
21	71573501	8	8,000,000	11.2	13	13,000,000	18.2
22	61379134	38	38,000,000	61.9	0	0	0.0
23	52465632	16	16,000,000	30.5	2	2,000,000	3.8
24	62685898	25	25,000,000	39.9	1	1,000,000	1.6
25	42851121	13	13,000,000	30.3	1	1,000,000	2.3
26	51663776	26	26,000,000	50.3	1	1,000,000	1.9
27	45388171	25	25,000,000	55.1	0	0	0.0
28	46248750	10	10,000,000	21.6	0	0	0.0
29	51484561	12	12,000,000	23.3	3	3,000,000	5.8

hormone regulation (such as the Estrogen signaling pathway). For example, the system development (GO:0048731) category's related child terms include system development, such as central nervous, respiratory, and endocrine. We also found many genes associated with carcass and meat quality traits, such as *PPP1R3B* (Edwards et al., 2003; Cinar et al., 2012), *ASXL1* (Grigoletto et al., 2020), *DNMT3B* (Liu et al., 2012), and *TMEM68* (Lindholm-Perry et al., 2012; Terakado et al., 2018; Edea et al., 2020), just to list a few. Our gene list also included LEP on chromosome four and PLAG1 on chromosome 14. Paim T. D. P. et al. (2020) previously found these two genes in high Angus regions. The LEP gene is expressed in adipose tissue and codes for leptin, a hormone known to regulate feed intake and energy balance in mammals (Woronuk et al., 2012). This gene is associated with marbling, fat thickness, rib eye area, and feed intake in several beef cattle breeds (Souza et al., 2010; Woronuk et al., 2012; Kononoff et al., 2017). Leptin is an essential gene for puberty onset (Williams et al., 2002). This gene could be inherited from Angus ancestors, or its frequency was increased by the selection of Brangus for early puberty since breed formation because *Bos indicus* heifers have challenges achieving puberty early in life (Sartori et al., 2010; Fortes et al., 2012). PLAG1 is involved in regulating stature and weight (Littlejohn et al., 2012; Pryce et al., 2012; Song et al., 2016). This gene is associated with yearling weight in Australian Tropical Composite breeds (Porto-Neto et al., 2014). There is still another gene, XKR4, which is close to PLAG1. The XKR4 gene is associated with subcutaneous rump fat thickness, scrotal circumference, serum concentration of prolactin, and sexual precocity (Fortes et al., 2012; Porto Neto et al., 2012; Bastin et al., 2014; Takada et al., 2018). The genes in the high (≥50%) Brahman regions are primarily responsible for biosynthetic-related biological processes (e.g., cytolysis), molecular biological functions related to enzyme activity (e.g., lysozyme activity, peptidoglycan muralytic activity, hydrolase activity, and serine-type endopeptidase activity), and diseases and immunity (e.g., MHC class II protein complex, MHC protein complex, type I diabetes mellitus, allograft rejection, graft-versus-host disease, autoimmune thyroid disease, and pathogenic *Escherichia coli* infection). For example, peptidoglycan muralytic activity (GO:0061783), which contributes to the degradation of peptidoglycan, is a major structural component of bacterial cell walls (Nelson et al., 2012); Another example is the MHC Class II protein complex (GO:0042613). MHC is involved in the immune process and plays the role of transmitting antigens (Rodgers and Cook, 2005). Still, the gene functionalities in chromosomal regions with high Angus breed proportions were diverse, possibly due to genomic hitchhiking of genes linked to the favorable alleles under selection. For example, there was a category of genes called “cellular response to stress”. Cells respond to stress in various ways, from activating pathways that promote survival to eliciting programmed cell death that eliminates damaged cells. Cell death research has attracted much attention in the last 2 decades also because of its relevance to development, degenerative diseases, and cancer in human (Fulda et al., 2010). Overall, the Brangus cattle have successfully combined the favorable traits of the two highly successful parent breeds.

QTLs were extracted from the chromosomal regions with the top 1% highest Angus (Angus SBC ≥91.1%) and Brahman (Braham SBC ≥59.4%) breed proportions, respectively (Supplementary Table S3). The QTLs in the top 1% Angus regions are related to meat, carcass, production, and reproduction. This list included, for example, birth weight QTL (chromosomes three and 8), weaning weight QTL (chromosomes 1, 3, 15, 18, and 23), yearling weight QTL (chromosomes 3 and 10), mature body weight QTL (chromosomes 3, 4, 10, 17, and 18), carcass weight QTL (chromosomes three and 8), marbling score QTL (chromosomes 15, 18, and 22), fat thickness QTL (chromosomes 3, 4, 17, 18, and 23), calving ease QTL (chromosomes 3, 8, 10, and 17), and QTL for Longissimus muscle area (chromosomes 3, 8, 17, and 18). Some QTL have pleiotropic effects. The QTL in the regions with 1% highest Brahman proportions are related to health, such as bovine respiratory disease susceptibility (chromosome 2: 117.3–134.4 Mb), insulin-like growth factor 1 level on (chromosome 5: 40.3–59.7 Mb; chromosome 14: 26.5–27.0 Mb), and blood cortisol level (chromosome 16: 29.7 Mb).

Conclusions

We have revisited the genomic breed compositions in 3,605 Brangus cattle from three perspectives, genome-wise, per chromosome, and per chromosome segment. The K-means clustering analysis revealed population stratification. The B-1 cluster consisted of ½ UB (i.e., first generation from crossing Brangus with Angus), with 81.6% Angus genomic lineage. The non-UB Brangus animals (B-2), on average, were 67.4% Angus and 32.6% Brahman. It was likely that selecting Brangus for traits where Angus had advantages led to an approximately 5% average deviation in the estimated GBC toward Angus lineage. Further deviation of Brangus breed compositions from the officially expected ancestral breed ratios is worth paying attention to because it could diminish the complementarity of Brahman and Angus over time.

The ancestral breed proportions varied substantially by chromosomes and by chromosomal segments, likely shaped by cross-breeding and subsequent inter-se mating and selection over time. There were strong hitchhiking effects on genomic regions where favorable Angus alleles resided. The functions of genes identified in the chromosomal regions with high Angus proportions were diverse, yet many were responsible for biological processes related to growth and body development. The genes identified in the regions with high Brahman compositions were primarily responsible for biosynthetic-related processes, molecular biological functions related to enzyme activity, diseases, and immunity. The co-presence of segments of high-Angus and high-Brahman lineage, respectively, on the same chromosomes, exemplified the successful complementary of favorable alleles from both ancestral breeds.

While genomic-estimated breed compositions depicted an overall picture of the “mosaic” genome of animals attributable to their ancestors, local ancestry genomic compositions were visualized through chromosomal- and segmental-estimated breed compositions. Precisely, the admixture coefficients based on the admixture model inferred the probability of alleles identical in state when estimating the ancestral breed compositions for individual animals. Nevertheless, they corresponded approximately to the probability of alleles identical by descent when confined to long chromosome chunks. We noted that alternative approaches existed, such as path analysis (Wu et al., 2020) and the breed-of-origin (BOA) approach (Calus et al., 2022), that could better handle the situation when ancestral breeds were correlated. In the present study, however, the correlation between Angus and Brahman was low (0.05–0.10). Hence, we used the admixture model because is theoretically instructive and easily implemented in practice.

Data availability statement

All data were taken from the databases of the Neogen Global Laboratories. Example data and programs are available at: https://

doi.org/10.6084/m9.figshare.22303009.v1. Inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the animal study because All data were taken from the databases of the Neogen Global Laboratories. No animal experiments were involved.

Author contributions

X-LW and JH conceived and planned this study in discussion with WG, CS, RT, and SB. ZL extracted the data and carried out the data analyses, with assistance provided by FY, SY, ZG, and WC. ZL drafted this manuscript, revised by X-LW, JH, and WG. All authors proofread and approved this manuscript.

Funding

This research was funded by the Special Fund for the Construction of Innovative Provinces in Hunan (Grant number 2021NK1009), and Hunan Provincial Natural Science Foundations of China (Grant numbers 2022JJ30286, 2021JJ40254).

Conflict of interest

CS, RT, and SB are employees of Neogen GeneSeek, an agri-genomics service provider. X-LW is an employee of the Council of Dairy Cattle Breeding (CDCB), a non-profit organization responsible for calculating and distributing genetic evaluations and genomic predictions and for managing genomic data storage.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1080279/full#supplementary-material

References

Akanno, E. C., Chen, L., Abo-Ismail, M. K., Crowley, J. J., Wang, Z., Li, C., et al. (2017). Genomic prediction of breed composition and heterosis effects in Angus, Charolais, and Hereford crosses using 50K genotypes. *Can. J. Anim. Sci.* 97 (3), 431–438. doi:10.1139/cjas-2016-0124

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bansal, V., and Libiger, O. (2015). Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinf* 16 (1), 4. doi:10.1186/s12859-014-0418-7

Bastin, B. C., Houser, A., Bagley, C. P., Ely, K. M., Payton, R. R., Saxton, A. M., et al. (2014). A polymorphism in XKR4 is significantly associated with serum prolactin concentrations in beef cows grazing tall fescueficantly associated with serum prolactin concentrations in beef cows grazing tall fescue. *Anim. Genet.* 45, 439–441. doi:10.1111/age.12134

Berry, D. P. (2019). Genomic information in livestock has multiple uses in precision breeding and management. *Livestock* 24 (1), 30–33. doi:10.12968/live.2019.24.1.30

Briggs, H. M., and Briggs, D. M. (1980). *Modern breeds of livestock*. 4th Ed. New York: Macmillan Publishing Co.

Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 178 (4), 2123–2132. doi:10.1534/genetics.107.084624

Calus, M. P. L., Henshall, J. M., Hawken, R., and Vandenplas, J. (2022). Estimation of dam line composition of 3-way crossbred animals using genomic information. *Genet. Sel. Evol.* 54, 44. doi:10.1186/s12711-022-00728-4

Cinar, M. U., Kayan, A., Uddin, M. J., Jonas, E., Tesfaye, D., Phatsara, C., et al. (2012). Association and expression quantitative trait loci (eQTL) analysis of porcine AMBP, GC and PPP1R3B genes with meat quality traits. *Mol. Biol. Rep.* 39, 4809–4821. doi:10.1007/s11033-011-1274-4

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

Edwards, D. B., Bates, R. O., and Osburn, W. N. (2003). Evaluation of Duroc-vs. Pietrain-sired pigs for carcass and meat quality measures. *J. Anim. Sci.* 81, 1895–1899. doi:10.2527/2003.8181895x

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. 4th ed. Essex, England: Longman Group Limited.

Fortes, M. R. S., Reverter, A., Hawken, R. J., Bolormaa, S., and Lehnert, S. A. (2012). Candidate genes associated with testicular development, sperm quality, and hormone levels of inhibin, luteinizing hormone, and insulin-like growth factor 1 in Brahman bulls. *Biol. Reprod.* 87 (3), 58–8. doi:10.1095/biolreprod.112.101089

Fulda, S., Gorman, A. M., Hori, O., and Samali, A. (2010). Cellular stress responses: Cell survival and cell death. *Int. J. Mol. Biol.* 2010, 214074. Article ID 214074. doi:10.1155/2010/214074

Funkhouser, S. A., Bates, R. O., Ernst, C. W., Newcomb, D., and Steibel, J. P. (2017). Estimation of genome-wide and locus-specific breed composition in pigsfic breed composition in pigs. *Transl. Anim. Sci.* 1, 36–44. doi:10.2527/tas2016.0003

Gobena, M., Elzo, M. A., and Mateescu, R. G. (2018). Population structure and genomic breed composition in an angus–brahman crossbred cattle population. *Front. Genet.* 9, 90. doi:10.3389/fgene.2018.00090

Goszczynski, D. E., Corbi-Botto, C. M., Durand, H. M., Rogberg-Mu~noz, A., Munilla, S., Peral-Garcia, P., et al. (2017). Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. *Animal* 12, 215–223. doi:10.1017/S1751731117001380

Grigoletto, L., Ferraz, J. B. S., Oliveira, H. R., Eler, J. P., Bussiman, F, O., Abreu Silva, B. C., et al. (2020). Genetic architecture of carcass and meat quality traits in Montana Tropical® composite beef cattle. *Front. Genet.* 11, 123. doi:10.3389/fgene.2020.00123

Hartigan, J. A., and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc. C* 28 (1), 100–108. doi:10.2307/2346830

He, J., Guo, Y., Xu, J., Li, H., Fuller, A., Tait, R. G., et al. (2018). Comparing SNP panels and statistical methods for estimating genomic breed composition of individual animals in ten cattle breeds. *BMC Genet.* 19 (1), 56. doi:10.1186/s12863-018-0654-3

Howard, J. T., Kachman, S. D., Snelling, W. M., Pollak, E. J., Ciobanu, D. C., Kuehn, L. A., et al. (2014). Beef cattle body temperature during climatic stress: A genome-wide association study. *Int. J. Biometeorol.* 58 (7), 1665–1672. doi:10.1007/s00484-013-0773-5

Hu, Z-L., Park, C. A., and Reecy, J. M. (2022). Bringing the animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services. *Nucleic Acids Res.* 50 (D1), D956–D961. doi:10.1093/nar/gkab1116

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31 (8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jannink, J.-L., and Wu, X.-L. (2003). Estimating allelic number and identity in state of QTLs in interconnected families. *Genet. Res.* 81 (2), 133–144. doi:10.1017/s0016672303006153

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44 (D1), D457–D462. doi:10.1093/nar/gkv1070

Kizilkaya, K., Tait, R. G., Garrick, D. J., Fernando, R. L., and Reecy, J. M. (2013). Genome-wide association study of infectious bovine keratoconjunctivitis in Angus cattle. *BMC Genet.* 14 (1), 23. doi:10.1186/1471-2156-14-23

Kononoff, P. J., Defoor, P. J., Engler, M. J., Swingle, R. S., Gleghorn, J. F., James, S. T., et al. (2017). Impacts of a leptin SNP on growth performance and carcass characters in finishing steers studied over timefinishing steers studied over time. *J. Anim. Sci.* 95, 194–200. doi:10.2527/jas2016.0926

Lawson, D. J., Hellenthal, G., Myers, S., Falush, D., and Zhang, F. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* 8, e1002453. doi:10.1371/journal.pgen.1002453

Leach, R. J., Craigmile, S. C., Knott, S. A., Williams, J. L., and Glass, E. J. (2010). Quantitative trait loci for variation in immune response to a Foot-and-Mouth Disease virus peptide. *BMC Genet.* 11 (1), 107. doi:10.1186/1471-2156-11-107

Li, Z., Wu, X.-L., Guo, W., He, J., Li, H., Rosa, G. J. M., et al. (2020). Estimation of genomic breed composition of individual animals in composite beef cattle. *Anim. Genet.* 51 (3), 457–460. doi:10.1111/age.12928

Lindholm-Perry, A. K., Kuehn, L. A., Smith, T. P. L., Ferrell, C. L., Jenkins, T. G., Freetly, H. C., et al. (2012). A region on BTA14 that includes the positional candidate genes LYPLA1, XKR4 and TMEM68 is associated with feed intake and growth phenotypes in cattle. *Anim. Genet.* 43, 216–219. doi:10.1111/j.1365-2052.2011.02232.x

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

Liu, X., Guo, X. Y., Xu, X. Z., Wu, M., Zhang, X., Li, Q., et al. (2012). Novel single nucleotide polymorphisms of the bovine methyltransferase 3b gene and their association with meat quality traits in beef cattle. *Genet. Mol. Res.* 11, 2569–2577. doi:10.4238/2012.June.29.1

Ma, L., Sonstegard, T. S., Cole, J. B., VanTassell, C. P., Wiggans, G. R., Crooker, B. A., et al. (2019). Genome changes due to artificial selection in U.S. Holstein cattle. *BMC Genomics* 20 (1), 128. doi:10.1186/s12864-019-5459-x

Machado, M. A., S Azevedo, A. L., Teodoro, R. L., Pires, M. A., Cd Peixoto, M. G., de Freitas, C., et al. (2010). Genome wide scan for quantitative trait loci affecting tick resistance in cattle (*Bos taurus* x *Bos indicus*). *BMC Genomics* 11 (1), 280. doi:10.1186/1471-2164-11-280

Merchant, S., Wood, D. E., and Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2, e675. doi:10.7717/peerj.675

Nelson, D. C., Schmelcher, M., Rodriguez-Rubio, L., Klumpp, J., Pritchard, D. G., Dong, S., et al. (2012). Endolysins as antimicrobials. *Adv. Virus Res.* 83, 299–365. doi:10.1016/B978-0-12-394438-2.00007-4

Nocedal, J., and Wright, S. J. (2006). *Numerical optimization*. New York, NY: Spinger.

Paim, T. D. P., Hay, E. H. A., Wilson, C., Thomas, M. G., Kuehn, L. A., Paiva, S. R., et al. (2020a). Dynamics of genomic architecture during composite breed development in cattle. *Anim. Genet.* 51, 224–234. doi:10.1111/age.12907

Paim, T. D. P., Hay, E. H. A., Wilson, C., Thomas, M. G., Kuehn, L. A., Paiva, S. R., et al. (2020b). Genomic breed composition of selection signatures in Brangus beef cattle. *Front. Genet.* 11 (710), 710. doi:10.3389/fgene.2020.00710

Porto Neto, L. R., Bunch, R. J., Harrison, B. E., and Barendse, W. (2012). Variation in the XKR4 gene was significantly associated with subcutaneous rump fat thickness in indicine and composite cattleficantly associated with subcutaneous rump fat thickness in indicine and composite cattle. *Anim. Genet.* 43, 785–789. doi:10.1111/j.1365-2052.2012.02330.x

Porto-Neto, L. R., Reverter, A., Prayaga, K. C., Chan, E. K. F., Johnston, D. J., Hawken, R. J., et al. (2014). The genetic architecture of climatic adaptation of tropical cattle. *PLoS One* 9, e0113284. doi:10.1371/journal.pone.0113284

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155 (2), 945–959. doi:10.1093/genetics/155.2.945

Pryce, J. E., Arias, J., Bowman, P. J., Davis, S. R., Macdonald, K. A., Waghorn, G. C., et al. (2012). Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. *J. Dairy Sci.* 95, 2108–2119. doi:10.3168/jds.2011-4628

Rodgers, J. R., and Cook, R. G. (2005). MHC class Ib molecules bridge innate and acquired immunity. *Nat. Rev. Immunol.* 5 (6), 459–471. doi:10.1038/nri1635

Rolf, M. M., Taylor, J. F., Schnabel, R. D., McKay, S. D., McClure, M. C., Northcutt, S. L., et al. (2012). Genome-wide association analysis for feed efficiency in Angus cattle. *Anim. Genet.* 43 (4), 367–374. doi:10.1111/j.1365-2052.2011.02273.x

Sabeti, P. C., Varilly, P., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164), 913–918. doi:10.1038/nature06250

Sartori, R., Bastos, M., Baruselli, P., Gimenes, L., Ereno, R., and Barros, C. (2010). Physiological differences and implications to reproductive management of *Bos taurus* and *Bos indicus* cattle in a tropical environment. *Soc. Reprod. Fertil. Suppl.* 67, 357–375. doi:10.7313/upo9781907284991.028

Singh, A., Mehrotra, A., Gondro, C., Romero, A. R., Pandey, A. K., Karthikeyan, A., et al. (2020). Signatures of selection in composite vrindavani cattle of India. *Front. Genet.* 11, 589496. doi:10.3389/fgene.2020.589496

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43 (W1), W589–W598. doi:10.1093/nar/gkv350

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23 (1), 391–403. doi:10.1017/S0016672308009579

Snelling, W. M., Allan, M. F., Keele, J. W., Kuehn, L. A., McDaneld, T., Smith, T. P. L., et al. (2010). Genome-wide association study of growth in crossbred beef cattle. *J. Anim. Sci.* 88 (3), 837–848. doi:10.2527/jas.2009-2257

Song, Y., Xu, L., Chen, Y., Zhang, L., Gao, H., Zhu, B., et al. (2016). Genome-wide association study reveals the PLAG1 gene for knuckle, biceps and shank weight in Simmental beef cattle. *PLoS One* 11, e0168316. doi:10.1371/journal.pone.0168316

Souza, F. R. P. P., Mercadante, M. E. Z. Z., Fonseca, L. F. S. S., Ferreira, L. M. S. S., Regatieri, I. C., Ayres, D. R., et al. (2010). Assessment of DGAT1 and LEP gene polymorphisms in three Nelore (*Bos indicus*) lines selected for growth and their relationship with growth and carcass traits. *J. Anim. Sci.* 88, 435–441. doi:10.2527/jas.2009-2174

Takada, L., Barbero, M. M. D., Oliveira, H. N., de Camargo, G. M. F., Fernandes Júnior, G. A., Aspilcueta-Borquis, R. R., et al. (2018). Genomic association for sexual precocity in beef heifers using pre-selection of genes and haplotype reconstruction. *PLoS One* 13, e0190197. doi:10.1371/journal.pone.0190197

Terakado, P. N., Costa, R. B., de Camargo, G. M. F., Irano, N., Bresolin, T., Takada, L., et al. (2018). Genome-wide association study for growth traits in Nelore cattle. *Animal* 12, 1358–1362. doi:10.1017/S1751731117003068

Wang, Y., Wu, X.-L., Li, Z., Bao, Z., Tait, R. G., Jr., Bauck, S., et al. (2020). Estimation of genomic breed composition for purebred and crossbred animals using sparsely regularized admixture models. *Front. Genet.* 11, 576. doi:10.3389/fgene.2020.00576

Weng, Z., Su, H., Saatchi, M., Lee, J., Thomas, M. G., Dunkelberger, J. R., et al. (2016). Genome-wide association study of growth and body composition traits in Brangus beef cattle. *Livest. Sci.* 183, 4–11. doi:10.1016/j.livsci.2015.11.011

Williams, G. L., Amstalden, M., Garcia, M. R., Stanko, R. L., Nizielski, S. E., Morrison, C. D., et al. (2002). Leptin and its role in the central regulation of reproduction in cattle. *Domest. Anim. Endocrinol.* 23, 339–349. doi:10.1016/S0739-7240(02)00169-8

Woronuk, G. N., Marquess, F. L., James, S. T., Palmer, J., Berryere, T., Deobald, H., et al. (2012). Association of leptin genotypes with beef cattle characteristics. *Anim. Genet.* 43, 608–610. doi:10.1111/j.1365-2052.2012.02320.x

Wu, X.-L., Li, Z., Wang, Y., He, J., Rosa, G. J. M., Ferretti, R., et al. (2020). A causality perspective of genomic breed composition for composite animals. *Front. Genet.* 11 (1369), 546052. doi:10.3389/fgene.2020.546052

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for comparing biological themes among gene clusters. *OMICS* 16 (5), 284–287. doi:10.1089/omi.2011.0118

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fgene.2023.1168150

Genome-wide association study for carcass weight in pasture-finished beef cattle in Hawai’i

Mandeep Adhikari[1], Michael B. Kantar[2], Ryan J. Longman[3,4], C. N. Lee[5], Melelani Oshiro[5], Kyle Caires[5] and Yanghua He[1,5]*

[1]Department of Molecular Biosciences and Bioengineering, University of Hawai’i at Mānoa, Honolulu, HI, United States, [2]Department of Tropical Plant and Soil Sciences, University of Hawai’i at Mānoa, Honolulu, HI, United States, [3]East West Center, Honolulu, HI, United States, [4]Department of Geography and Environment, University of Hawai’i at Mānoa, Honolulu, HI, United States, [5]Department of Human Nutrition, Food, and Animal Sciences, University of Hawai’i at Mānoa, Honolulu, HI, United States

OPEN ACCESS

EDITED BY
Li Ma,
University of Maryland, United States

REVIEWED BY
Dailu Guan,
University of California, Davis, United States
Xiaojuan Fan,
National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH), United States

*CORRESPONDENCE
Yanghua He,
✉ yanghua.he@hawaii.edu

RECEIVED 17 February 2023
ACCEPTED 11 April 2023
PUBLISHED 09 May 2023

CITATION
Adhikari M, Kantar MB, Longman RJ, Lee CN, Oshiro M, Caires K and He Y (2023), Genome-wide association study for carcass weight in pasture-finished beef cattle in Hawai’i.
Front. Genet. 14:1168150.
doi: 10.3389/fgene.2023.1168150

Introduction: Genome-wide association studies (GWAS) have identified genetic markers for cattle production and reproduction traits. Several publications have reported Single Nucleotide Polymorphisms (SNPs) for carcass-related traits in cattle, but these studies were rarely conducted in pasture-finished beef cattle. Hawai’i, however, has a diverse climate, and 100% of its beef cattle are pasture-fed.

Methods: Blood samples were collected from 400 cattle raised in Hawai’i islands at the commercial harvest facility. Genomic DNA was isolated, and 352 high-quality samples were genotyped using the Neogen GGP Bovine 100 K BeadChip. SNPs that did not meet the quality control standards were removed using PLINK 1.9, and 85 k high-quality SNPs from 351 cattle were used for association mapping with carcass weight using GAPIT (Version 3.0) in R 4.2. Four models were used for the GWAS analysis: General Linear Model (GLM), the Mixed Linear Model (MLM), the Fixed and Random Model Circulating Probability Unification (FarmCPU), the Bayesian-Information and Linkage-Disequilibrium Iteratively Nested Keyway (BLINK).

Results and Discussion: Our results indicated that the two multi-locus models, FarmCPU and BLINK, outperformed single-locus models, GLM and MLM, in beef herds in this study. Specifically, five significant SNPs were identified using FarmCPU, while BLINK and GLM each identified the other three. Also, three of these eleven SNPs (“BTA-40510-no-rs”, “BovineHD1400006853”, and “BovineHD2100020346”) were shared by multiple models. The significant SNPs were mapped to genes such as *EIF5*, *RGS20*, *TCEA1*, *LYPLA1*, and *MRPL15*, which were previously reported to be associated with carcass-related traits, growth, and feed intake in several tropical cattle breeds. This confirms that the genes identified in this study could be candidate genes for carcass weight in pasture-fed beef cattle and can be selected for further breeding programs to improve the carcass yield and productivity of pasture-finished beef cattle in Hawai’i and beyond.

KEYWORDS

GWAS, pasture-finished beef, carcass weight, Hawai’i, SNPs

Introduction

In Hawai'i, there is a considerable amount of land classified as pasture, much of which is suitable grazing land (Fukumoto et al., 2015). Year-round availability of forage favors cattle production and the beef industry in Hawai'i which significantly contributes to the state's economy (Asem-Hiablie et al., 2018). In Hawai'i there are two distinct climatic seasons: The relatively hot dry season runs from May to October, and the cool wet season runs from November to April (Giambelluca et al., 2014). Despite differences in the season, the tropical location of the Islands provides ample sunlight and moisture that supports year-round forage growth necessary for pasture-finished cattle farming which is less common in the continental United States and other regions. However, cattle stocking rates are often decreased during the dry season due to limited forage growth (Adhikari et al., 2022). At present, the total pasture area in the state is about 448,513 hectares, most of which (> 50%) is located on the Island of Hawai'i (Melrose et al., 2015). The limited availability of land is the main constraint for pasture-finished beef production in Hawai'i. Approximately 85% of calves are exported to the continental United States at weaning age and only 15% remain for local food supplies, which satisfied only 10%–13% of the local meat demand, and the gap is fulfilled by the imported meat (Asem-Hiablie et al., 2018; National Agricultural Statistics Service, USDA, 2021). There is a growing demand for pasture-finished beef among tourists and local consumers in Hawai'i, and genetic improvement of beef cattle is necessary for Hawai'i to increase production and improve productivity, leading to a larger local supply.

Carcass weight is a critical factor that affects beef cattle production and its economic returns (S.-H. Lee et al., 2014). This trait is impacted by both genetic and environmental factors (Irshad et al., 2013). Despite its significance, a limited number of studies have been conducted on Hawai'i cattle. A survey by Asem-Hiablie et al. (2018) explored the management practices, herd size, feeding, and marketing strategies under various production systems. Fukumoto and Kim (2007) examined carcass characteristics of forage-finished cattle in Hawai'i. Additionally, pasture-finished beef from Hawai'i competes with feed-lot-finished beef from the continental United States in terms of tenderness and marbling score (Kim, Fukumoto, and Kim, 2012). Information on cattle genetics and dedicated research on genes governing carcass yield and meat quality is an emerging scientific field that at present is under study in Hawai'i.

Genetic interventions hold the potential to significantly boost net carcass production and expand the local supply without additional retention of cattle heads for finishing. With the development of genetic testing technologies and the reduction in the cost of genotyping, significant advances in research have been made over the past two decades to improve the breeding and genetics of domestic cattle (Van Tassell et al., 2008; Weller et al., 2017; Tam et al., 2019). Affordable genotyping cost and the tendency of single nucleotide polymorphism (SNPs) to follow a pattern of linkage disequilibrium (LD) across the genome has opened up the arena for several genomic studies such as ancestry analysis, genomic selection (GS) and genome-wide association studies (GWAS) (Gautier et al., 2010; Goddard and Hayes, 2012; Decker et al., 2014). SNPs have been efficiently used to evaluate economic traits in cattle, such as carcass weight (Rolf et al., 2012; Costa et al., 2015; Keogh et al., 2021). Using SNP markers to evaluate cattle production and productivity is becoming essential in commercial cattle farming (Smith et al., 2019; Keogh et al., 2021). Hawai'i cattle, which are raised exclusively on pasture and thrive in the diverse geography and environment of the state, may possess unique genetic markers for carcass weight. Previous studies have successfully reported SNPs related to carcass weight, growth traits, feed intake, environmental adaptation, and meat quality traits in grain-finished or grass-finished production systems (Rolf et al., 2012; Costa et al., 2015; Silva et al., 2017; Smith et al., 2019; Keogh et al., 2021). However, there is a shortage of research on genetic markers, candidate genes, and their expression in pasture-finished beef in the tropical Pacific environment, and no specific genetic markers have been identified for Hawai'i cattle. (S. H. Lee et al., 2013; Silva et al., 2017; Edea et al., 2018; Hay and Roberts, 2018). Thus, focused genome-wide association studies (GWAS) in Hawai'i cattle can be valuable in discovering unique markers or verifying existing SNP markers for pasture-finished beef cattle in the tropical environment of Hawai'i.

One of the main challenges for the GWAS analysis is managing false positives and false negatives that may occur due to population structure and familial relationships. To address this issue, mixed linear models (MLMs) are commonly used, incorporating covariates for structure and kinship to control for false positives. LD is the non-random association of SNPs markers at different chromosome loci and is mainly determined by the physical distance between the markers, which can influence false positive and false negative rates. Several factors, such as population stratification, migration, recombination, mutation, and selection, affect the pattern of LD in a population (Goddard and Hayes, 2012; Karimi et al., 2015; Singh et al., 2021). In this study, four different statistical models, were compared for GWAS analysis for carcass weight in Hawaiian beef herds, including single-locus models: GLM—General Linear Model and MLM—Mixed Linear Model, and multi-locus models: FarmCPU and BLINK (Huang et al., 2019; Miao et al., 2019). GLM uses population structure (Q) as a covariate, and MLM uses both population structure and kinship (Q + K) as covariates. The FarmCPU model consists of the fixed-effect model (FEM) and the random-effect model (REM), which is evaluated iteratively. The effects in the FEM include the significant principal components, sex, and pseudo-quantitative trait nucleotides (Liu et al., 2016; Tang et al., 2019). Unlike GLM and MLM, which rely on one-to-one markers and trait correlation, FarmCPU trains multiple markers and builds correlations with significant SNP markers simultaneously. Additionally, markers from other loci are utilized as covariates to partially eliminate the confounding effects of the markers and kinship (Liu et al., 2016). BLINK is an enhanced version of FarmCPU (Wang and Zhang, 2021), which replaced REM with FEM for the model selection of the pseudo QTNs. Consequently, the iterations are eliminated to optimize the genetic-to-residual variance ratio, generating a higher statistical power than FarmCPU (Liu et al., 2016; Huang et al., 2019).

Therefore, this study aims to test several statistical models and determine the most appropriate model for association mapping in Hawai'i beef cattle by using carcass weight as a phenotype. The main objective of this paper is to identify SNP markers and candidate genes related to carcass weight for pasture-finished beef cattle. The goal of this research is to advance the genetic enhancement of beef cattle by incorporating the identified candidate genes in breeding programs in Hawai'i.

Methodology

Sample collection and genotyping

Samples were collected from 400 beef cattle at the commercial harvest facility in Kapolei, on the Island of Oahu. The sampling was random and included representatives from the major eleven ranches on four Islands in Hawaii: the Island of Hawaii (Big Island), Maui, Oahu, and Kauai. Both male and female cattle were included, and the age of the cattle ranged from younger than 30 months to older than 30 months to ensure that the samples were truly representative of the Hawai'i cattle population. The original source farms for the samples are mapped in the base map of Hawaii. However, to maintain the confidentiality of the commercial farms, we assigned them a nickname with Alphabet A to K, such as Hawaii Ranch A (HWRA) and Hawaii Ranch B (HWRB), as shown in Figure 1. All details about the source farms' location in the island chain and the number of cattle heads sampled from each farm can be found in Supplementary Table S1. At the slaughter plant, 10 mL of whole blood sample was taken from the jugular vein of each cattle using EDTA anticoagulant tubes, placed on ice, and eventually brought back to the laboratory where it was stored at −80°C. Genomic DNA was isolated from the blood samples using Quick-DNA Miniprep Plus Kit (Zymo Research, D4069), followed by the concentration measurement using NanoDrop™ One Microvolume UV-Vis Spectrophotometers. A total of 352 DNA samples with a concentration higher than 100 ng/uL and OD values between 1.8 and 2.2 were loaded for genotyping using the Neogen GGP Bovine 100 K BeadChip with the ARS-UCD1.2 assembly (Rosen et al., 2020). The raw genotypic data were checked and filtered for missing genotypes greater than 10%, minor allele frequencies (MAF) less than 0.05, Hardy Weinberg Equilibrium *p*-value less than 10^{-6}, and missingness per individual greater than 10%. SNPs on mitochondrial DNA and sex chromosomes were removed using PLINK 1.9. Therefore, 85 K SNPs in 351 cattle after the quality control were used in further analysis to identify the genetic markers associated with carcass weight in Hawai'i beef cattle herds. All animal experiments were conducted in accordance with NIH guidelines for housing and care of laboratory animals and under the University of Hawaii (UH) regulations UH IACUC Policy 18.0 after review and approval by the UH Animal Welfare and Biosafety Programs Committee (Assurance number A3423-01).

GWAS analysis

To test for normality, the Shapiro test was performed on the phenotypic data, specifically the carcass weight. Since the original data are not normally distributed, a log transformation was performed to normalize it. The normalized data on carcass weight were then subjected to a one-way analysis of variance (ANOVA), followed by *post hoc* Fisher's least significant difference (LSD) test using the Agricolae package (Steel, 1997) in R 4.2 (https://www.R-project.org/). Statistical significance was evaluated using an alpha value of 0.05, and any relationships falling below this value were considered significant. The log-transformed phenotypic data were eventually used in the GWAS analysis. GAPIT (version 3.0) R package (Wang and Zhang, 2021) was used to conduct association mapping with four different models: i) General Linear Model (GLM) (Price et al., 2006), ii) Mixed Linear Model (MLM) (Yu et al., 2006), iii) Fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016), and iv) Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Huang et al., 2019). Carcass weight was the phenotype to be tested, where age, sex, and farms were incorporated as covariates in all models to avoid confounding effects. All other parameters were set as default. GAPIT comes with a built-in function for intermediate analysis, which includes principal component (PC) analysis, kinship matrix calculation, and linkage disequilibrium decay (LD). To compute the kinship matrix, the algorithm developed by VanRaden (2008) was used. The developer team of GAPIT has provided a user-friendly manual (https://zzlab.net/GAPIT/gapit_help_document.pdf) that includes fundamental codes and pipelines with a brief explanation of the models and algorithms used for association mapping. The pipeline was followed in our study, and the LD decay plot was displayed over distance from the LD results from TASSEL 5 using R software. To check our population stratification, the genomic inflation factor, Lambda (λgc), was estimated, which is determined by comparing the median of the chi-squared test statistics obtained from a GWAS to the anticipated median of the chi-squared distribution. The median value of a chi-squared distribution with one degree of freedom is 0.4549364. The approach used to calculate lambda can differ based on the association analysis output, such as z-scores, chi-square statistics, or *p*-values (Devlin and Roeder, 1999)**.** In this study, we used *p*-values from the GWAS outcomes of all four models (GLM, MLM, FarmCPU, and BLINK) and computed λgc using the *qchisq*() function in R Studio 4.2. To correct for population structure, the first PC was fitted in the model, and the kinship matrix was added to account for the confounding effect of ancestry. Marker-trait associations were considered significant if they met an exploratory threshold of $p <$ 10–6 (-log10*p* > 6), which were displayed in Manhattan plots. The validity of associations was verified using the quantile-quantile (QQ) plot to distinguish between true and spurious results, such as false positives and false negatives (Stich et al., 2008; Kristensen et al., 2018). Additionally, the *p*-value threshold was adjusted for False Discovery Rate (FDR < 0.05), and any additional SNPs were listed in Supplementary Table S2.

Candidate gene selection and functional annotations

The result of linkage disequilibrium decay (LD) was used as a sliding window to find the genes within a certain distance from the identified SNP markers using the latest reference genome assembly for cows (*bosTau9 or ARS-UCD 1.2 genome assembly*) (Rosen et al., 2020). Candidate genes were selected within 100 kb upstream and downstream of the significant SNPs based on LD value for our cattle

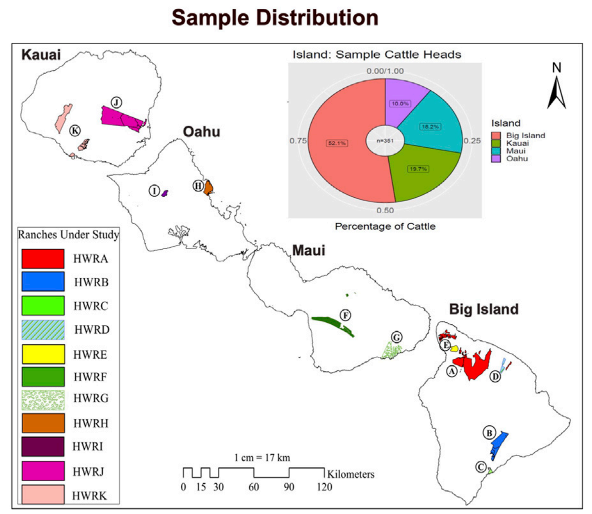


FIGURE 1
The cattle in this study and their distribution on Hawaiian islands. Eleven Hawaiian ranches coded as HWRA-HWRK are highlighted on the map of Hawai'i. The pie chart represents the percentage of cattle heads ($N = 351$) in each island (Big Island, Maui, Oahu, and Kauai) sampled in this study.

population. Three major genome browsers, UCSC (https://genome.ucsc.edu/), NCBI (https://www.ncbi.nlm.nih.gov/), and Ensembl (https://uswest.ensembl.org/index.html), were used to annotate the significant SNPs identified for carcass weight in Hawai'i cattle population. The function of the candidate genes, their homologs, and their roles in carcass-related traits and other mammals were explored through GeneCards (http://www.genecards.org/) and UniProt/Swiss-Prot browser (http://uniprot.org).

Results

Population distribution and phenotypic analysis

The population in our study included cows, steers, and heifers ($N = 351$ cattle heads) from eleven diversified farms located across the Islands of Hawaii (Big Island), Maui, Oahu, and Kauai (Figure 1). Over 52% of the cattle in this study are located on the Big Island, and the remaining 48% of the cattle were from neighboring islands, with at least two farms from each island. As reported, Hawaii and Maui have the largest cattle herds, contributing to more than 70% of the cattle heads in the Hawai'i island chain, followed by Kauai and Oahu (National Agricultural Statistics Service, USDA, 2021). Most of the ranching activities in Hawai'i were concentrated on the Big Island and Maui due to the availability of suitable pasture and their favorable climate for forage growth. In our study, 70% of the cattle heads were taken from these two islands (Figure 1), making our sample population proportionately represent the beef cattle distribution in Hawai'i.

The raw phenotypic data for carcass weight failed the Shapiro test for normality ($p < 0.05$), indicating a deviation from a normal distribution. The results of the raw data revealed a right-skewed distribution with a few extreme values (Figures 2A, B). Thus, we log-transformed the raw data before conducting further analysis. The normality test on transformed data showed a normal distribution ($p > 0.05$), with the majority of the data concentrated around the central value (Figure 2C). Analysis of variance conducted on the log-transformed carcass values demonstrated a significant difference ($p < 0.05$) in carcass yield among the islands. The pair-wise multiple comparisons mean tests revealed that cattle from Maui Island had significantly lower carcass weight ($p < 0.05$) than those from other neighboring islands (Figure 2D; Supplementary Tables S3, S4), whereas the carcass weight of cattle from the other three islands did not differ significantly.

Population structure and linkage disequilibrium decay

The present study employed principal component analysis (PCA) to investigate the genetic structure of 351 Hawai'i cattle, utilizing quality-controlled single nucleotide polymorphism (SNP) data. The analysis revealed that there were no distinct genetic clusters among the sampled accessions, as shown in Figure 3A. This observation is consistent with a uniform genetic background of cattle across the Island chain, leading to the presence of a single linear cluster with no population structure. Furthermore, the results demonstrated that the cattle herds across the Hawai'i island chain shared similarities, with no distinct clusters based on allelic SNPs explained by the first two principal components. The scree plot for eigenvalues revealed that the first two principal components explained 5% of the total variability. The sharply declined elbow at the second PC indicates that the first PC was the major source of variability used to correct for possible population structure, while subsequent PCs explained less variability after reaching the lowest elbow point (Figure 3B). Additionally, the heatmap and dendrogram of the kinship matrix confirmed the absence of clear clusters in the population, indicating that the cattle population in this study is unrelated by family (Figure 3C). The square of the correlation coefficient between the markers at two loci (r^2) was used to evaluate the LD (linkage disequilibrium) estimate. When LD reaches an r^2 value below 0.2, it is typically expected to decay by half (Vos et al., 2017; Singh et al., 2021). In the case of the Hawai'i cattle population under investigation, the LD decay reached an r^2 value of 0.15 at approximately 100 kb (Figure 4). Thus, to identify candidate genes associated with the carcass weight of beef cattle, we searched for genes within a 100 kb range upstream and downstream of the identified SNPs using the UCSC genome browser.

Association analysis

In this study, four different models were used to compare their strengths in controlling false positives and false negatives in the population. Among them, two multi-locus models, BLINK and FarmCPU, coincided with the expected straight line diagonally with a sharp deviation at the tail, indicating true association by controlling both false positives and false negative markers. The genomic inflation factors (λgc) were calculated for four models to check our population stratification, where GLM had a λgc of 1.55,

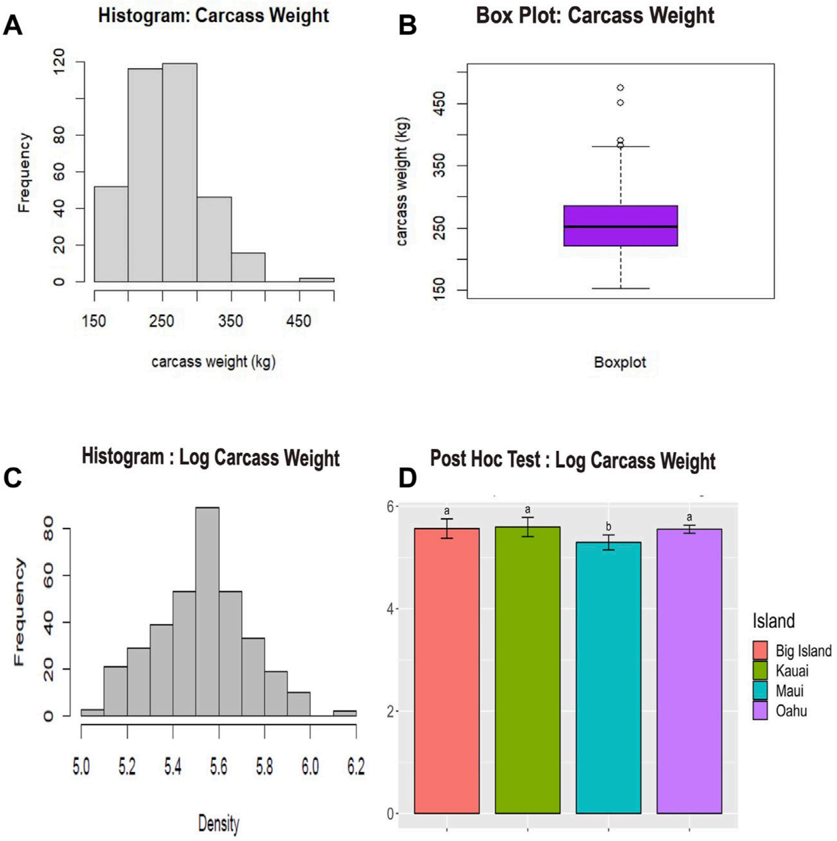


FIGURE 2
Phenotypic data analysis for carcass weight of cattle. **(A)** The histogram displays the distribution of unadjusted carcass weight. **(B)** The boxplot summarizes the statistical information of the unadjusted carcass weight, including outliers. **(C)** The histogram shows the distribution of log-transformed carcass weight. **(D)** Fisher's *post hoc* least significant difference (LSD) test (*p < 0.05*) conducted across the islands. This *post hoc* test followed an Analysis of Variance (ANOVA) test *(p < 0.05*). The lowercase letters above the bar plot indicate significant differences among the groups. Groups with the same letter are not significantly different from each other, while groups with different letters are significantly different.

MLM had a λgc of 1.002, FarmCPU had a λgc of 1.03, and BLINK had a λgc of 1.11 (Figure 5). The quantile-quantile (qq) plot showed a significant upward deviation from the straight line in GLM, indicating a high incidence of false positives. Conversely, the qq plots for the MLM, FarmCPU, and BLINK models exhibited a normal distribution of *p*-values, with λgc values close to 1, indicating effective control of spurious results and a high likelihood of true associations. Among the four models used in the association study, three models (GLM, FarmCPU, and BLINK) identified significant SNPs associated with carcass weight. Five significant SNPs were identified with the FarmCPU model while the other three significant SNPs were identified with the model BLINK ($p < 10^{-6}$) (Figures 5, 6). Two of these SNPs, "*BTA-40510-no-rs*" and "*BovineHD2100020346*", were shared between the two models and are considered to have a stronger association with carcass weight in Hawai'i beef cattle herds (Table 1). Additionally, a few SNP markers are unique with each model: "*BovineHD0100011931*", and "*BovineHD0200007999*" on chromosomes 1 and 2 were found with FarmCPU solely, and "*BovineHD0500025848*" on chromosome 5 was found only with BLINK. Three SNP markers identified with GLM are shared with other models. Without validation from other more robust models, results from GLM alone would have been inconclusive. However, the presence of the same markers in multi-locus models erased the doubt for false positives by GLM. The MLM was too conservative and was not able to identify any associated markers for carcass weight (Figures 5, 6). False negatives may have arisen due to model overfitting, as this population was free from population structure and family relatedness. In addition, adjusting the *p*-value threshold to False Discovery Rate (FDR <0.05) resulted in the identification of a greater number of significant SNPs. Specifically, 58 significant SNPs were identified after FDR adjustment, and these are listed in Supplementary Table S2.

Identification of candidate genes

All the significant SNPs identified in our population were explored for their biological function annotation. Genes within 100 kb upstream and downstream relative to the identified SNPs were scanned, and eleven genes (*ZMAT3, CERS6, PLEKHA5, MYCT1, RGS20, TCEA1, LYPLA1, MRPL15, EIF5, CKB, and MARK3*) were identified to be associated with carcass weight in Hawai'i cattle. Also, eight of these genes (*ZMAT3, RGS20, TCEA1, LYPLA1, MRPL15, EIF5, CKB, and MARK3*) overlapped with significant SNPs identified by at least two models (FarmCPU,

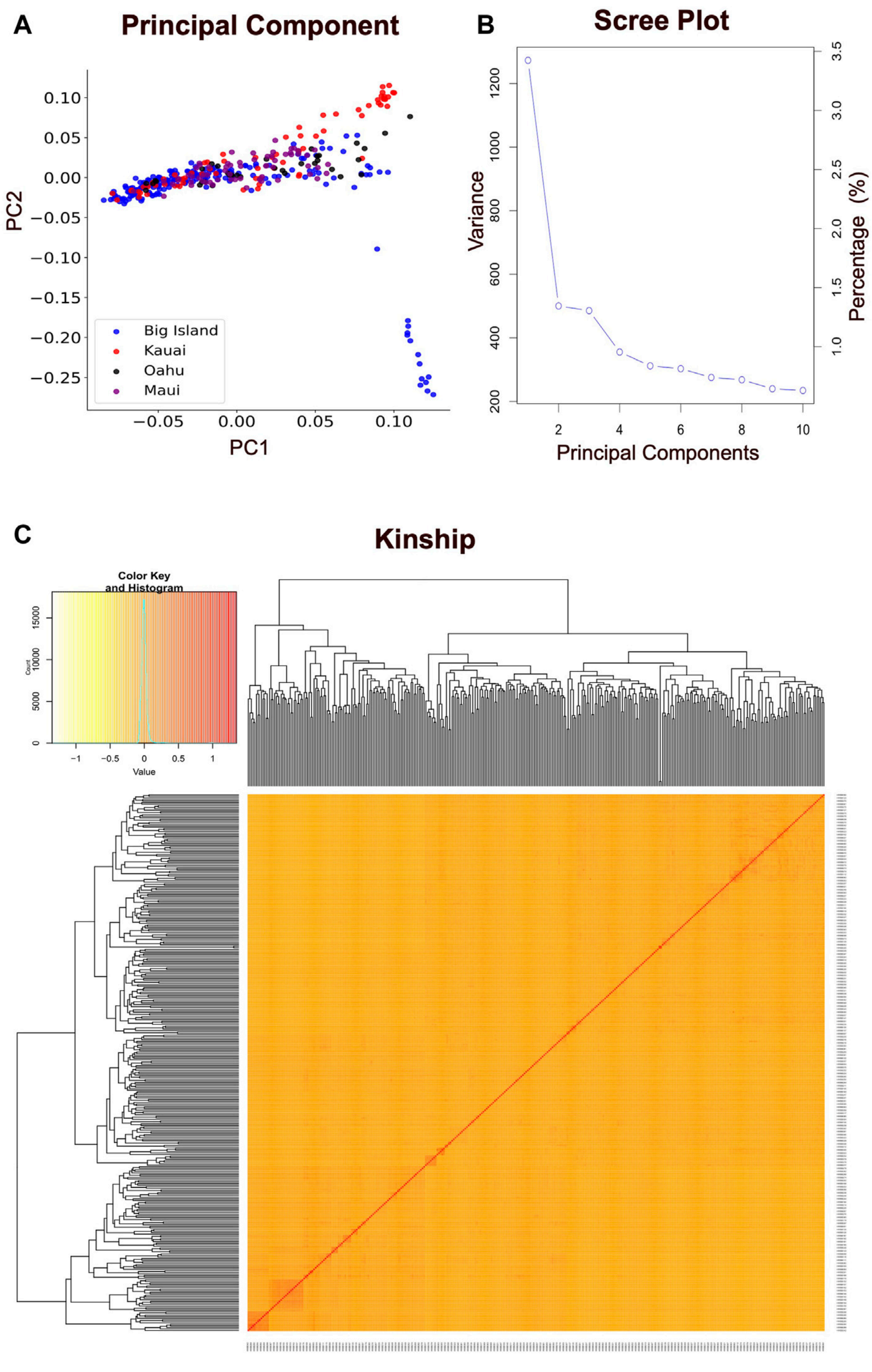


FIGURE 3
Population structure analysis. **(A)** The principal component analysis (PCA) results for 351 Hawai'i cattle from four islands, using 85K quality-controlled SNP data. Each colored dot represents an individual animal located on a different island. The horizontal and vertical axes represent the first and second principal components, respectively, contributing to 3.5% and 1.5% of the total variability in the data. The PCA analysis provides insight into the genetic relationships among cattle populations and identifies patterns of genetic variation. **(B)** The scree plot illustrates the variance accumulation of the top ten principal components (PCs). The *x*-axis represents the top ten PCs, while the *y*-axis represents eigenvalues that signify the amount of variation. The accumulated variance for each PC is denoted by an empty circle. The "elbow" point of the curve, where the slope begins to level off, signifies the number of factors that the analysis should generate. **(C)** The hierarchical clustering and heat map of the pairwise kinship matrix values, calculated based on 85K quality-controlled SNPs from 351 Hawaiian cattle. The color histogram illustrates the distribution of the coefficient of coancestry, with stronger red colors indicating higher levels of relatedness among individuals.

BLINK, and GLM) (Table 1). Details of candidate genes (Gene name, Chromosome number, Ensemble ID) and their biological functions in other mammals including humans, mice, sheep, pigs, and dairy cows are presented in Supplementary Table S5. It is worth mentioning that five of these genes (*RGS20, TCEA1, LYPLA1, MRPL15*, and *EIF5*) have been previously identified in several

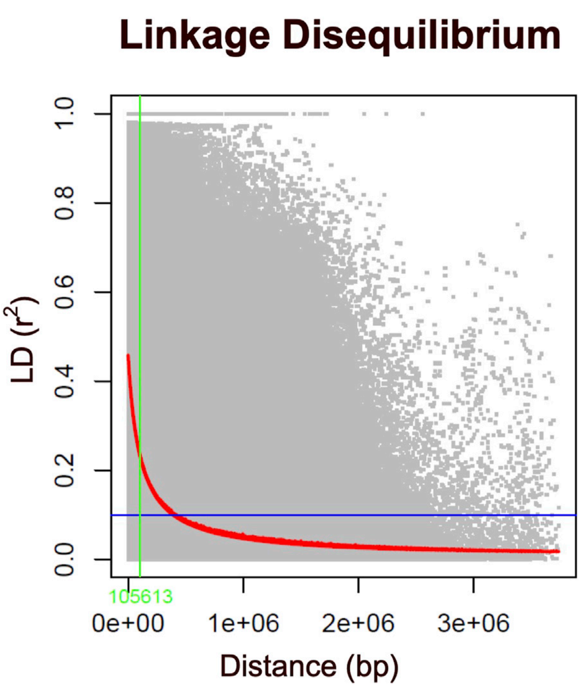


FIGURE 4
Genome-wide linkage disequilibrium (LD) decay plot for 351 Hawaiian cattle based on 85K SNP markers. The LD, measured as the squared correlation coefficient (r^2) between pairs of polymorphic markers, is plotted against their genetic distance (bp) across the chromosomes. The red line represents the moving average of the 10 adjacent markers, while each gray dot represents a pair of distances between two markers on the window and their corresponding squared correlation coefficient. The blue line denotes the LD cutoff of 0.1, and the green line indicates the critical LD at a distance of less than 100 kb.

studies to be associated with carcass traits and growth traits in cattle (Lindholm-Perry et al., 2012; Hay and Roberts, 2018; Doyle et al., 2020). *ZMAT3* has been reported to be correlated to conception rate and fertility in Brangus cattle (Fortes et al., 2012), and *CKB* and *MARK3* are related to milk production and somatic cell score in Holstein dairy cattle (Buaban et al., 2022). *PLEKHA5*, *MYCT1*, and *CERS6* were identified for the first time in cattle in our study, however, these genes are conserved in more than 200 other organisms, vertebrates, and mammals, including humans, chimpanzees, rhesus monkeys, dogs, mice, rats, and cows (https://www.ncbi.nlm.nih.gov/gene/). The specific roles of these genes in cattle have not been defined yet and may serve as supporting and maintenance functions.

Discussion

Association mapping and efficient model for GWAS in cattle

Out of the four models used in this study, BLINK, FarmCPU, and GLM performed well in predicting significant SNP markers for carcass weight, while MLM failed to identify the markers with the same trait. The GLM commonly produced false positives and the MLM commonly produced false negatives. These results were consistent with other published results (Tamba et al., 2017; Wen et al., 2018). Based on the results from this study, it is possible to conclude that, a conventional single-locus model like MLM was too stringent to identify the SNP markers. MLM can produce better results when there is evidence of population structure due to the geography diversity or family relatedness, but the cattle in this study were from the same geographic area (Hawai'i Pacific) and represented several farms, ruling out the alleles to have family relatedness, which was reflected in our results of no kinship clusters observed (Figure 3C). However, weaker family relatedness observed as small patches across the diagonals at multiple spots in the kinship heat map plot could not be ignored, and therefore, the kinship matrix was fitted as covariates to adjust the confounding effect due to family relatedness in our model. MLM accounts for both covariates due to PC and kinship; therefore, the model got overfitted and might have resulted in false negative SNPs. In contrast, QQ plots with the FarmCPU and BLINK models showed a straight line close to the 1:1 with a slightly deviated tail, indicating that FarmCPU and BLINK controlled false positives and false negatives without compromising the results for associated markers (Figure 5). Our main results i.e., identified SNPs were primarily from FarmCPU and BLINK, while GLM identified the same markers as did by the other two models. The number of PCs seems to affect less multilocus models such as FarmCPU and there is no concrete gold standard for how many PCs to be included to correct for the possible population stratification (Wang and Zhang, 2021). We followed a general convention of using the number of PC-based observations of the elbow on the scree plot. Our results had an elbow on the second PC, indicating the first PC is the major source of variability. Therefore, we used a single PC to correct for population stratification, which is almost equivalent to the results of using a sole kinship model with no PCs as covariates. However, including one PC helped to elaborate the model making it a full model for GWAS. Further, the genomic inflation factor (λgc) ranged between 1.002 and 1.11 among MLM, FarmCPU, and BLINK, indicating that these three models best fitted for GWAS in Hawaiian cattle herds, while GLM did not fit properly with λgc above 1.5. Results from GLM would have been questionable, as the QQ plot deviated sharply from the expected line (λgc > 1.1). However, three shared SNP markers identified by multi-locus models increase the validity of the true association between SNP markers and the trait of carcass weight in our study, which is similar to a single-trait GWAS study in wheat, FarmCPU and BLINK performed better than MLM in identifying the associated markers (Merrick et al., 2022).

In this study, FarmCPU identified five significant markers, which are more than BLINK identified; two of those significant SNP markers are common in both models. Researchers are increasingly using multi-locus models in association mapping, more exclusively in plants and some in animal studies. In a similar study in plants, researchers compared several qualitative traits in soybean and maize flowers using eight popular models where FarmCPU performed better for most of the traits than other models, including GLM and MLM (Kaler et al., 2020). FarmCPU is gaining popularity today due to improved statistical power when compared to other methods. The problem of model overfitting is minimized when using the FarmCPU model because of a two-step

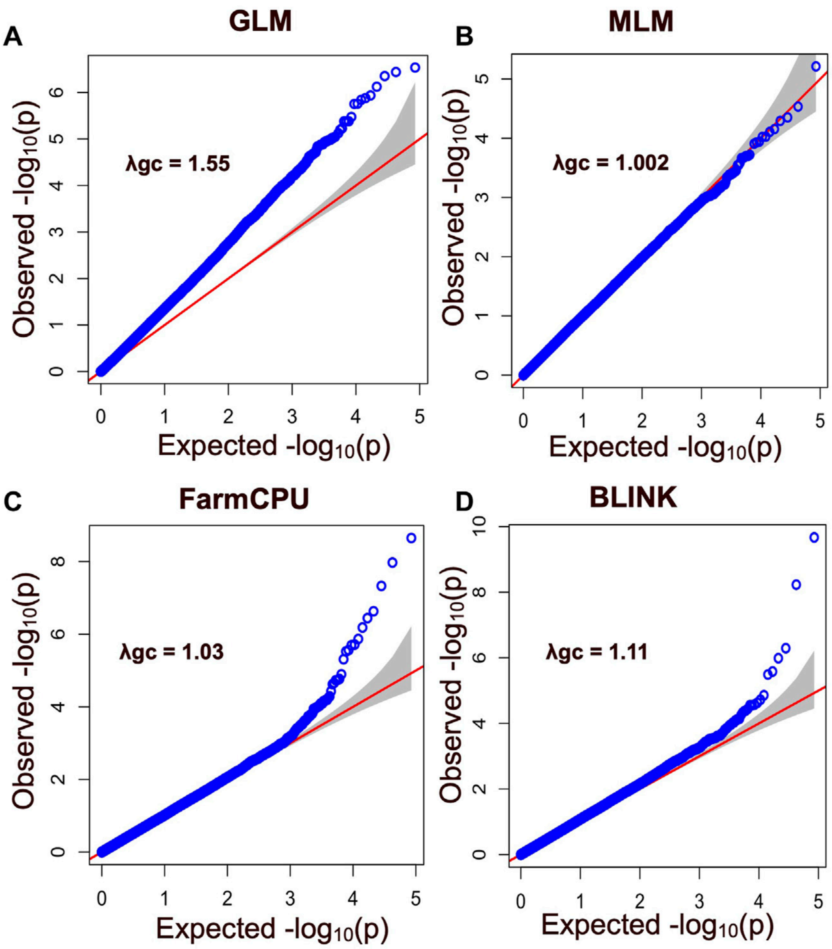


FIGURE 5
Quantile - Quantile (QQ) plot showing SNP markers with their observed and expected *p values* for different models including **(A)** General Linear Model (GLM), **(B)** the Mixed Linear Model (MLM), **(C)** the Fixed and Random Model Circulating Probability Unification (FarmCPU), and **(D)** the Bayesian-Information and Linkage-Disequilibrium Iteratively Nested Keyway (BLINK). Blue circles correspond to the *p*-values derived from the principal components + kinship model, while the red line indicates the expected *p*-value distribution under the null hypothesis that the *p*-values follow a uniform [0, 1] distribution. The gray shadow area represents the 95% confidence interval for the QQ plot under the null hypothesis of no association between the SNP and the trait. The -log 10 (p) negative base 10 logarithms of the *p*-values (probability of type-I error made in GWAS hypotheses testing) are also shown. This plot provides insight into the distribution of significant associations between the SNP and the trait, with deviations from the expected distribution indicating the presence of false-positive associations or other factors affecting the association analysis.

adjustment. The first adjustment involves fitting the covariates from population structure, family relatedness, and pseudo-quantitative trait nucleotides. The second adjustment involves either refining the covariates or selective inclusion or exclusion of pseudo-quantitative traits based on their relationship with the testing markers (Liu et al., 2016). Therefore, model selection becomes a crucial step in GWAS to prevent the loss of valuable markers as false negatives and to control the biased associations that are not truly associated with the traits and appear as false positives. In addition, single-locus models, which only consider a one-to-one independent relationship between markers and traits, do not accurately apply to the biological phenomena as the interaction of genes is a common phenomenon in trait expression. In contrast, multi-locus models simultaneously test the association of multiple markers for a given trait (Liu et al., 2016), which is closer to biological phenomena involving gene action and interactions. MLM initially gained popularity over GLM due to its higher statistical power, however, multi-locus models like FarmCPU and BLINK surpass both GLM and MLM in statistical power and computational efficiency (Liu et al., 2016; Huang et al., 2019). MLM only considers population structure and kinship as covariates so markers in various loci that are not significant sometimes may appear as false positives or false negatives. In contrast, FarmCPU establishes a relationship with the marker at one locus and treats all other markers at different loci as covariates, reiterating again and again and completing the K iteration, where K = SNPs ((Liu et al., 2016). This way assigning the non-significant markers at multiple loci as covariates minimizes the chances of their appearance as false positives or false negatives with multi-locus models (Liu et al., 2016; Tang et al., 2019; Kaler et al., 2020), which was also observed in our results where FarmCPU and BLINK outperformed over MlM. Based on our results, it can be suggested that using multi-locus models for association mapping in

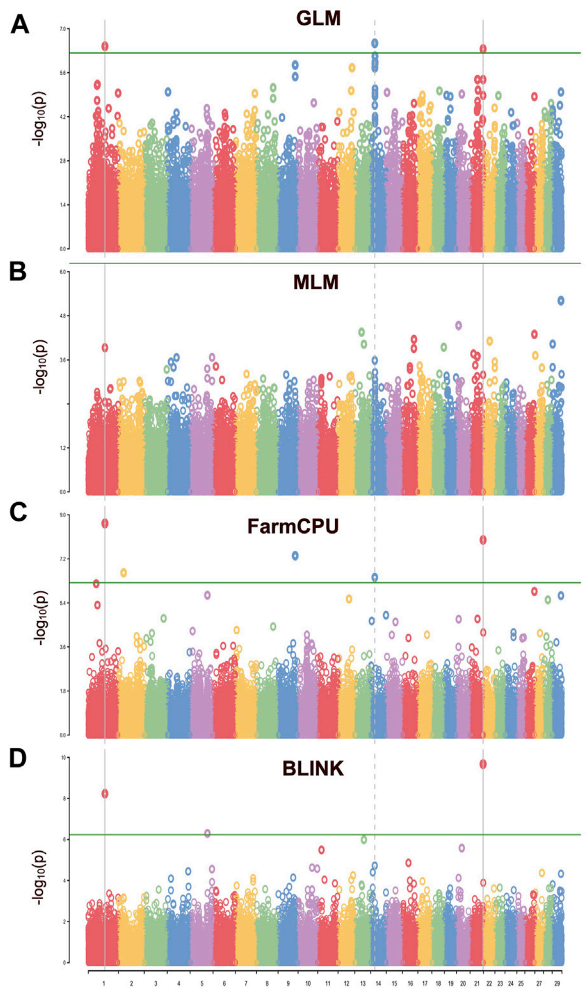


FIGURE 6
The Manhattan plot for the GWAS of carcass weight in the Hawaiian cattle population. Each chromosome (*x*-axis) is represented by a different color, and the plot is based on -log 10 (*p*-value) from GWAS against chromosome position. The green line indicates the genome-wide significant threshold at *p* = 10e-6, while vertical lines highlight the chromosomes containing significant SNPs. **(A)** SNPs identified by single-locus GLM. **(B)** No SNPs were identified by single-locus MLM. **(C)** SNPs identified by multi-locus FarmCPU model **(D)** SNPs identified by multi-locus BLINK model. This plot provides an overview of the distribution of significant associations across the genome, with peaks indicating regions of potential interest for further investigation.

animal-modeled research studies may be a better option than relying solely on single-locus models, as is commonly done in recent association studies in plant models.

Candidate genes related to carcass traits in cattle

Bovine chromosome 14 (BTA14) has been widely explored for quantitative trait loci (QTL) and genes related to feed intake, weight gain, and carcass traits in dairy and beef cattle (Smith et al., 2019; Srikanth et al., 2020). The genes, located in a conserved region on BTA14, have been reported as a selective sweep region in dairy and beef cattle breeds (Zhao et al., 2015), and the DNA regions on BTA14 have been associated with backfat thickness, rib eye muscle area, marbling, and other carcass traits in beef cattle (Lindholm-Perry et al., 2012; S. H. Lee et al., 2013; Zhang et al., 2019). In another recent study, *RGS20, TCEA1, LYPLA1*, and *MRPL15* on BTA14 have been associated with the back fat thickness (BFT) and Intra Muscular Fat (IMF) in a composite beef cattle breed (Hay and Roberts, 2018). *RGS20* has also been found to be involved in actin cytoskeleton organization which governs meat tenderness in European beef cattle breeds (Mengistie et al., 2022). *TCEA1* and *LYPLA1* have been associated with average daily feed intake and average daily weight gain in composite cattle breeds (Lindholm-Perry et al., 2012; Grigoletto et al., 2019; Hay et al., 2022). Furthermore, *RGS20* was associated with thigh width in Angus breeds (Doyle et al., 2020) and average daily weight gain in Yorkshire pig breeds (Cai et al., 2022). *EIF5* gene on chromosome 21 was also found to be associated with marbling and carcass traits in Nellore cattle (Carvalho et al., 2019). Expression of the *EIF5* gene positively contributes to the growth of the longissimus thoracis muscle in *Bos. Indicus* (Bruscadin et al., 2021). Candidate genes identified in this study and their roles in some other breeds of discussed above presented in Table 2. From these findings and discussions, it can be concluded that *RGS20, TCEA1, LYPLA1, MRPL15*, and *EIF5* genes are strongly associated with carcass weight in Hawai'i beef cattle. Additionally, *MYCT1* and *CERS6* are possibly candidate genes for carcass weight in cattle, as their roles have been identified in pigs and sheep (Wu et al., 2020; Xu et al., 2021). Further studies are required to ascertain the association of *MYCT1* and *CERS6* genes with carcass weight and to better understand their biological roles in cattle.

Candidate genes in other mammals

Previous work identified *CERS6* and *MYCT1* genes have an association with carcass-related traits in other animals (Wu et al., 2020; Xu et al., 2021; Buaban et al., 2022) but not in beef cattle. Few of the genes (*ZMAT3*, *PLEKHA5*, *CKB,* and *MARK3)* identified in our study were reported in dairy cattle (Buaban et al., 2022). Very little information on above mentioned six genes within beef cattle is available, however, some details about their biological processes and homologs in other mammals including humans, mice, sheep, pigs, and dairy cows are listed in Supplementary Table S5. *CERS6* and *MYCT1* have been studied to be associated with obesity, weight gain, and subcutaneous fats in several mammalian species, including humans, mice, sheep, and pigs. *CERS6* enables sphingosine N-acyltransferase activity and is involved in the membrane's ceramide biosynthetic process. In an association study, *CSER6* was associated with fat deposition in sheep (Xu et al., 2021). Another study found that *CERS6* was associated with subcutaneous fat in lamb in response to a concentrate-supplemented diet (González-Calvo et al., 2017). *CERS6* expression positively correlates with BMI, body fat content, and obesity in humans. Upregulation of *CERS6* and subsequent increase in specific acyl-chain ceramides contributes to both murine and human obesity (Turpin et al., 2014). *MYCT1* gene was predicted to regulate specific *MYC* target genes. The role of the *MYCT1* gene in

TABLE 1 List of the significant SNPs ($p < 10^{-6}$) associated with carcass weight in the Hawaiian beef cattle population.

SNP	Position	MAF	Allele	Effect	Model	Candidate genes
BTA-40510-no-rs	chr1: 88012422	0.45441595	G/A	0.054	FarmCPU, BLINK, GLM	*ZMAT3*
BovineHD0200007999	chr2: 27466243	0.36324786	C/T	−0.045	FarmCPU	*CERS6*
BovineHD0500025848	chr5: 90643657	0.16524217	A/C	0.052	BLINK	*PLEKHA5*
BTB-01839335	chr9: 89664367	0.48005698	A/G	−0.043	FarmCPU	*MYCT1*
BovineHD1400006853	chr14: 21949250	0.09116809	T/C	0.084	FarmCPU, GLM	*RGS20, TCEA1, LYPLA1, MRPL15*
BovineHD2100020346	chr21: 68056605	0.43589744	A/C	0.047	FarmCPU, BLINK, GLM	*EIF5, CKB, MARK3*

Notes: MAF, minor allele frequency; Allele: The first allele is the nucleotide of the reference allele; The second allele is the nucleotide of the alternate allele; Effect: the contributing weightage of SNPs to carcass weight; Model, the different models that successfully identified SNPs associated with carcass weight; Candidate Genes, the genes that correspond to the significant SNPs in the range of upstream 100 kb and downstream 100 kb (reference genome, ARS-UCD1.2/bosTau9).

TABLE 2 Candidate genes and their roles in beef cattle.

Candidate gene	Role in cattle	Literature
RGS20	Back fat thickness, Intramuscular fat, and meat tenderness in composite beef cattle breeds	Hay and Roberts (2018)
TCEA1	Growth traits in Montana tropical composite cattle	Grigoletto et al. (2019)
LYPLA1	Feed intake, growth, and average daily weight gain in composite beef cattle and cross breeds	Lindholm-Perry et al. (2012); Hay et al. (2022)
MRPL15	Residual feed intake in Australian Angus; Muscle growth in cattle	Cassar-Malek et al. (2007); Heras-Saldana et al. (2019)
EIF5	Muscle growth, marbling, and meat quality traits in Nellore cattle	Carvalho et al. (2019); Bruscadin et al. (2021)

cattle has yet to be studied, but it is associated with meat quality and pH value in Qingyu pigs, specifically, *MYCT1* is involved in skeletal muscle development, regulation of Ca^{2+} release in the muscle, and anaerobic respiration, governing superior meat quality traits in Qingyu pigs (Wu et al., 2020). These genes (*CERS6* and *MYCT1*) are mostly related to muscles, subcutaneous fat, and obesity in humans, mice, sheep, and pig's meat, placing them in the list of possible candidate genes for carcass-related traits in beef cattle. CKB (Creatine Kinase B) is conserved in 315 mammals, including humans, mice, monkeys, and cattle. A previous study revealed that the *CKB* gene located in BTA 21 governs the fertility of cattle (Han and Peñagaricano, 2016). The *CKB* gene encodes an enzyme creatine kinase, and the elevated level of creatine kinase in the sperm causes oligospermia and male sterility (Gergely et al., 1999). *MARK3* is associated with bone mineral density in humans and mice (Calabrese et al., 2017). The protein encoded by this gene is activated by phosphorylation and in turn, is involved in the phosphorylation of tau proteins MAP2 and MAP4. *PLEKHA5* has been associated with milk-fat yield in Holstein cattle (Jiang et al., 2019; Pedrosa et al., 2021). *CKB* and *MARK3* genes located together in the same genomic region on chr21 have been reported to be associated with milk production and somatic cell score in Holstein cattle (Buaban et al., 2022). Therefore, genes *PLEKHA5, CKB,* and *MARK3* identified from our study might indirectly contribute to weight gain and carcass traits during the early stage of growth under cow-calf operation, ensuring optimum milk supply before weaning age. However, there is no clear evidence yet for these genes (*ZMAT3*, *CERS6, PLEKHA5, MYCT1, MARK3*, and *CKB*) regarding their association with carcass-related traits in beef cattle, and further research on their functional validations is required to confirm whether these genes are indeed associated with carcass weight and meat quality traits in beef cattle.

Conclusion

Multi-locus models such as FarmCPU and BLINK were found to be superior to single-locus models (GLM and MLM) in identifying SNP markers and minimizing false positives. Two SNP markers were identified using both multi-locus models. Three other markers were identified using GLM and FarmCPU models, strengthening the correlation of these SNPs with carcass weight in beef cattle. The *EIF5* gene on chromosome 21 and four other genes in the BTA14 region (*RGS20, TCEA1, LYPLA1,* and *MRPL15*) were found to be associated with carcass weight in Hawai'i beef cattle, and these results align with the previous findings showing their correlation with carcass weight and related traits in other cattle breeds. Future work incorporating selection pressures using these genes (*EIF5, RGS20, TCEA1, LYPLA1,* and *MRPL15*) will facilitate genetic improvement in Hawaiian beef cattle, enhancing productivity while utilizing limited resources without harming the delicate ecosystem of the island.

Data availability statement

Our SNP genotyping data reported are available in the DDBJ Genomic Expression Archive under the accession number PRJDB15706.

Ethics statement

The animal study was reviewed and approved by The University of Hawaii (UH) regulations UH IACUC Policy 18.0 after review and approval by the UH Animal Welfare and Biosafety Programs Committee (Assurance number A3423-01).

Author contributions

YH and CL contributed to the conception and design of the study. MO and KC contributed to sample collection. MK defined methods and contributed to data analysis. MA organized the data and performed the statistical analysis, interpreted the results, and wrote the first draft of the manuscript. YH, CL, MK, and RL visualized results and supervised editing. All authors contributed to the manuscript revision and approved the submitted version.

Funding

This work was supported by the National Institute of Food and Agriculture (NIFA) through the Hatch research agencies under the United States Department of Agriculture (USDA) (HAW02062-H) and administered by the College of Tropical Agriculture and Human Resources (CTAHR) at the University of Hawai'i at Mānoa. The authors would like to extend their gratitude to Hawai'i Meats LLC (Kapolei, Oahu, Hawai'i) and its staff, particularly Kamuela Barr and Christopher Cravalho, for their invaluable contribution to this study during the period of July-December 2021. Their support in providing essential samples was crucial for the successful completion of the research, and we appreciate their cooperation and dedication to the project. The technical support and advanced computing resources from the University of Hawaii Information Technology Services – Cyberinfrastructure, funded in part by the National Science Foundation CC* awards # 2201428 and # 2232862 are gratefully acknowledged.

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.2023.1168150/full#supplementary-material

References

Adhikari, M., Longman, R. J., Giambelluca, T. W., Lee, C., and He, Y. (2022). Climate change impacts shifting landscape of the dairy industry in Hawaii. *Transl. Animal Sci.* 6, txac064. doi:10.1093/tas/txac064

Asem-Hiablie, S., Rotz, C. A., Sandlin, J. D., M'Randa, R. S., and Stout, R. C. (2018). Management characteristics of beef cattle production in Hawaii. *Prof. Animal Sci.* 34 (2), 167–176. doi:10.15232/pas.2017-01691

Bruscadin, J. J., de Souza, M. M., de Oliveira, K. S., Rocha, M. I. P., Afonso, J., Cardoso, T. F., et al. (2021). Muscle allele-specific expression QTLs may affect meat quality traits in Bos indicus. *Sci. Rep.* 11 (1), 7321. doi:10.1038/s41598-021-86782-2

Buaban, S., Lengnudum, K., Boonkum, W., and Phakdeedindan, P. (2022). Genome-wide association study on milk production and somatic cell score for Thai dairy cattle using weighted single-step approach with random regression test-day model. *J. Dairy Sci.* 105 (1), 468–494. doi:10.3168/jds.2020-19826

Cai, Z., Christensen, O. F., Lund, M. S., Ostersen, T., and Sahana, G. (2022). Large-scale association study on daily weight gain in pigs reveals overlap of genetic factors for growth in humans. *BMC Genomics* 23 (1), 133. doi:10.1186/s12864-022-08373-3

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and Co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst.* 4 (1), 46–59. doi:10.1016/j.cels.2016.10.014

Carvalho, M. E., Baldi, F., Alexandre, P. A., de Almeida Santana, M. H., Ventura, R. V., Bueno, R. S., et al. (2019). Research Article Genomic regions and genes associated with carcass quality in Nelore cattle. *Genet. Mol. Res.* 18 (1), 1–15. doi:10.4238/gmr18226

Cassar-Malek, I., Passelaigue, F., Bernard, C., Léger, J., and Hocquette, J-F. (2007). Target genes of myostatin loss-of-function in muscles of late bovine fetuses. *BMC Genomics* 8 (1), 1–11. doi:10.1186/1471-2164-8-63

Costa, R. B., Camargo, G. M. F., Diaz, I. D. P. S., Irano, N., Dias, M. M., Carvalheiro, R., et al. (2015). Genome-wide association study of reproductive traits in Nellore heifers using bayesian inference. *Genet. Sel. Evol.* 47 (1), 1–9. doi:10.1186/s12711-015-0150-4

Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Molina Alcalá, A., Sonstegard, T. S., et al. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* 10 (3), e1004254. doi:10.1371/journal.pgen.1004254

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55 (4), 997–1004. doi:10.1111/j.0006-341x.1999.00997.x

Doyle, J. L., Berry, D. P., Veerkamp, R. F., Carthy, T. R., Evans, R. D., Walsh, S. W., et al. (2020). Genomic regions associated with muscularity in beef cattle differ in five contrasting cattle breeds. *Genet. Sel. Evol. GSE* 52, 2. doi:10.1186/s12711-020-0523-1

Edea, Z., Jeoung, Y. H., Shin, S-S., Ku, J., Seo, S., Kim, I-H., et al. (2018). Genome–wide association study of carcass weight in commercial hanwoo cattle. *Asian-Australasian J. Animal Sci.* 31 (3), 327–334. doi:10.5713/ajas.17.0276

Fortes, M. R. S., Snelling, W. M., Reverter, A., Nagaraj, S. H., Lehnert, S. A., Hawken, R. J., et al. (2012). Gene network analyses of first service conception in Brangus heifers: Use of genome and trait associations, hypothalamic-transcriptome information, and transcription factors. *J. Animal Sci.* 90 (9), 2894–2906. doi:10.2527/jas.2011-4601

Fukumoto, G. K., and Kim, Y. S. (2007). Carcass characteristics of forage-finished cattle produced in hawai'i. Available at: https://www.hicattle.org/Media/HICattle/Docs/fst-25.pdf.

Fukumoto, G. K., Thorne, M. S., Silva, J. H., Deenik, J. L., and Stevenson, M. H. (2015). *Suitability map for foragefinished beef production using GIS technology: Hawaii island." pasture and range management PRM-7.* Honolulu, Hawai'i: University of Hawai 'i at Mānoa, CTHAR Publication, 1–6. Available at: https://www.ctahr.hawaii.edu/oc/freepubs/pdf/PRM-7.pdf.

Gautier, M., Laloë, D., and Moazami-Goudarzi, K. (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PloS One* 5 (9), e13038. doi:10.1371/journal.pone.0013038

Gergely, A., Szöllösi, J., Falkai, G., Resch, B., Kovacs, L., and Huszar, G. (1999). Sperm creatine kinase activity in normospermic and oligozospermic Hungarian men. *J. Assisted Reproduction Genet.* 16 (1), 35–40. doi:10.1023/A:1022545612784

Giambelluca, T. W., Shuai, X., Barnes, M. L., Alliss, R. J., Longman, R. J., Miura, T., et al. (2014). Evapotranspiration of Hawai 'i. Final report submitted to the US army corps of engineers—honolulu district, and the commission on water resource management, state of Hawai 'i. Available at: http://evapotranspiration.geography.hawaii.edu/assets/files/PDF/ET%20Project%20Final%20Report.pdf.

Goddard, M. E., and Hayes, B. J. (2012). Genome-wide association studies and linkage disequilibrium in cattle. *Bov. Genomics*, 192–210. doi:10.1002/9781118301739.ch13

González-Calvo, L., Dervishi, E., Joy, M., Sarto, P., Martin-Hernandez, R., Serrano, M., et al. (2017). Genome-wide expression profiling in muscle and subcutaneous fat of lambs in response to the intake of concentrate supplemented with vitamin E. *BMC Genomics* 18 (1), 92. doi:10.1186/s12864-016-3405-8

Grigoletto, L., Brito, L. F., Mattos, E. C., Eler, J. P., Bussiman, F. O., da Conceição Abreu Silva, B., et al. (2019). Genome-wide associations and detection of candidate genes for direct and maternal genetic effects influencing growth traits in the Montana Tropical® composite population. *Livest. Sci.* 229, 64–76. doi:10.1016/j.livsci.2019.09.013

Han, Y., and Peñagaricano, F. (2016). Unravelling the genomic architecture of bull fertility in Holstein cattle. *BMC Genet.* 17 (1), 143. doi:10.1186/s12863-016-0454-6

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

Hay, E. H., Toghiani, S., Roberts, A. J., Paim, T., Alexander Kuehn, L., and Blackburn, H. D. (2022). Genetic architecture of a composite beef cattle population. *J. Animal Sci.* 100 (9), skac230. doi:10.1093/jas/skac230

Heras-Saldana, S., Clark, S. A., Duijvesteijn, N., Gondro, C., van der Werf, J. H. J., and Chen, Y. (2019). Combining information from genome-wide association and multi-tissue gene expression studies to elucidate factors underlying genetic variation for residual feed intake in Australian Angus cattle. *BMC Genomics* 20, 939. doi:10.1186/s12864-019-6270-4

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). Blink: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* 8 (2), giy154. doi:10.1093/gigascience/giy154

Irshad, A., Kandeepan, G., Kumar, S., Ashish, K. A., Vishnuraj, M. R., Shukla, V., et al. (2013). Factors influencing carcass composition of Livestock: A review. *J. Animal Prod. Adv.* 3 (5), 177–186. doi:10.5455/JAPA.20130531093231

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

Kaler, A. S., Gillman, J. D., Beissinger, T., and Purcell, L. C. (2020). Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Front. Plant Sci.* 10, 1794. doi:10.3389/fpls.2019.01794

Karimi, K., Koshkoiyeh, A. E., and Gondro., C. (2015). Comparison of linkage disequilibrium levels in Iranian indigenous cattle using whole genome SNPs data. *J. Animal Sci. Technol.* 57, 1–10. doi:10.1186/s40781-015-0080-2

Keogh, K., Carthy, T. R., McClure, M. C., Waters, S. M., and Kenny, D. A. (2021). Genome-wide association study of economically important traits in charolais and limousin beef cows. *Animal* 15 (1), 100011. doi:10.1016/j.animal.2020.100011

Kim, Y. S., Fukumoto, G. K., and Kim, S. (2012). Carcass quality and meat tenderness of Hawaii pasture-finished cattle and Hawaii-originated, mainland feedlot-finished cattle. *Trop. Animal Health Prod.* 44 (7), 1411–1415. doi:10.1007/s11250-012-0080-x

Kristensen, P. S., Jahoor, A., Andersen, J., Cericola, R. F., Orabi, J., Janss, L. L., et al. (2018). Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. *Front. Plant Sci.* 9, 69. doi:10.3389/fpls.2018.00069

Lee, S. H., Choi, B. H., Lim, D., Gondro, C., Cho, Y. M., Dang, C. G., et al. (2013). Genome-wide association study identifies major loci for carcass weight on BTA14 in hanwoo (Korean cattle). *PloS One* 8 (10), e74677. doi:10.1371/journal.pone.0074677

Lee, S-H., Park, B-H., Sharma, A., Dang, C-G., Lee, S-S., Choi, T-J., et al. (2014). Hanwoo cattle: Origin, domestication, breeding strategies and genomic selection. *J. Animal Sci. Technol.* 56 (1), 2. doi:10.1186/2055-0391-56-2

Lindholm-Perry, A. K., Kuehn, L. A., Smith, T. P. L., Ferrell, C. L., Jenkins, T. G., Freetly, H. C., et al. (2012). A region on BTA14 that includes the positional candidate genes LYPLA1, XKR4 and TMEM68 is associated with feed intake and growth phenotypes in cattle 1. *Anim. Genet.* 43 (2), 216–219. doi:10.1111/j.1365-2052.2011.02232.x

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12 (2), e1005767. doi:10.1371/journal.pgen.1005767

Melrose, J., Perroy, R., and Cares, S.University of Hawaii at Hilo; Spatial Data Analysis & Visualization Research Lab (2015). *Statewide agricultural land use baseline*. Hawai'i, United States: Hawaii State Department of Agriculture. Available at: https://hdoa.hawaii.gov/wp-content/uploads/2016/02/StateAgLandUseBaseline2015.pdf.

Mengistie, D., Edea, Z., T s TesemaDejene, G., Dessie, T., Jemal, J., et al. (2022). *Genome-wide signature of positive selection and linkage disequilibrium in Ethiopian indigenous and European beef cattle breeds*. North Carolina, United States: Research Square. doi:10.21203/rs.3.rs-1554212/v1

Merrick, L. F., Burke, A. B., Zhang, Z., and Carter, A. H. (2022). Comparison of single-trait and multi-trait genome-wide association models and inclusion of correlated traits in the dissection of the genetic architecture of a complex trait in a breeding program. *Front. Plant Sci.* 12, 772907. doi:10.3389/fpls.2021.772907

Miao, C., Yang, J., and Schnable, J. C. (2019). Optimising the identification of causal variants across varying genetic architectures in crops. *Plant Biotechnol. J.* 17 (5), 893–905. doi:10.1111/pbi.13023

National Agricultural Statistics Service, USDA (2021). National agricultural statistics Service, USDA. United States: USDA. Available at: http://quickstats.nass.usda.gov/.

Pedrosa, V. B., Schenkel, F. S., Chen, S-Y., Oliveira, H. R., Casey, T. M., Melka, M. G., et al. (2021). Genomewide association analyses of lactation persistency and milk production traits in Holstein cattle based on imputed whole-genome sequence data. *Genes* 12 (11), 1830. doi:10.3390/genes12111830

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–909. doi:10.1038/ng1847

Rolf, M. M., Taylor, J. F., Schnabel, R. D., McKay, S. D., McClure, M. C., Northcutt, S. L., et al. (2012). Genome-wide association analysis for feed efficiency in Angus cattle. *Anim. Genet.* 43 (4), 367–374. doi:10.1111/j.1365-2052.2011.02273.x

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9 (3), giaa021. doi:10.1093/gigascience/giaa021

Silva, R. M. O., Stafuzza, N. B., Oliveira Fragomeni, B. de, F de Camargo, G. M., Ceacero, T. M., dos Santos Gonçalves Cyrillo, J. N., et al. (2017). Genome-wide association study for carcass traits in an experimental nelore cattle population. *PLOS ONE* 12 (1), e0169860. doi:10.1371/journal.pone.0169860

Singh, A., Kumar, A., Mehrotra, A., Pandey, A. K., Mishra, B. P., Dutt, T., et al. (2021). Estimation of linkage disequilibrium levels and allele frequency distribution in crossbred vrindavani cattle using 50K SNP data. *Plos One* 16 (11), e0259572. doi:10.1371/journal.pone.0259572

Smith, J. L., Wilson, M. L., Nilson, S. M., Rowan, T. N., Oldeschulte, D. L., Schnabel, R. D., et al. (2019). Genome-wide association and genotype by environment interactions for growth traits in US gelbvieh cattle. *BMC Genomics* 20 (1), 926–1013. doi:10.1186/s12864-019-6231-y

Srikanth, K., Lee, S-H., Chung, K-Y., Park, J-E., Jang, G-W., Park, M-R., et al. (2020). A gene-set enrichment and protein–protein interaction network-based GWAS with regulatory SNPs identifies candidate genes and pathways associated with carcass traits in hanwoo cattle. *Genes* 11 (3), 316. doi:10.3390/genes11030316

Steel, R. G. D. (1997). *Analysis of variance II: Multiway classifications*. McGraw-Hill, NY: Principles and Procedures of Statistics: A Biometrical Approach, 204–252.

Stich, B., Möhring, J., Piepho, H-P., Heckenberger, M., Buckler, E. S., and E Melchinger, A. (2008). Comparison of mixed-model approaches for association mapping. *Genetics* 178 (3), 1745–1754. doi:10.1534/genetics.107.079707

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20 (8), 467–484. doi:10.1038/s41576-019-0127-1

Tamba, C. L., Ni, Y-L., and Zhang, Y-M. (2017). Iterative sure independence screening EM-bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13 (1), e1005357. doi:10.1371/journal.pcbi.1005357

Tang, Z., Xu, J., Yin, L., Yin, D., Zhu, M., Yu, M., et al. (2019). Genome-wide association study reveals candidate genes for growth relevant traits in pigs. *Front. Genet.* 10, 302. doi:10.3389/fgene.2019.00302

Turpin, S. M., Nicholls, H. T., Willmes, D. M., Mourier, A., Brodesser, S., Wunderlich, C. M., et al. (2014). Obesity-induced CerS6-dependent C16:0 ceramide production promotes weight gain and glucose intolerance. *Cell Metab.* 20 (4), 678–686. doi:10.1016/j.cmet.2014.08.002

Van Tassell, C. P., L Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., et al. (2008). SNP Discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5 (3), 247–252. doi:10.1038/nmeth.1185

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130, 123–135. doi:10.1007/s00122-016-2798-8

Wang, J., and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Bioinforma. Commons* 19 (4), 629–640. doi:10.1016/j.gpb.2021.08.005

Weller, J. I., Ezra, E., and Ron, M. (2017). Invited review: A perspective on the future of genomic selection in dairy cattle. *J. Dairy Sci.* 100 (11), 8633–8644. doi:10.3168/jds.2017-12879

Wen, Y-J., Zhang, H., Ni, Y-L., Huang, B., Zhang, J., Feng, J-Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings Bioinforma.* 19 (4), 700–712. doi:10.1093/bib/bbw145

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, X., Jiang, A., et al. (2020). Whole-genome sequencing association analysis reveals the genetic architecture of meat quality traits in Chinese Qingyu pigs. *Genome* 63 (10), 503–515. doi:10.1139/gen-2019-0227

Xu, S-S., Gao, L., Shen, M., and Lyu, F. (2021). Whole-genome selective scans detect genes associated with important phenotypic traits in sheep (Ovis aries). *Front. Genet.* 12, 738879. doi:10.3389/fgene.2021.738879

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi:10.1038/ng1702

Zhang, R., Miao, J., Song, Y., Zhang, W., Xu, L., Chen, Y., et al. (2019). Genome-wide association study identifies the PLAG1-OXR1 region on BTA14 for carcass meat yield in cattle. *Physiol. Genomics* 51 (5), 137–144. doi:10.1152/physiolgenomics.00112.2018

Zhao, F., McParland, S., Kearney, F., Du, L., and Berry, D. P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genet. Sel. Evol.* 47 (1), 49. doi:10.1186/s12711-015-0127-3

TYPE Original Research
PUBLISHED 16 August 2023
DOI 10.3389/fvets.2023.1184764

OPEN ACCESS

EDITED BY
Kefei Chen,
Curtin University, Australia

REVIEWED BY
Peipei Ma,
Shanghai Jiao Tong University, China
Doreen Becker,
Research Institute for Farm Animal Biology (FBN), Germany

*CORRESPONDENCE
Shi-Yi Chen
✉ sychensau@gmail.com;
✉ chensy@sicau.edu.cn
Hanzhong Liu
✉ hanzhong6688@sina.com

RECEIVED 12 March 2023
ACCEPTED 31 July 2023
PUBLISHED 16 August 2023

CITATION
Zhang K, Wang G, Wang L, Wen B, Fu X, Liu N, Yu Z, Jian W, Guo X, Liu H and Chen S-Y (2023) A genome-wide association study of coat color in Chinese Rex rabbits. *Front. Vet. Sci.* 10:1184764. doi: 10.3389/fvets.2023.1184764

A genome-wide association study of coat color in Chinese Rex rabbits

Kai Zhang[1], Guozhi Wang[2], Lihuan Wang[1], Bin Wen[1], Xiangchao Fu[1], Ning Liu[1], Zhiju Yu[1], Wensu Jian[1], Xiaolin Guo[1], Hanzhong Liu[1]*and Shi-Yi Chen[2]*

[1]Sichuan Academy of Grassland Sciences, Chengdu, Sichuan, China, [2]Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, Sichuan, China

Coat color is an important phenotypic characteristic of the domestic rabbit (*Oryctolagus cuniculus*) and has specific economic importance in the Rex rabbit industry. Coat color varies considerably among different populations of rabbits, and several causal genes for this variation have been thoroughly studied. Nevertheless, the candidate genes affecting coat color variation in Chinese Rex rabbits remained to be investigated. In this study, we collected blood samples from 250 Chinese Rex rabbits with six different coat colors. We performed genome sequencing using a restriction site-associated DNA sequencing approach. A total of 91,546 single nucleotide polymorphisms (SNPs), evenly distributed among 21 autosomes, were identified. Genome-wide association studies (GWAS) were performed using a mixed linear model, in which the individual polygenic effect was fitted as a random effect. We detected a total of 24 significant SNPs that were located within a genomic region on chromosome 4 (OCU4). After re-fitting the most significant SNP (OCU4:13,434,448, $p = 1.31e\text{-}12$) as a covariate, another near-significant SNP (OCU4:11,344,946, $p = 7.03e\text{-}07$) was still present. Hence, we conclude that the 2.1-Mb genomic region located between these two significant SNPs is significantly associated with coat color in Chinese Rex rabbits. The well-studied coat-color-associated agouti signaling protein (*ASIP*) gene is located within this region. Furthermore, low genetic differentiation was also observed among the six coat color varieties. In conclusion, our results confirmed that *ASIP* is a putative causal gene affecting coat color variation in Chinese Rex rabbits.

KEYWORDS

Rex rabbits, coat color, GWAS, RAD-seq, genome

Introduction

Among all farm animals, modern rabbits (*Oryctolagus cuniculus*) are among the most recently domesticated species, although the exact domestication date of the species remains controversial when examined on the basis of archeological records and genetic evidence (1, 2). However, it has been widely acknowledged that modern rabbits have a single domestication origin, resulting in lower genetic variation in comparison with other farm animals (3–5). More than 200 rabbit breeds have been officially registered in the Domestic Animal Diversity

Information System (DAD-IS),[1] and these show considerable morphological variation, such as in body size, coat color, and hair phenotype (6, 7). Among them, the Rex rabbit is well known for its short, dense, and smooth hair. This Rex rabbit phenotype is believed to have genetically originated from normal hair (8). Coat color is an important phenotypic trait in the fur industry, and at least 16 color varieties of Rex rabbits have been recognized by the American Rabbit Breeders Association (ARBA);[2] however, the preferred coat color differs between different markets.

Coat color in mammals is determined by the relative amounts of eumelanin and phaeomelanin in melanocytes, and many studies have been conducted to identify coat-color-associated genes and causal mutations in domestic animals during the past two decades (9). In domestic rabbits, melanocortin 1 receptor (*MC1R*) is the first gene to have been thoroughly studied, and several causal mutations have been successfully identified as affecting coat color (10, 11). As a competitive ligand to *MC1R* in the melanin synthesis pathway (12), a premature stop mutation of the agouti signaling protein (*ASIP*) gene has been reported to be responsible for the non-agouti black coat color in rabbits (13). Additionally, a premature stop mutation of the tyrosinase-related protein 1 (*TYRP1*) gene is associated with brown coat color in rabbits (14). Based on the gene expression patterns observed in Rex rabbits of various colors, it has been suggested that the POU class 2 homeobox 1 gene (*POU2F1*) affects fur color formation in Rex rabbits (15). The variability of the tyrosinase (*TYR*) gene has been studied in domestic and wild European rabbits; this work has confirmed the effects of missense mutations on coat colors (16). The genetic polymorphisms of five candidate genes were genotyped to investigate their associations with different coat colors in rabbits (17). In addition to loci determining different coat colors, it was found that both eumelanic and pheomelanic pigmentations can be further diluted more or less under genetical control of the *dilute* locus: for example, black can be diluted to gray. Fontanesi et al. (18) successfully mapped the *dilute* locus of rabbits to the melanophilin (*MLPH*) gene and identified a frameshift mutation associated with the dilute coat color.

Due to the wide application of high-throughput sequencing technologies, large numbers of genome-wide variants can now be discovered and genotyped at an affordable cost (19). Among these technologies, restriction-site-associated DNA sequencing (RAD-seq) is a cost-efficient approach for investigating genome-wide variants, especially in non-model species (20); the approach was first proposed in 2008 and is characterized by sequencing of small genomic fragments that are randomly digested by restriction enzyme (s). The RAD-seq approach has been widely used for population genetics and genome-wide association studies (GWAS) [such as by (21–24)]. In rabbits, genetic diversity and population structure have been investigated using genome-wide single nucleotide polymorphisms (SNPs) that have been generated using the RAD-seq approach (25, 26). In this study, we similarly employed the RAD-seq approach to identify genome-wide SNPs, which we subsequently used for GWAS with six coat color varieties of Chinese Rex rabbits. The results could help us to better understand the underlying genetic basis of coat colors in Chinese Rex rabbits.

Materials and methods

Animals and genomic DNA

Venous blood was collected from the marginal ear veins of 250 Rex rabbits raised at the Research Farm of Sichuan Academy of Grassland Sciences. The rabbits consisted of six coat color varieties: 40 White Rex (WT), 42 Californian Rex (CL), 42 Black Rex (BL), 42 Chinchilla Rex (CC), 42 Dark Chinchilla Rex (DC), and 42 Light Chinchilla Rex (LC). Among these varieties, WT, CL, BL, and CC exhibit different coat colors, and there are two varieties of CC with a darker (DC) and lighter (LC) coat color, respectively (Figure 1). There was no genetic relationship within three generations among any of the sampled animals according to pedigree information. Genomic DNA was extracted using the Axy-Prep Genomic DNA Miniprep Kit (Axygen Bioscience, USA).

Genome sequencing

Based on preliminary investigation on the reference genome sequences of rabbits, the restriction enzyme *EcoRI* (NEB, Beijing) was successfully used to digest genomic DNA (~1 μg per sample used). Sequencing libraries were constructed according to the recommended pipeline (20). In brief, P1 adapter sequence was first added to the digested fragments; this was followed by sequential steps of sample pooling, random shearing, and fragment size-based selection using agarose gel. Subsequently, DNA was ligated to a second adapter (P2) with divergent ends. DNA fragments of ~400 bp in length were selected to construct the sequencing libraries. Finally, the libraries were sequenced on an Illumina HiSeq platform and 150 bp paired-end reads were generated (Novogene Co. Ltd., Beijing).

Reads mapping and SNP genotyping

After the initial sequencing images were converted into sequence files in the FASTQ format using a standard pipeline, we first investigated the Q_{phred} value-based error rate. Using the fastp software package (27), low-quality reads were discarded according to three criteria (26): (1) reads containing adaptor sequences, (2) reads containing ambiguous bases for more than 10% of the total length, and (3) reads containing low-quality bases (Q_{phred} value <5) for more than 50% of the total length. If either member of the paired reads was marked as low quality, both pairs were discarded. After these steps, we obtained clean reads and subjected them to the following analyses.

All clean reads were mapped to a rabbit reference genome (UM_NZW_1.0) using the BWA software with default parameters (28). Subsequently, we employed the GATK toolkit v3.8 (29) for discovery and genotyping of small variants (SNPs and InDels) among all samples according to GATK Best Practices recommendations (30, 31); in this process, the duplicate removal, realignment, and hard filtering steps were performed with default parameters. After exclusion of all InDels, a raw set of SNPs was obtained. SNPs were removed if they had a coverage depth < 3, calling rate < 90% for the genotypes or individuals, minor allele frequency (MAF) < 0.05, and extreme deviation from Hardy–Weinberg equilibrium (HWE, $p < 10^{-8}$). Finally, we extracted biallelic SNPs and generated a clean set of SNPs. The missing

[1] https://www.fao.org/dad-is

[2] www.arba.net

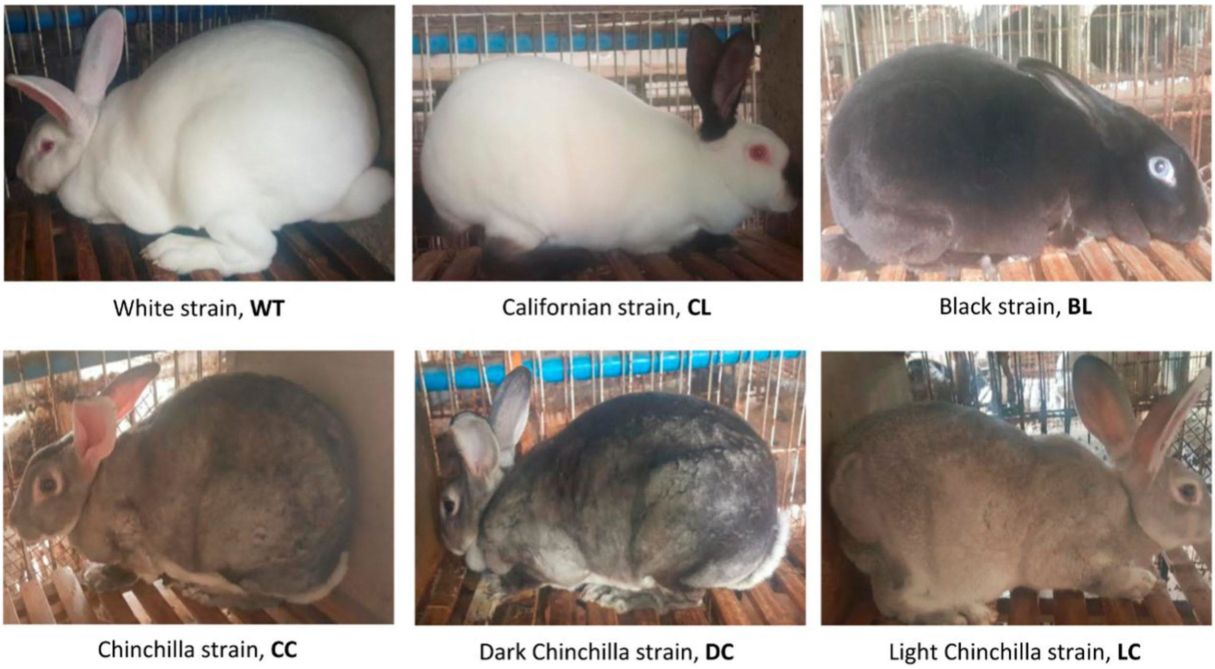

White strain, **WT**
Californian strain, **CL**
Black strain, **BL**
Chinchilla strain, **CC**
Dark Chinchilla strain, **DC**
Light Chinchilla strain, **LC**

FIGURE 1
Phenotypes of the six coat color varieties of Rex rabbit included in this study.

genotypes were further imputed using the Beagle software package v5.4 with default parameters (32).

Population and association analyses

We investigated the genomic distribution of clean SNPs and the transition/transversion ratio using the ANNOVAR software package (33). Nucleotide diversity (π) for each locus was calculated using the vcftools software package (34). The PopSc toolkit (35) was used to calculate the polymorphism information content (PIC), inter-variety Wright's F_{ST}, and intra-variety Wright's F_{IS} (36). The pairwise *p*-distances among all samples were calculated from all SNPs using the TreeBeST software package (TreeSoft) and then subjected to the construction of a phylogenetic tree according to the neighbor-joining method (37); this phylogenetic tree was visualized using the ggtree R package (38).

In GWAS, the six coat color varieties of Rex rabbits were arbitrarily coded using the ordinal values of WT = 1, LC = 2, CC = 3, DC = 4, CL = 5, and BL = 6. To avoid potential bias arising from the arbitrary coding of coat colors, the reverse order was employed in independent repeat performance of GWAS. The effect of each SNP was estimated using a mixed linear model implemented in the GCTA software package (39):

$$y = 1 + Z\beta + W\mu + e$$

where y is the vector of coat colors coded above; 1 is the mean term; β is the fixed effect of the SNP tested for association; Z is a vector containing the genotype score for the tested SNP; μ is the vector of individual random polygenic effects with $\mu \sim N\left(0, G\sigma_u^2\right)$, where G is the genomic relationship matrix and σ_u^2 is the additive genetic variance; W is the incidence matrix for μ; and e is a vector of random residual effects with $e \sim N\left(0, I\sigma_e^2\right)$, where I is an identity matrix and σ_e^2 is the residual variance. After estimation of the SNP effects, the most significant SNP was selected and further added as a covariate to the mixed linear model described above. A Bonferroni approach was used for correction of multiple comparisons in the GWAS results (40).

Results

Sequencing and SNPs

We obtained 208.51 Gb raw paired-end reads (approximately 1.5 billion reads) across all the sequenced samples, from which 208.48 Gb clean reads (0.83 Gb per sample) were generated after the quality control steps. On average, 98.9% of the clean reads were successfully aligned against the reference genome. A total of 5,162,522 raw SNPs were generated on 21 autosomes, and we finally obtained 91,546 high-quality biallelic SNPs according to our custom filtering process. These SNPs were distributed across the whole genome, and an average of 42.3 SNPs per Mb genomic region was comparably observed among all autosomes (Figure 2A). The mean MAF was ~0.25 (Figure 2B). There were 64,311 transitions and 27,235 transversions (transition/transversion ratio = 2.36). Using the reference annotation of the rabbit genome (UM_NZW_1.0), we inferred the locations of the SNPs. SNPs were distributed within exons (*N* = 1,648), introns (*N* = 34,037), 1 kb upstream/downstream regions of genes (*N* = 1,411), and intergenic regions (*N* = 54,450).

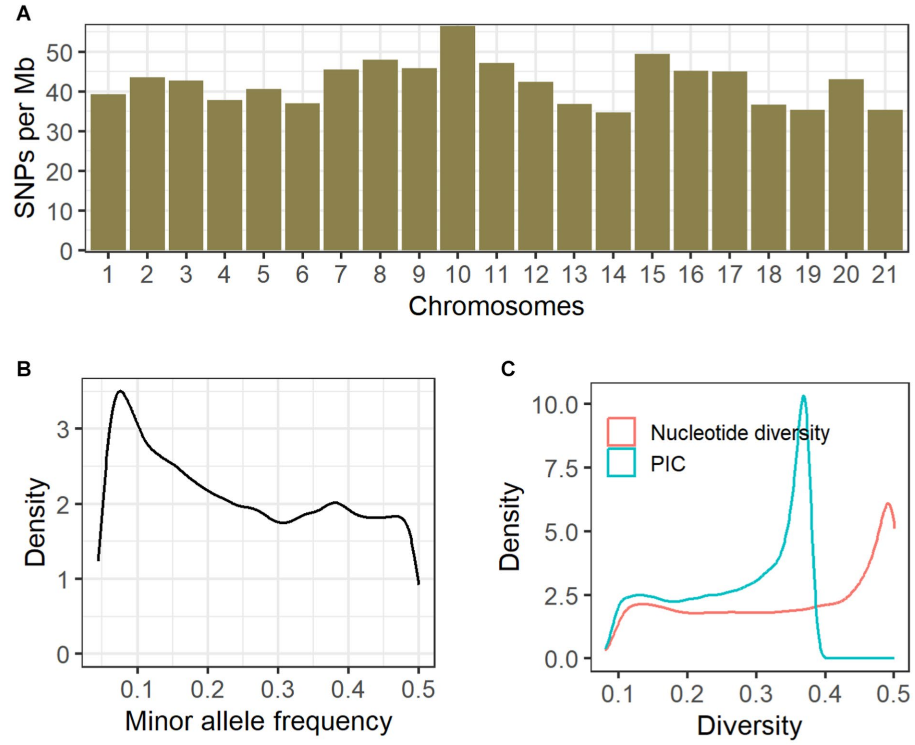


FIGURE 2
Genomic distribution and genetic diversity of SNPs. For all clean SNPs, we investigated the genomic distribution **(A)**, minor allele frequencies **(B)**, and the density distribution of nucleotide diversity and polymorphism information content **(C)**.

Genetic diversity and population structure

Among all clean SNPs, the distribution density of nucleotide diversity exhibited a single peak close to 0.5, and a similar pattern was also observed for PIC (Figure 2C). The median and mean values for nucleotide diversity across the six coat color varieties were 0.3518 and 0.3185, respectively (Table 1); among these, the Black Rex showed the highest degree of nucleotide diversity, with a median of 0.3672 and a mean of 0.3340. The Black Rex and Californian Rex had the highest and lowest PIC, with mean values of 0.2649 and 0.2440, respectively. Furthermore, there were no obvious differences among the coat color varieties in relation to genetic diversity.

The highest and lowest degrees of inter-variety differentiation were observed between the Dark Chinchilla Rex and the Californian Rex ($F_{ST} = 0.0962$), and between the Chinchilla Rex and the Dark Chinchilla Rex ($F_{ST} = -0.0002$), respectively (Figure 3A). Intra-variety inbreeding coefficients (F_{IS}) ranged from −0.1221 in the Californian Rex to −0.0522 in the Black Rex. According to the phylogenetic tree for all samples (Figure 3B), both the White Rex and the Californian Rex formed their own clusters and were separated from the other breeds. Next, most of the Black Rex rabbits were clustered together and were almost distinguishable. However, there was no obvious clustering pattern among the Chinchilla Rex and the other two breeds.

TABLE 1 Nucleotide diversity (π) and polymorphism information content (PIC) in different coat color varieties of Rex rabbit.

Coat color variety	π		PIC	
	Median	Mean	Median	Mean
White Rex	0.3532	0.3236	0.2879	0.2575
Californian Rex	0.3408	0.3069	0.2800	0.2440
Black Rex	0.3672	0.3340	0.2970	0.2649
Chinchilla Rex	0.3543	0.3169	0.2888	0.2520
Dark Chinchilla Rex	0.3408	0.3116	0.2800	0.2478
Light Chinchilla Rex	0.3543	0.3179	0.2888	0.2526
Overall	**0.3518**	**0.3185**	**0.2871**	**0.2531**

Association with coat colors

The association analysis results are shown in Figure 4. A total of 24 SNPs were detected as statistically significant; all of these were located within a 3.01-Mb genomic region on chromosome 4 (OCU4). After fitting the most significant SNP (OCU4:13,434,448; $p = 1.31\text{e-}12$) as a covariate, the association signal within this region noticeably decreased, but it still almost reached the threshold for significance (OCU4:11,344,946; $p = 7.03\text{e-}07$). The allelic frequencies of the two SNPs within each population are shown in Table 2; notably, OCU4:13,434,448 was completely fixed in the three non-Chinchilla populations. When both SNPs (OCU4:13,434,448 and OCU4:11,344,946) were simultaneously fitted as covariates, there was no longer any significant association signal within this region. Upon reverse-coding of the coat color, the association results did not change noticeably (Supplementary Figure S1).

We further investigated the annotated genes within this candidate genomic region (including 500 kb upstream of OCU4:11,344,946 and

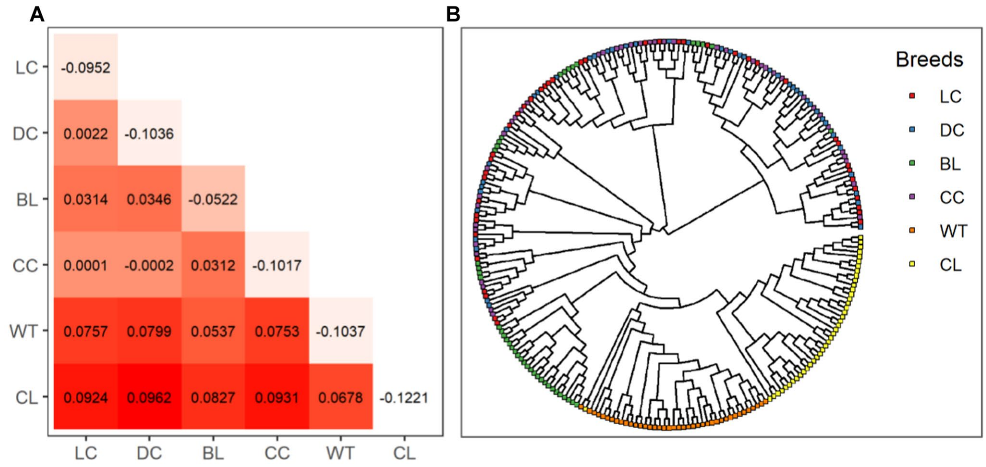


FIGURE 3
Genetic structures among the six coat color varieties of Rex rabbit. The matrix **(A)** shows pairwise Wright's F_{ST} values in the lower triangle and F_{IS} values in diagonal cells. The phylogenetic tree for all 250 animals is shown in **(B)**. WT, White Rex; CL, Californian Rex; BL, Black Rex; CC, Chinchilla Rex; DC, Dark Chinchilla Rex; LC, Light Chinchilla Rex.

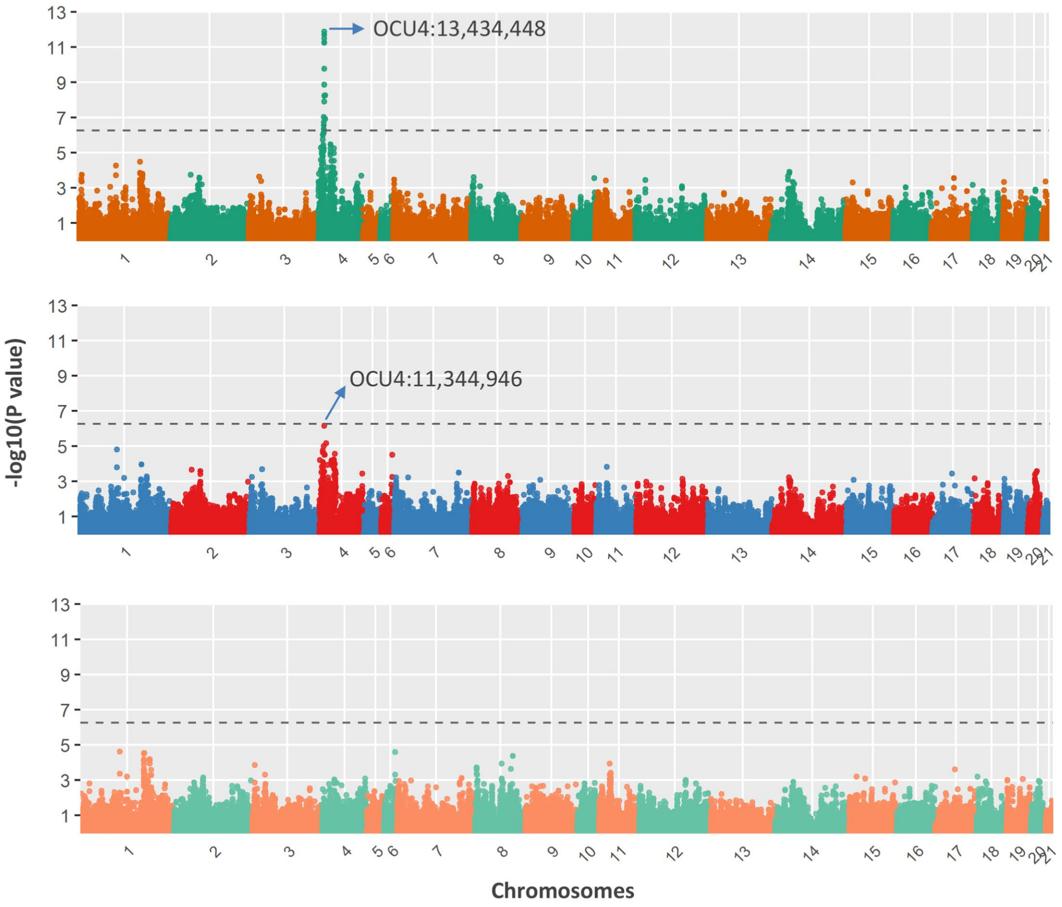


FIGURE 4
Genome-wide association with coat colors of Chinese Rex rabbits. After testing all SNP effects with a mixed linear model (top panel), the most significant SNP (OCU4:13,434,448) was fitted as a covariate for re-testing of SNP effects (middle panel). Both significant SNPs (OCU4:13,434,448 and OCU4:11,344,946) were simultaneously fitted as covariates for re-testing of SNP effects (bottom panel). The dashed line represents the genome-wide significance threshold.

TABLE 2 Frequencies of reference alleles for the two (near-) significant SNPs.

Coat color variety	OCU4:13,434,448 (T > A)	OCU4:11,344,946 (G > A)
White Rex	1.00	0.64
Californian Rex	1.00	0.98
Black Rex	1.00	0.82
Chinchilla Rex	0.73	0.45
Dark Chinchilla Rex	0.92	0.44
Light Chinchilla Rex	0.55	0.54

500 kb downstream of OCU4:13,434,448), identifying 51 positional candidate genes in total. Among these genes, the well-studied *ASIP* gene, which is significantly associated with agouti coat color in rabbits, was located 150 kb upstream of the most significant SNP (OCU4:13,434,448). However, the second-most independently significant SNP (OCU4:11,344,946) was located at a large distance, 1.7 Mb upstream of the *ASIP* gene. Besides *ASIP*, no other known coat-color-associated gene was found within this candidate genomic region.

Discussion

Coat color is an important phenotypic characteristic in domestic animals and has been directly subjected to artificial selection (41). It has also been proposed that hundreds of loci/genes play a role in affecting coat color, which (in combination with diverse selection preferences among humans) has ultimately resulted in considerable variation in a wide range of domestic animals (9). In addition to being farmed for the production of meat, wool, and fur, modern rabbits have been kept as a pet animal worldwide, with specific emphasis on the subjective selection of coat color. Therefore, rabbits could represent an ideal case for the identification of candidate genes and causal mutations affecting the expression of different coat colors. With the use of a cost-efficient method, genome-wide genetic variants could be discovered *de novo* through implementation high-throughput surveys, such as GWAS, for economically important traits and for the investigation of population genetic structures. In this study, we collected six coat color varieties of Rex rabbits raised in China and employed a high-throughput approach to successfully identify genome-wide and evenly distributed SNPs.

Coat color in mammals is generally considered to be a qualitative trait, although the phenotypic variations are genetically determined by polygenes. Therefore, the genome-wide scanning approach has been increasingly widely used to reveal coat-color-associated candidate genes and causal mutations. For example, Li et al. (42) genotyped ~50 k SNPs and employed a GWAS approach to identify three known pigmentation genes in sheep. In the Iranian Markhoz goat, a total of six genes have been identified as being associated with black, brown, and white coat colors using a GWAS approach (43). Based on the newly discovered SNPs in this study, we also conducted the first GWAS for coat color in Chinese Rex rabbits. Our results revealed that a 2.1-Mb genomic region (OCU4:11,344,946 – 13,434,448) containing *ASIP*, which has been shown in previous studies to be significantly associated with coat color (13), is also significantly associated with coat color in Rex rabbits. In a previous study of Rex rabbits with different coat colors, Yang et al. (44) found that *ASIP* had three alleles and was extensively expressed in all analyzed tissues. Recently, an 11-kb deletion spanning the promoter and first exon of *ASIP* has been suggested to be the most likely causal variant for the black-and-tan phenotype in rabbits (45). In the present study, we confirmed that *ASIP* is a putative causal gene affecting coat color in Chinese Rex rabbits. In the melanocytes of the hair follicle, *ASIP* encodes a paracrine signaling molecule that promotes the synthesis of pheomelanin (46). However, further studies are needed to explore whether the two candidate SNPs identified in this study are causal variants or not; although both of them are located more than 100 kb away from *ASIP* (upstream and downstream), possible roles for these SNPs in regulating gene expression cannot be excluded. Another possibility is that the two candidate SNPs are closely linked to the potential causal variant(s).

In addition to the discovery of coat-color-associated candidate genes, both genetic diversity and population structures among the six coat color varieties of Rex rabbits were investigated using the set of genome-wide SNPs generated in this study. Our results revealed the differential genetic diversity among these coat color varieties, with the highest genetic diversity observed in the Black Rex. This result is consistent with those presented in a previous report on genetic diversity patterns among 29 domestic and wild rabbit populations, examined using microsatellite markers (47). Liu et al. (25) also investigated population structure among eight Chinese rabbit breeds (not including the Rex rabbit), whose F_{ST} values were significantly higher than our estimates in this study; this may suggest that genetic differentiation among different populations of Rex rabbits is relatively low in comparison with other indigenous breeds. In accordance with this possibility, less inter-variety genetic differentiation was observed, with lower Fst values, than in former reports (47, 48). Meanwhile, our clustering analysis revealed that only individuals of the White and Californian Rex rabbit varieties could be clustered together and distinguished from individuals of other varieties. Overall, our results revealed using genome-wide SNP information that there is low genetic differentiation among different coat color varieties of Chinese Rex rabbits.

Conclusion

In this study, we discovered a genome-wide set of SNPs for Chinese Rex rabbits and used these to perform association analyses for the coat color phenotype. Our results revealed a single genomic region that is significantly associated with Rex coat color, and confirmed that the previously known coat-color-associated gene *ASIP* is a putative causal gene affecting coat color variation in Chinese Rex rabbits. Furthermore, low genetic differentiation was revealed among the six coat color varieties of Rex rabbit studied.

Data availability statement

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

Ethics statement

The animal study was approved by the Animal Care and Use Committee of Sichuan Academy of Grassland Sciences (YTS20B03). The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

KZ, HL, and S-YC: conceptualization. KZ and GW: formal analysis. LW, BW, XF, NL, ZY, WJ, and XG: resources. KZ: writing—original draft preparation. HL and S-YC: writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the Earmarked Fund for the China Agriculture Research System (CARS-43-A-3) and Science and Technology Department of Sichuan Province (2021YFYZ0033 and 2021YFYZ0009).

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.2023.1184764/full#supplementary-material

References

1. Carneiro M, Afonso S, Geraldes A, Garreau H, Bolet G, Boucher S, et al. The genetic structure of domestic rabbits. *Mol Biol Evol*. (2011) 28:1801–16. doi: 10.1093/molbev/msr003

2. Irving-Pease EK, Frantz LAF, Sykes N, Callou C, Larson G. Rabbits and the specious origins of domestication. *Trends Ecol Evol*. (2018) 33:149–52. doi: 10.1016/j.tree.2017.12.009

3. Frantz LA, Bradley DG, Larson G, Orlando L. Animal domestication in the era of ancient genomics. *Nat Rev Genet*. (2020) 21:449–60. doi: 10.1038/s41576-020-0225-0

4. Lai SJ, Liu YP, Liu YX, Li XW, Yao YG. Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. *Mol Phylogenet Evol*. (2006) 38:146–54. doi: 10.1016/j.ympev.2005.06.013

5. Liu YP, Cao SX, Chen SY, Yao YG, Liu TZ. Genetic diversity of Chinese domestic goat based on the mitochondrial DNA sequence variation. *J Anim Breed Genet*. (2009) 126:80–9. doi: 10.1111/j.1439-0388.2008.00737.x

6. Dorożyńska K, Maj D. Rabbits-their domestication and molecular genetics of hair coat development and quality. *Anim Genet*. (2021) 52:10–20. doi: 10.1111/age.13024

7. Fontanesi L. Rabbit genetic resources can provide several animal models to explain at the genetic level the diversity of morphological and physiological relevant traits. *Appl Sci*. (2021) 11:373. doi: 10.3390/app11010373

8. Diribarne M, Mata X, Chantry-Darmon C, Vaiman A, Auvinet G, Bouet S, et al. A deletion in exon 9 of the LIPH gene is responsible for the rex hair coat phenotype in rabbits (*Oryctolagus cuniculus*). *PLoS One*. (2011) 6:e19281. doi: 10.1371/journal.pone.0019281

9. Linderholm A, Larson G. The role of humans in facilitating and sustaining coat colour variation in domestic animals. *Semin Cell Dev Biol*. (2013) 24:587–93. doi: 10.1016/j.semcdb.2013.03.015

10. Fontanesi L, Scotti E, Colombo M, Beretti F, Forestier L, Dall'olio S, et al. A composite six bp in-frame deletion in the melanocortin 1 receptor (MC1R) gene is associated with the Japanese brindling coat colour in rabbits (*Oryctolagus cuniculus*). *BMC Genet*. (2010) 11:59. doi: 10.1186/1471-2156-11-59

11. Fontanesi L, Tazzoli M, Beretti F, Russo V. Mutations in the melanocortin 1 receptor (MC1R) gene are associated with coat colours in the domestic rabbit (*Oryctolagus cuniculus*). *Anim Genet*. (2006) 37:489–93. doi: 10.1111/j.1365-2052.2006.01494.x

12. Lu D, Willard D, Patel IR, Kadwell S, Overton L, Kost T, et al. Agouti protein is an antagonist of the melanocyte-stimulating-hormone receptor. *Nature*. (1994) 371:799–802. doi: 10.1038/371799a0

13. Fontanesi L, Forestier L, Allain D, Scotti E, Beretti F, Deretz-Picoulet S, et al. Characterization of the rabbit agouti signaling protein (ASIP) gene: transcripts and phylogenetic analyses and identification of the causative mutation of the nonagouti black coat colour. *Genomics*. (2010) 95:166–75. doi: 10.1016/j.ygeno.2009.11.003

14. Utzeri VJ, Ribani A, Fontanesi L. A premature stop codon in the TYRP 1 gene is associated with brown coat colour in the European rabbit (*Oryctolagus cuniculus*). *Anim Genet*. (2014) 45:600–3. doi: 10.1111/age.12171

15. Yang N, Zhao B, Hu S, Bao Z, Liu M, Chen Y, et al. Characterization of POU2F1 gene and its potential impact on the expression of genes involved in fur color formation in rex rabbit. *Genes*. (2020) 11:575. doi: 10.3390/genes11050575

16. Utzeri VJ, Ribani A, Schiavo G, Fontanesi L. Describing variability in the tyrosinase (TYR) gene, the albino coat colour locus, in domestic and wild European rabbits. *Ital J Anim Sci*. (2021) 20:181–7. doi: 10.1080/1828051X.2021.1877574

17. Jia X, Ding P, Chen S-Y, Zhao S, Wang J, Lai S-J. Analysis of MC1R, MITF, TYR, TYRP1, and MLPH genes polymorphism in four rabbit breeds with different coat colors. *Animals*. (2021) 11:81. doi: 10.3390/ani11010081

18. Fontanesi L, Scotti E, Allain D, Dall'olio S. A frameshift mutation in the melanophilin gene causes the dilute coat colour in rabbit (*Oryctolagus cuniculus*) breeds. *Anim Genet*. (2014) 45:248–55. doi: 10.1111/age.12104

19. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*. (2011) 12:499–510. doi: 10.1038/nrg3012

20. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. (2008) 3:e3376. doi: 10.1371/journal.pone.0003376

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

22. Masharing N, Sodhi M, Chanda D, Singh I, Vivek P, Tiwari M, et al. ddRAD sequencing based genotyping of six indigenous dairy cattle breeds of India to infer

existing genetic diversity and population structure. *Sci Rep*. (2023) 13:9379. doi: 10.1038/s41598-023-32418-6

23. Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol*. (2013) 22:2841–7. doi: 10.1111/mec.12350

24. Xu P, Xu S, Wu X, Tao Y, Wang B, Wang S, et al. Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *Plant J*. (2014) 77:430–42. doi: 10.1111/tpj.12370

25. Liu C, Wang S, Dong X, Zhao J, Ye X, Gong R, et al. Exploring the genomic resources and analysing the genetic diversity and population structure of Chinese indigenous rabbit breeds by RAD-seq. *BMC Genomics*. (2021) 22:573. doi: 10.1186/s12864-021-07833-6

26. Ren A, Du K, Jia X, Yang R, Wang J, Chen S-Y, et al. Genetic diversity and population structure of four Chinese rabbit breeds. *PLoS One*. (2019) 14:e0222503. doi: 10.1371/journal.pone.0222503

27. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. (2018) 34:i884–90. doi: 10.1093/bioinformatics/bty560

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

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

30. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. (2011) 43:491–8. doi: 10.1038/ng.806

31. Van Der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. (2013) 43:11.10.1. doi: 10.1002/0471250953.bi1110s43

32. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet*. (2018) 103:338–48. doi: 10.1016/j.ajhg.2018.07.015

33. 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

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

35. Chen SY, Deng F, Huang Y, Li C, Liu L, Jia X, et al. PopSc: computing toolkit for basic statistics of molecular population genetics simultaneously implemented in web-based calculator, Python and R. *PLoS ONE*. (2016) 11:e0165434. doi: 10.1371/journal.pone.0165434

36. Wright S. The genetical structure of populations. *Ann Eugenics*. (1949) 15:323–54. doi: 10.1111/j.1469-1809.1949.tb02451.x

37. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. (2016) 33:1870–4. doi: 10.1093/molbev/msw054

38. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. (2017) 8:28–36. doi: 10.1111/2041-210X.12628

39. 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

40. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. (1995) 310:170. doi: 10.1136/bmj.310.6973.170

41. Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. *Proc R Soc Lond B Biol Sci*. (2011) 278:3161–70. doi: 10.1098/rspb.2011.1376

42. Li MH, Tiirikka T, Kantanen J. A genome-wide scan study identifies a single nucleotide substitution in ASIP associated with white versus non-white coat-colour variation in sheep (*Ovis aries*). *Heredity*. (2014) 112:122–31. doi: 10.1038/hdy.2013.83

43. Nazari-Ghadikolaei A, Mehrabani-Yeganeh H, Miarei-Aashtiani SR, Staiger EA, Rashidi A, Huson HJ. 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

44. Yang C, Ge J, Chen S, Liu Y, Chen B, Gu Z. Sequence and gene expression analysis of the agouti signalling protein gene in rex rabbits with different coat colours. *Ital J Anim Sci*. (2015) 14:3810. doi: 10.4081/ijas.2015.3810

45. Letko A, Ammann B, Jagannathan V, Henkel J, Leuthard F, Schelling C, et al. A deletion spanning the promoter and first exon of the hair cycle-specific ASIP transcript isoform in black and tan rabbits. *Anim Genet*. (2020) 51:137–40. doi: 10.1111/age.12881

46. Del Bino S, Duval C, Bernerd F. Clinical and biological characterization of skin pigmentation diversity and its consequences on UV impact. *Int J Mol Sci*. (2018) 19:2668. doi: 10.3390/ijms19092668

47. Alves JM, Carneiro M, Afonso S, Lopes S, Garreau H, Boucher S, et al. Levels and patterns of genetic diversity and population structure in domestic rabbits. *PLoS One*. (2015) 10:e0144687. doi: 10.1371/journal.pone.0144687

48. Jochová M, Novák K, Kott T, Volek Z, Majzlík I, Tůmová E. Genetic characterization of Czech local rabbit breeds using microsatellite analysis. *Livest Sci*. (2017) 201:41–9. doi: 10.1016/j.livsci.2017.03.025

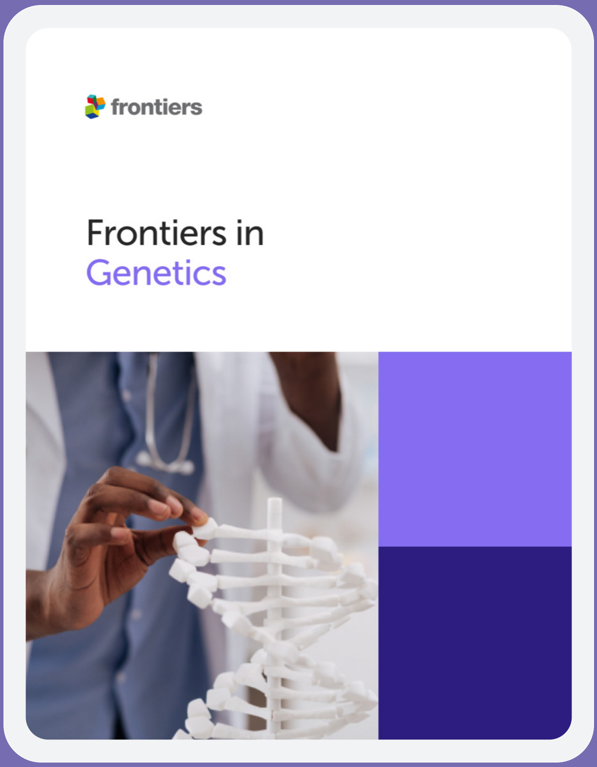
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact