MULTI-LAYERED GENOME-WIDE ASSOCIATION/PREDICTION IN ANIMALS

EDITED BY: Ruidong Xiang, Hao Cheng, Lingzhao Fang, Zhe Zhang and Marie-Pierre Sanchez

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-047-3
DOI 10.3389/978-2-88976-047-3

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

MULTI-LAYERED GENOME-WIDE ASSOCIATION/PREDICTION IN ANIMALS

Topic Editors:
Ruidong Xiang, The University of Melbourne, Australia
Hao Cheng, University of California, Davis, United States
Lingzhao Fang, University of Edinburgh, United Kingdom
Zhe Zhang, South China Agricultural University, China
Marie-Pierre Sanchez, Institut National de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), France

Citation: Xiang, R., Cheng, H., Fang, L., Zhang, Z., Sanchez, M.-P., eds. (2022). Multi-Layered Genome-Wide Association/Prediction in Animals. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-047-3

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 08 April 2022
doi: 10.3389/fgene.2022.877748

Editorial: Multi-Layered Genome-Wide Association/Prediction in Animals

Ruidong Xiang[1,2]*, *Lingzhao Fang*[3], *Marie-Pierre Sanchez*[4], *Hao Cheng*[5] and *Zhe Zhang*[6]

[1]Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, VIC, Australia, [2]Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, VIC, Australia, [3]MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, United Kingdom, [4]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France, [5]Department of Animal Science, University of California, Davis, Davis, CA, United States, [6]College of Animal Science, South China Agricultural University, Guangzhou, China

Keywords: multi-omics, biological priors, genome-wide association studies, genomic prediction, genomic selection

Editorial on the Research Topic

Multi-Layered Genome-Wide Association/Prediction in Animals

DNA mutations are the fundamental source of genomic variations that lead to phenotypic differences between individuals. Genomic variations in a population are usually assayed by single nucleotide polymorphism (SNP) arrays or whole-genome sequencing (WGS) to obtain genotype counts. If phenotypic measurements are also available on genotyped individuals in this population, genotype counts can be statistically linked to phenotypic measurements, i.e., genome-wide association studies (GWAS). Decades of GWAS in humans (Visscher et al., 2017) and animals (Hayes and Daetwyler, 2019) have shown that causal variants for complex traits are largely located at non-coding regions of the genome. This has been further supported by recent human studies of genetic variations with roles in gene regulation, e.g., those that are gene expression quantitative trait loci (eQTL) (Consortium, 2020) are enriched in causal variants of complex traits. Due to the vast availability of data in humans, such as proteomics and metabolomics, great efforts have been invested in the integration of multi-omics information and GWAS results (Hasin et al., 2017). The effort of functional annotation of animal genomes only started recently (Clark et al., 2020), although the size of multi-omics data has been increasing (Liu et al., 2021).

Unlike genomics research in humans, GWAS in animals is usually carried out amongst related individuals with small effective population sizes. This results in many SNPs in high linkage disequilibrium (LD) from a locus being associated with a trait, and it is difficult to distinguish which ones are causal. This is particularly difficult when the GWAS used imputed sequence variants (Hayes and Daetwyler, 2019) where a large number of variants are in very strong LD. Therefore, external information, such as multi-omics datasets independent of GWAS, is needed to pinpoint causal signals. Apart from the use of multi-omics data, multi-trait meta-analyses of GWAS (Xiang et al., 2020; Xiang et al., 2021) and large-scale GWAS of intermediate traits like milk composition (Sanchez et al., 2021) also improve the detection of causal variants. In addition, multi-breed meta-analyses can help to pinpoint causal mutations as LD is conserved over shorter distances across breeds (van den Berg et al., 2020).

The genomic information of domestic animals is used to improve animal breeding. In particular, genomic selection or genomic prediction (GP) (Meuwissen et al., 2001) using genome-wide marker information has greatly benefited animal breeding. GP was primarily designed to use all available markers to estimate genomic breeding values (gEBVs) reflecting the genetic merit of animals. However, the accuracy of GP, approximated as the correlation between gEBVs and phenotype in the validation population is far from being perfect. There are

Edited and reviewed by:
Martino Cassandro, University of Padua, Italy

***Correspondence:**
Ruidong Xiang
ruidong.xiang@unimelb.edu.au

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

Received: *17 February 2022*
Accepted: *07 March 2022*
Published: *08 April 2022*

Citation:
Xiang R, Fang L, Sanchez M-P, Cheng H and Zhang Z (2022) Editorial: Multi-Layered Genome-Wide Association/Prediction in Animals. Front. Genet. 13:877748. doi: 10.3389/fgene.2022.877748

many ways to improve the accuracy of GP and emerging evidence shows that the use of functional information can enhance GP (MacLeod et al., 2016; Xiang et al., 2019; Teng et al., 2020). With the growing size of functional genomics data, it is anticipated that functional genomics priors will be routinely integrated into GP to improve its accuracy. This will then need the development of suitable methodologies that effectively fuse multi-omics data together with the genotype-phenotype association analysis in the GP model (Cheng et al., 2021).

The 'Multi-layered Genome-wide Association and Prediction in Animals' research topic intends to collect high-quality articles on the emerging area of integrating multi-omics datasets into GWAS and GP. In its conclusion, 9 articles from 58 authors have been collected, ranging from data generation, integrative analysis of multi-omics with GWAS and GP, and new method development across multiple domestic species.

Developing new methods for the integrative analysis of multi-omics and GWAS/GP is one of the key research areas in genetics. Due to flexibility in incorporating priors in the model, several new Bayesian methods have been proposed, including BayesHP and BayesHE (Shi et al.) that incorporate "global-local" shrinkage priors, and multi-class Bayesian Alphabet methods (Wang et al.) that incorporate biological information into multi-trait Bayesian analysis. The application of these methods into simulated and real data supports that incorporating biological priors into GP training improves its accuracy.

GWAS or GP using WGS is another emerging area. Due to high costs of in-depth WGS, there is a new shift toward using low-pass WGS which provides cost-effective options for GWAS or GP to use millions of sequence variants. By analyzing simulated data, Deng et al. show that imputation using low-pass WGS is more accurate than using SNP arrays. This was also found by Zhao et al. where real low-pass WGS from donkeys were generated, analyzed, and applied to GP.

GWAS in animals has been largely used to dissect causative loci associated with complex traits. Jiang et al. present such an effort in detecting loci associated with body size in Hu sheep. Also, Yang et al. identified loci associated with meat production in chicken. Apart from the standard linear mixed model, GWAS can also be carried out using single-step Bayesian regression, and Naserkheil et al. present such an effort in identifying loci associated with meat production traits of beef cattle.

In fact, loci prioritized by GWAS may be used as biological priors to enhance GP. However, Gebreyesus et al. found that adding GWAS-prioritized variants had no improvement in GP for survival traits of dairy cattle which have very low heritability estimates. This emphasizes that more studies are needed in this area. Other lowly heritable but important traits in cattle included female fertility. Chen et al. found that accounting for sire genetic effects improves the genetic evaluation of fertility of Holstein cows.

In conclusion, integrating multi-omics data with GWAS and GP in animals is an important and emerging research area in livestock genomics. We anticipate that the development and application of efficient methods, increased use of WGS, and integration of more types of multi-omics data will be future directions of this area. Understanding how DNA mutations shape complex traits not only furthers our understanding of biology, but also provides practical benefits in animal breeding.

AUTHOR CONTRIBUTIONS

RX drafted the manuscript, and revised with all authors. All authors have proof-read the final version.

REFERENCES

Cheng, H., Zhao, T., and Zeng, J. (2021). Extend Mixed Models to Multi-Layer Neural Networks for Genomic Prediction Including Intermediate Omics Data. *bioRxiv* 2021, 472186. doi:10.1101/2021.12.10.472186

Clark, E. L., Archibald, A. L., Daetwyler, H. D., Groenen, M. A. M., Harrison, P. W., Houston, R. D., et al. (2020). From FAANG to fork: Application of Highly Annotated Genomes to Improve Farmed Animal Production. *Genome Biol.* 21, 285–289. doi:10.1186/s13059-020-02197-8

Consortium, G. (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics Approaches to Disease. *Genome Biol.* 18, 83–15. doi:10.1186/s13059-017-1215-1

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

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., et al. (2021). A Comprehensive Catalogue of Regulatory Variants in the Cattle Transcriptome. *bioRxiv* 2020, 406280. doi:10.1101/2020.12.01.406280

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting Biological Priors and Sequence Variants Enhances QTL Discovery and Genomic Prediction of Complex Traits. *BMC genomics* 17, 144. doi:10.1186/s12864-016-2443-6

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Sanchez, M.-P., Rocha, D., Charles, M., Boussaha, M., Hozé, C., Brochard, M., et al. (2021). Sequence-based GWAS and post-GWAS Analyses Reveal a Key Role of SLC37A1, ANKH, and Regulatory Regions on Bovine Milk mineral Content. *Scientific Rep.* 11, 1–15. doi:10.1038/s41598-021-87078-1

Teng, J., Huang, S., Chen, Z., Gao, N., Ye, S., Diao, S., et al. (2020). Optimizing Genomic Prediction Model Given Causal Genes in a Dairy Cattle Population. *J. Dairy Sci.* 103, 10299–10310. doi:10.3168/jds.2020-18233

van den Berg, I., Xiang, R., Jenko, J., Pausch, H., Boussaha, M., Schrooten, C., et al. (2020). Meta-analysis for Milk Fat and Protein Percentage Using Imputed Sequence Variant Genotypes in 94,321 Cattle from Eight Cattle Breeds. *Genet. Sel Evol.* 52, 37. doi:10.1186/s12711-020-00556-4

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

Xiang, R., Berg, I. v. d., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the Contribution of Sequence Variants with Regulatory and Evolutionary Significance to 34 Bovine Complex Traits. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19398–19408. doi:10.1073/pnas.1904159116

Xiang, R., MacLeod, I. M., Daetwyler, H. D., de Jong, G., O'Connor, E., Schrooten, C., et al. (2021). Genome-wide fine-mapping Identifies Pleiotropic and Functional Variants that Predict many Traits across Global Cattle Populations. *Nat. Commun.* 12, 860. doi:10.1038/s41467-021-21001-0

Xiang, R., van den Berg, I., MacLeod, I. M., Daetwyler, H. D., and Goddard, M. E. (2020). Effect Direction Meta-Analysis of GWAS Identifies Extreme, Prevalent and Shared Pleiotropy in a Large Mammal. *Commun. Biol.* 3, 88. doi:10.1038/s42003-020-0823-6

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 in Genetics

ORIGINAL RESEARCH
published: 30 March 2021
doi: 10.3389/fgene.2021.645107

Identification of Major Loci and Candidate Genes for Meat Production-Related Traits in Broilers

*Xinting Yang[1†], Jiahong Sun[1†], Guiping Zhao[1], Wei Li[1], Xiaodong Tan[1], Maiqing Zheng[1], Furong Feng[2], Dawei Liu[2], Jie Wen[1] and Ranran Liu[1]**

[1] State Key Laboratory of Animal Nutrition, Key Laboratory of Animal (Poultry) Genetics Breeding and Reproduction, Ministry of Agriculture, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] Foshan Gaoming Xinguang Agricultural and Animal Industrials Corporation, Foshan, China

Background: Carcass traits are crucial characteristics of broilers. However, the underlying genetic mechanisms are not well understood. In the current study, significant loci and major-effect candidate genes affecting nine carcass traits related to meat production were analyzed in 873 purebred broilers using an imputation-based genome-wide association study.

Results: The heritability estimates of nine carcass traits, including carcass weight, thigh muscle weight, and thigh muscle percentage, were moderate to high and ranged from 0.21 to 0.39. Twelve genome-wide significant SNPs and 118 suggestively significant SNPs of 546,656 autosomal variants were associated with carcass traits. All SNPs for six weight traits (body weight at 42 days of age, carcass weight, eviscerated weight, whole thigh weight, thigh weight, and thigh muscle weight) were clustered around the 24.08 Kb region (GGA24: 5.73–5.75 Mb) and contained only one candidate gene (*DRD2*). The most significant SNP, rs15226023, accounted for 4.85–7.71% of the estimated genetic variance of the six weight traits. The remaining SNPs for carcass composition traits (whole thigh percentage and thigh percentage) were clustered around the 42.52 Kb region (GGA3: 53.03–53.08 Mb) and contained only one candidate gene (*ADGRG6*). The most significant SNP in this region, rs13571431, accounted for 11.89–13.56% of the estimated genetic variance of two carcass composition traits. Some degree of genetic differentiation in *ADGRG6* between large and small breeds was observed.

Conclusion: We identified one 24.08 Kb region for weight traits and one 42.52 Kb region for thigh-related carcass traits. *DRD2* was the major-effect candidate gene for weight traits, and *ADGRG6* was the major-effect candidate gene for carcass composition traits. Our results supply essential information for causative mutation identification of carcass traits in broilers.

Keywords: carcass composition, thigh meat, weight trait, genome-wide association study, imputation, candidate genes

OPEN ACCESS

Edited by:
Zhe Zhang,
South China Agricultural University, China

Reviewed by:
Luke Kramer,
Iowa State University, United States
Andrea Talenti,
University of Edinburgh, United Kingdom

****Correspondence:***
Ranran Liu
liuranran@caas.cn

† These authors have contributed equally to this work

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

Received: *22 December 2020*
Accepted: *02 March 2021*
Published: *30 March 2021*

Citation:
Yang X, Sun J, Zhao G, Li W, Tan X, Zheng M, Feng F, Liu D, Wen J and Liu R (2021) Identification of Major Loci and Candidate Genes for Meat Production-Related Traits in Broilers. Front. Genet. 12:645107. doi: 10.3389/fgene.2021.645107

INTRODUCTION

For global meat consumption, chicken meat is the second largest and provide almost 1/3 of meat resource[1]. Improvements in the weight and carcass traits are major goals in modern broiler breeding programs. In particular, thigh development and meat production are closely related to the efficiency of the broiler industry. These traits have moderate to high heritability and are controlled by multiple genes (Claire D'Andre et al., 2013; Flisar et al., 2014).

Many studies have been performed to identify quantitative trait loci (QTLs), genes, and/or causative mutation. In pigs, a causative mutation of *IGF2* has a major effect on muscle growth (Van Laere et al., 2003). In large dog breeds, variants in *IRS4*, *ACSL4*, and *IGSF1* were strongly associated with skeletal size and body mass (Plassais et al., 2017). In mice and humans, changes in the Neurobeachin abundance or activity significantly affect the body weight (Olszewski et al., 2012). A non-synonymous *FGD3* variant was identified as a positional candidate for disproportional tall stature, accounting for a carcass weight QTL and skeletal dysplasia in Japanese Black cattle (Takasuga et al., 2015). Many QTL and genes have been found in chickens, including *LCORL1* and *LDB2* (Gu et al., 2011; Liu et al., 2013, 2015). Recently, multiple haplotypes at the distal end of chromosome 1 were identified as a major-effect QTL for chicken growth traits (Wang et al., 2020). However, the genetic mechanisms of carcass traits are not well understood.

An imputation-based genome-wide association study was conducted to identify significant loci and candidate genes affecting multiple weight and carcass composition traits in fast-growing white-feathered broilers.

MATERIALS AND METHODS

Experimental Birds

Chickens were obtained from a fast-growing white-feathered broiler line B. The chickens were produced and raised by Foshan Gaoming Xinguang Agricultural and Animal Industrials Co., Ltd. (Foshan, China). In generation 5 (G5), 873 breeders (412 males and 462 females) were randomly selected and slaughtered at 42 day of age. All birds were raised in stair-step cages under the same recommended environmental (Li et al., 2009) and nutritional conditions (Feeding Standard of Chickens, China, NY 33-2004).

In addition, thirty Chahua and 24 Daweishan mini chickens were used for the phylogenetic and selective sweep analysis. Chahua chickens are similar to red junglefowl. The blood samples were supplied by the Aquatic Animal Husbandry Association of Yulin City in the Guangxi Zhuang Autonomous Region. The individuals are small, and the body weight rarely exceeds 1 kg. The blood samples of the Daweishan mini chickens were obtained from the Yunnan Agricultural University. The Daweishan mini chickens are a small-sized Chinese indigenous breed similar to junglefowl distributing in the tropical and subtropical zone in Yunnan Province. The typical characteristics of Daweishan Mini chickens are a slow growth rate and low body weight compared with Chinese indigenous chickens of the same age (Rong et al., 2011).

Phenotypic Measurements

The weight traits and carcass composition traits were measured in line B as follows: body weight at 42 day of age (BW42), carcass weight (CW), eviscerated weight (EW), whole thigh weight (WThW, weight including feet, and single), thigh weight (ThW, weight with bone, and single), thigh muscle weight (ThMW, single), and the WThW, ThW, and ThMW as percentages of BW42 (WThP, ThP, and ThMP). In addition, the same method (Li D. et al., 2020) was used to calculate the feed intake from 28 to 42 days of age (FI).

Estimates of Genetic Parameters

A linear mixed model (LMM) was fitted using the residual maximum likelihood (REML). The ASReml software (Gilmour et al., 2009) was used to estimate the variance components of the carcass traits of 873 broilers. The Wald *F*-statistics showed that the fixed effect of sex was statistically significant ($P < 0.05$). The variance components were estimated using a univariate mixed liner animal model. The animal model was expressed as follows:

$$Y = Xb + Zu + e$$

where Y is the vector of observations; b is the vector of the fixed effects; u is the vector of the additive genetic effects, with u $\sim N(0, A\sigma_u^2)$, where A is the additive genetic relationship matrix (GRM), and σ_u^2 is the additive genetic variance; e is the vector of the random residual effects; and X and Z are the incidence matrices assigning observations to effects. The fixed effect in the model was sex.

The phenotypic variance is the sum of all variance components and is defined as follows:

$$\sigma_P^2 = \sigma_a^2 + \sigma_e^2$$

The heritability is the ratio of the additive genetic variance to the phenotypic variance and is defined as follows:

$$h^2 = \frac{\sigma_a^2}{\sigma_P^2}$$

where σ_P^2 is the phenotypic variance, σ_a^2 is the additive variance, and σ_e^2 is the residual fraction.

A bivariate animal model was fitted to estimate the phenotypic and genetic correlations between carcass traits using the ASReml software package. The description of the model terms was the same as that for the univariate mixed linear animal model.

A likelihood ratio test (LRT) was used to determine the significance of the heritability and the genetic correlations. The LRT compares the likelihood of a full model with that of a nested model (without the additive genetic component) and was used to test for nonzero additive genetic variance.

[1] http://www.fao.org/home/en/

Genotyping, Imputation, and Quality Control

Genomic DNA was extracted from blood samples using the phenol-chloroform method. The 873 broilers were genotyped with the customized chicken 55 K SNP array obtained from Beijing Compass Biotechnology Co., Ltd. (Beijing, China; Liu et al., 2019). A total of 873 broilers (412 males and 462 females) were used for genotype imputation for the target panel. The following quality control criteria were used for the target panel: individual call percentage ≥90%, SNP call percentage ≥90%, and minor allele frequency (MAF) ≥0.05. In addition, the SNPs located on the sex chromosomes and GGA16 were removed. We used 41,204 autosome variants and 873 broilers in the subsequent analyses.

Whole-genome sequences (WGS) of 230 broilers (101 males and 129 females) from the same line in G7 were used for the reference panel (Li D. et al., 2020). Briefly, an average depth of 10× was acquired, and variant calling and SNP filtering were performed according to a standardized bioinformatics pipeline. The quality control criteria of the reference panel included the individual call percentage ≥90%, SNP call percentage ≥90%, and MAF ≥ 0.05. After filtering, 9,760,228 autosome variants remained for 230 birds.

For the Chahua and Daweishan mini chickens, genome resequencing was conducted on the Illumina NovaSeq 6000 platform with an average depth of approximately 10× coverage. The sequencing was performed by Annoroad Gene Technology Co., Ltd. (Beijing, China). Variant calling was performed according to a standardized bioinformatics pipeline (DePristo et al., 2011; Van der Auwera et al., 2013). Specifically, clean sequencing data were aligned to the chicken reference genome (GRCg6a/galGal6; ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/315/GCF_000002315.6_GRCg6a/) with the Burrows-Wheeler Aligner (BWA)-MEM algorithm (Li, 2013). Then, PCR duplicates were removed with Picard-tools v1.115[2]. Variant calling was then performed via the HaplotypeCaller in GVCF mode with joint genotyping on all samples. We used ANNOVAR software (Wang et al., 2010) and the existing genome annotation file (gff) to annotate each detected SNP. Finally, the SNPs were filtered with the GATK Variant Filtration protocol. The filtering settings were as follows: variant confidence score < 30.0, QualByDepth < 2.0, ReadPosRankSum < −8.0, total depth of coverage < 4.0, and FisherStrand > 60.0. In addition, quality control was conducted using the following criteria: individual call rate ≥ 90%, SNP call rate ≥ 90%, and MAF ≥ 0.05. After filtering, a total of 9,235,705 autosome variants remained for the 54 sequenced birds.

Genotype imputation of the 55 K genotypes of the broilers to the imputed WGS level was performed with Beagle 5.0 (Browning et al., 2018). The effective population size (Ne) affects the accuracy of genotype imputation (Van den Berg et al., 2019) because it is much smaller in livestock than in humans (Hall, 2016). We used SNeP software (Barbato et al., 2015) to estimate the Ne; SNeP estimates the Ne using the following equation:

$$N_T(t) = \frac{1}{4f(c_t)} \times \left[\frac{1}{E(r^2_{adj} \nu c_t)} - a \right]$$

where $N_T(t)$ is the effective population size estimated t generations ago, $f(c_t)$ is the mapping function used to estimate the recombination rate (c_t) t generations ago, r^2_{adj} is the linkage disequilibrium (LD) estimate adjusted for the sampling bias ($r^2_{adj} = r^2 - \frac{1}{2N}$, where N is the population sample size), and "a" is a constant accounting for mutation. The results showed that the Ne was 432 before the 13th generation.

The reference panel was pre-phased with Beagle 5.0 (default settings, except that the Ne was 432; Browning et al., 2018). Then the imputation from 55 K to the WGS level was executed in Beagle 5.0 with the default parameters, except that the Ne was 432. To assess the imputation accuracy, the allelic R^2 values for each variant and the genotype concordance rate for a random 2% SNPs were determined. The allelic R^2 was calculated as the estimated squared correlation of the imputed sequence genotype on the true sequence, which was given by Beagle 5.0. The genotype concordance rate was calculated by comparing the imputed and real genotypes for a randomly masked 2% of the SNPs analyzed with both panels. We applied the following post-imputation filtering criteria for each SNP: allelic $R^2 \geq 0.8$ and MAF ≥ 0.05. Finally, 6,546,656 autosomal variants and 873 samples remained for the subsequent genome−wide association study (GWAS; **Supplementary Table 1**).

Genome-Wide Association Study

The GWAS was performed using the univariate LMM implemented in GEMMA version 0.98.1 software[3] (Zhou and Stephens, 2012). Due to the high genetic correlation between the weight traits and carcass composition traits, we fitted a multivariate linear mixed model (MLMM) for the weight traits and carcass composition traits. The genotype was the fixed effect, and the additive polygenic effect was the random effect. Sex was considered a covariate for all traits.

The univariate LMM had the following form:

$$y = W\alpha + x\beta + u + \varepsilon;\ u \sim MVN_n(0,\ \lambda\tau^{-1}K),$$
$$\varepsilon \sim MVN_n(0,\ \tau^{-1}I_n),$$

where y represents the vector of the phenotypic values; W represents the vector of the covariates, including a column of 1 s; α represents the vector of the corresponding coefficients, including the intercept; x represents the vector of the marker genotypes; β represents the effect size of the marker; u represents the vector of the random polygenic effects; ϵ represents the vector of errors; τ^{-1} represents the variance of the residual errors; λ represents the ratio between the two variance components; K represents the centered relatedness matrix estimated from 6,546,656 variants, and In represents the identity matrix. MVN_n denotes the n-dimensional multivariate normal distribution. The

[2]http://broadinstitute.github.io/picard/

[3]https://github.com/genetics-statistics/GEMMA/releases

Wald test was used as a criterion to select the SNPs associated with the carcass traits.

The MLMM had the following form:

$$Y = WA + x\beta^T + U + E;\ G \sim MN_{n\times d}(0,\ K,\ V_g),$$
$$E \sim MN_{n\times d}(0,\ I_{n\times n},\ V_e),$$

where Y is an $n \times d$ matrix of d phenotypes for n individuals; $W = (w_1, \cdots, w_c)$ is an $n \times c$ matrix of covariates, including a column of 1 s; A is a $c \times d$ matrix of the corresponding coefficients, including the intercept; x is an n-vector of the marker genotypes; β is a d vector of the marker effect sizes for the d phenotypes; U is an $n \times d$ matrix of the random effects; E is an $n \times d$ matrix of the errors; K is the centered relatedness matrix estimated from 6,546,656 variants, $I_{n\times n}$ is a $n \times n$ identity matrix, V_g is a $d \times d$ symmetric matrix of the genetic variance components, V_e is a $d \times d$ symmetric matrix of the environmental variance components, and $MN_{n\times d}(0, V_1, V_2)$ denotes the normal distribution of the $n \times d$ matrix with mean 0, a row covariance matrix V_1 (n by n), and a column covariance matrix V_2 (d by d). The Wald test was used as a criterion to select the SNPs associated with the carcass traits.

The threshold P-value of the 5% Bonferroni genome-wide significance was 7.63e-9 (0.05/6,546,656), and that of the significance of the "suggestive association" that allows a one-time false positive effect in the GWAS test was 1.52e-7 (1/6,546,656). It was calculated using the same method. Manhattan and quantile-quantile (Q-Q) plots were constructed for each trait using the CMplot package[4] in R (version 4.0.0). LD blocks of the target regions were identified using the Haploview version 4.2 software (Barrett et al., 2005). The SNP positions were updated using the newest release from the University of California-Santa Cruz (UCSC; GRCg6a/galGal6 genome version). The identification of genes in the genome-wide significant and suggestive regions was performed using the UCSC annotation of the GRCg6a/galGal6 genome version[5]. Boxplots were produced with the ggplot2 package in R (version 4.0.0).

Bayesian Analysis

A Bayesian approach called Bayes C_π (Habier et al., 2011) was used to obtain the average proportion of genetic variance and phenotypic variance explained by each 1-Mb genomic non-overlapping window (n = 1,024). The Bayesian analyses were performed using the hibayes package[6] in R (version 4.0.0). The number of iterations after the burn-in phase was 20,000, and that of the burn-in period was 10,000. Sex was fitted as a fixed covariate for all traits.

Estimation of SNP Effect Size

The estimation of SNP effect size was performed using MLMM analysis with the genotype data used to compute the GRM in DISSECT (Canela-Xandri et al., 2015)[7]. The approach is based on fitting the equation:

$$y = X\beta + Wu + \epsilon$$

where y is the vector of the phenotypes, β is the vector of fixed effects that included sex, u is the vector of the SNP effects distributed as $u \sim N(0, I\sigma_u^2)$, I is the identity matrix, and ϵ is the vector of the residual effects distributed as $\sim N(0, I\sigma^2)$. W is a genotype matrix defined by the equation:

$$w_{ik} = \frac{(s_{ik} - 2p_k)}{\sqrt{2p_k(1-p_k)}}$$

where s_{ik} is the number of copies of the reference allele for the SNP k of the individual i, and p_k is the frequency of the reference allele for the SNP k. In this model, the variance of y is:

$$\mathrm{var}(y) = A\sigma_g^2 + I\sigma^2$$

where A is the GRM, σ_g^2 is the genetic variance, and σ^2 is the residual variance. The variance components were estimated using the REML.

Phylogenetic Analysis

Using data from 80 white-feathered broilers with different-parents, 30 Chahua chickens, and 24 Daweishan mini chickens, the pair-wise genetic distance matrices were calculated based on whole-genome SNPs, two candidate genes, and five randomly selected 100 Kb regions, respectively. A total of eight neighbor-joining trees were then constructed using MEGA X (Kumar et al., 2018). The defaults were used for all parameters.

Selective Sweep Analysis

Selective sweep analysis was performed on 80 white-feathered broilers with different parents, 30 Chahua chickens, and 24 Daweishan mini chickens. The population differentiation index (F_{ST}) was calculated as described by Weir and Cockerham (1984) and was implemented in the VCFtools v0.1.14 program (Danecek et al., 2011). The average F_{ST} values were plotted in 20-kb overlapping genomic bins (for more than 10 SNPs) with a 10-kb step-size.

The nucleotide diversity (π) of each population was estimated using a 20-kb sliding window (for more than 10 SNPs) with a 10-kb step across the whole genome, and the ratio ($\pi_{\text{Chahua and MINI}}/\pi_{\text{B line}}$) was computed. The $\log_2(\pi$ ratio) was defined as $\log_2 \frac{\pi_{\text{Chahua and MINI}}}{\pi_{\text{line B}}}$.

The heterozygosity H_P was calculated only in the line B chickens as: $H_P = \frac{2\sum n_{MAJ}\sum n_{MIN}}{(\sum n_{MAJ}+\sum n_{MIN})^2}$, where $\sum n_{MAJ}$ is the sum of the major allele frequencies, and $\sum n_{MIN}$ is the sum of the minor allele frequencies within a window. The H_P value was Z-transformed as follows: $ZH_P = \frac{(H_P - \mu H_P)}{\sigma H_P}$, where μH_P is the overall average heterozygosity and σH_P is the standard deviation for all windows.

The cross-population extended haplotype homozygosity (XP-EHH; Sabeti et al., 2007) was calculated for each SNP in the

[4]https://cran.r-project.org/package=CMplot

[5]http://genome-asia.ucsc.edu/cgi-bin/hgGateway?hgsid=472768848_otkBtCHKhHMTV1xrxHuq737iivJ1

[6]https://github.com/YinLiLin/hibayes

[7]https://www.dissect.ed.ac.uk

dataset using Selscan v-1.3.0 (Szpiech and Hernandez, 2014). Prior to the XP-EHH test, Beagle 5.0 software (Browning et al., 2018) was used to estimate missing genotypes and reconstructing the haplotypes from the unphased SNP genotype data.

The overall experimental workflow is depicted in **Supplementary Figure 1**.

RESULTS

Descriptive Statistics of Phenotypes

The descriptive statistics of the carcass traits are listed in **Table 1**. At 42 days of age, the body weight (BW42), carcass weight (CW), and eviscerated weight (EW) were 1912.74 ± 196.49, 1728.24 ± 180.26, and 1229.60 ± 145.98. The whole thigh weight (WThW, weight including feet), thigh weight (ThW, weight with bone), and thigh muscle weight (ThMW) were 249.47 ± 31.13, 213.37 ± 26.65, and 157.19 ± 20.20. The whole thigh percentage (WThP), thigh percentage (ThP), and thigh muscle percentage (ThMP) were 13.03 ± 0.74, 11.15 ± 0.67, and 8.21 ± 0.54. Feed intake from 28 to 42 days of age was 1775.98 ± 147.83. The coefficients of variation of these traits in the population ranged from 5.64 to 12.85%.

Estimates of Genetic Parameters

The heritabilities and the phenotypic and genetic correlations for the carcass traits are presented in **Table 2**. The WThP and ThP had moderate heritabilities (0.21–0.22). The heritabilities of the other carcass traits were higher (0.30 to 0.39). All weight traits showed strong positive genetic correlations (0.78–0.99) and phenotypic correlations (0.71–0.99). All carcass composition traits showed strong positive genetic correlations (0.76–0.94) and phenotypic correlations (0.81–0.98).

Imputation Accuracy

The proportion of SNP markers on each chromosome based on different datasets are shown in **Figure 1A**. The numbers of SNPs in the different MAF classes for different datasets are shown in **Figure 1B**. In general, the proportion of SNP markers on each chromosome and the MAF distribution of the four datasets showed the same trends. Consistency was observed in the distribution of SNPs between the 55 K array data and the imputed WGS data after filtering and between the WGS data and imputed WGS data before filtering (MAF $\geq$ 0.05).

The allelic R^2 values and the average genotype concordance rate were used to evaluate the imputation accuracy of the imputed WGS data (**Figure 1C**). At the chromosome level, the allelic R^2 values of the imputed sequence before filtering ranged from 0.52 to 0.87, whereas the genotype concordance rate fluctuated between 0.60 and 0.89. After post-imputation filtering, the allelic R^2 values and the genotype concordance rate reached an average of 0.89 and 0.91, respectively. The distribution of the SNPs used in the GWAS after post-imputation filtering is summarized in **Supplementary Table 1**.

TABLE 1 | Descriptive statistics of the carcass traits of broilers.

Traits[1]	*N*	Mean	SD[2]	Min	Max	CV, %[3]
BW42, g	873	1912.74	196.49	1327.60	2502.00	10.27
CW, g	873	1728.24	180.26	1168.90	2278.60	10.43
EW, g	873	1229.60	145.98	786.80	1724.60	11.87
WThW, g	873	249.47	31.13	164.86	343.06	12.48
ThW, g	873	213.37	26.65	135.80	293.20	12.49
ThMW, g	873	157.19	20.20	99.50	224.50	12.85
WThP, %	873	13.03	0.74	10.80	15.33	5.64
ThP, %	873	11.15	0.67	9.25	13.10	5.98
ThMP, %	873	8.21	0.54	6.46	11.36	6.63
FI, g	873	1775.98	147.83	1235.20	2130.20	8.32

[1]*Body weight at 42 days of age (BW42), carcass weight (CW), eviscerated weight (EW), whole thigh weight (WThW), thigh weight (ThW), thigh muscle weight (ThMW), whole thigh percentage (WThP), thigh percentage (ThP), thigh muscle percentage (ThMP), and feed intake from 28 to 42 days of age (FI).*
[2]*Standard deviation.*
[3]*Coefficient of variation (%).*

GWAS Results

The Manhattan and Q-Q plots of the univariate GWAS results are presented in **Figure 2A** and **Table 2**. We detected 12 genome-wide significant SNPs and 118 suggestively significant SNPs. All SNPs for the weight traits (BW42, CW, EW, WThW, ThW, and ThMW) were clustered around the 242.29 Kb region (GGA24: 5.63–5.87 Mb). The most significant SNP, rs15226023, accounted for 4.85–7.71% of the genetic variance of the six weight traits. The remaining SNPs for the carcass composition traits (WThP and ThP) were clustered around the 42.52 Kb region (GGA3: 53.03–53.08 Mb). The most significant SNP in this region, rs13571431, accounted for 11.89 and 13.56% of the genetic variance of WThP and ThP, respectively.

The consistent significant loci were detected using the multivariate GWAS and univariate GWAS. The Manhattan and Q-Q plots of the multivariate GWAS are presented in **Figure 2B** and **Table 2**.

Based on $r^2 \geq 0.8$, the empirical confidence interval of the QTL in the GGA24 for the weight traits was 24.08 kb (GGA24: 5.73–5.75 Mb), and the unique gene *DRD2* was located in the region (**Figure 3A**). Another QTL in the GGA3 for carcass composition traits was 42.52 kb (GGA3: 53.03–53.08 Mb), and the unique gene *ADGRG6* was located in the region (**Figure 3B**).

For the weight traits, the 24.83 Kb region in GGA24 (GGA24: 5.73–5.75 Mb) in the significant region was detected by LD analysis (**Supplementary Figure 2A**). This LD block covered the exon1 and intron1 of *DRD2* and had a positive effect ($\beta < 0$) on all weight traits (**Supplementary Figure 2B**). Interestingly, this LD block also had a positive effect ($\beta < 0$) on FI (**Supplementary Figure 2B**).

For WThP and ThP, one 42.52 kb strong LD block in GGA3 (GGA3: 53.03–53.08 Mb) was detected by LD analysis and contained 69 significant SNPs (**Supplementary Figure 3A**). This LD block covered exon1, intron1, exon2, and intron2 of *ADGRG6* and had a negative effect ($\beta < 0$) on all carcass composition traits (**Supplementary Figure 3B**).

The effects of the significant LD block resulted in significant differences in the carcass traits, as shown in **Supplementary Figures 2**, **3**. The lowest and highest phenotypic values belonged

TABLE 2 | Overview of the significant QTLs of univariate GWAS and multivariate GWAS associated with target traits.

Traits[1]	GGA[2]	Base-pair region		nSNP[3]	Lead SNP	Alleles	MAF	*P*-value	Candidate gene	PVE(%)[4]	GVE(%)[5]	Position
		Start	End		(rsname)							
BW42	24	5658679	5835194	14	5741556	C/G	0.23	4.06E-09	DRD2	2.75	4.85	Intron 1
CW	24	5658679	5754835	13	5741556	C/G	0.23	3.97E-09	DRD2	2.79	5.00	Intron 1
EW	24	5632110	5835194	49	5741556	C/G	0.23	3.46E-10	DRD2	3.43	7.71	Intron 1
WThW	24	5741556	5874395	5	5741556	C/G	0.23	1.21E-08	DRD2	2.81	5.96	Intron 1
ThW	24	5741556	5744643	3	5741556	C/G	0.23	2.62E-08	DRD2	2.72	5.89	Intron 1
ThMW	24	5741556	5744643	2	5741556	C/G	0.23	4.94E-08	DRD2	2.52	4.95	Intron 1
WThP	3	53032570	53075086	65	53034833	T/C	0.35	2.40E-09	ADGRG6	3.55	11.89	Intron 2
ThP	3	53032570	53075086	68	53034833	T/C	0.35	4.61E-09	ADGRG6	3.43	13.56	Intron 2
BW42, CW, EW	24	5658679	5744643	4	5741556	C/G	0.23	1.15E-08	DRD2	/	/	Intron 1
BW42, CW, EW, WThW, ThW, ThMW	24	5741556	5741556	1	5741556	C/G	0.23	1.28E-07	DRD2	/	/	Intron 1
WThP, ThP, ThMP	3	53032570	53757085	58	53033387	C/T	0.33	8.63E-09	ADGRG6	/	/	Intron 2

[1]*Body weight at 42 days of age (BW42), carcass weight (CW), eviscerated weight (EW), whole thigh weight (WThW), thigh weight (ThW), thigh muscle weight (ThMW), whole thigh percentage (WThP), thigh percentage (ThP), thigh muscle percentage (ThMP).*
[2]*Gallus gallus chromosome.*
[3]*The number of significant SNPs in the interval.*
[4]*Variance in phenotype explained by the lead SNP.*
[5]*Genetic variance explained by the lead SNP.*

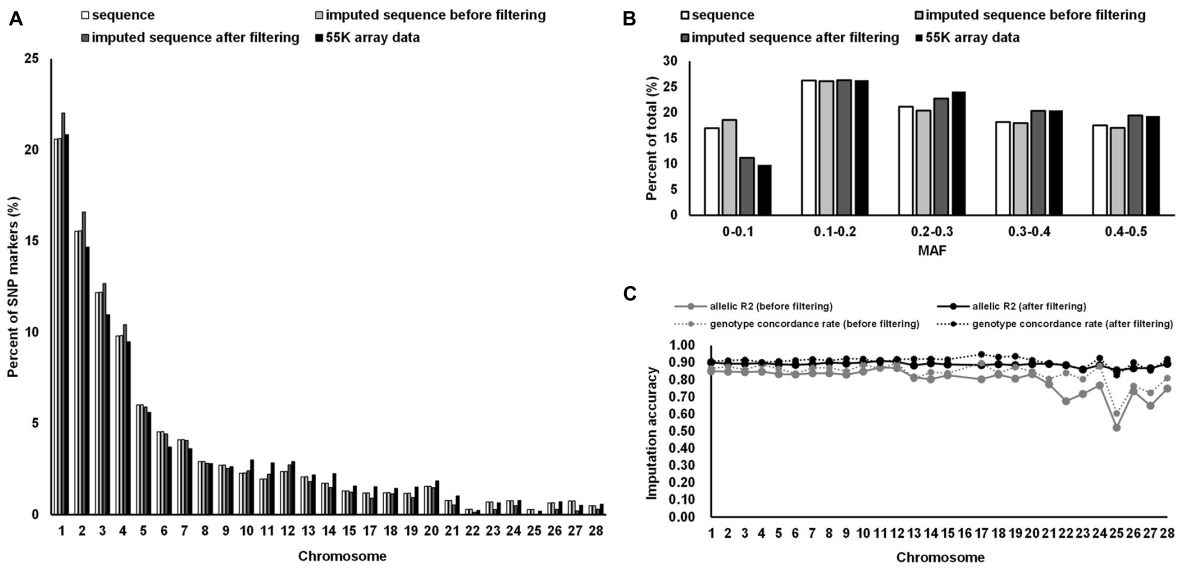

FIGURE 1 | Distribution of SNP markers and imputation accuracy. **(A)** Percentage of SNPs on the chromosome for the 55 K array data of 873 birds, WGS data of 230 birds, and imputed WGS data of 873 birds after imputation and post-imputation filtering. **(B)** Percentage of SNPs in each MAF class for different datasets. **(C)** The imputation accuracy of the imputed WGS data of 873 birds after imputation and post-imputation filtering. MAF, minor allele frequency.

to homozygotes, whereas heterozygotes had intermediate values. Broilers with homozygous mutations of the LD block on GGA24 (GGA24: 5.73–5.75 Mb) had higher carcass weights and higher FIs than those with homozygous wild type. Broilers with homozygous mutations of the LD block on GGA3 (GGA3: 53.03–53.08 Mb) had lower WThP, ThP, and ThMP than those with homozygous wild type.

Bayesian Analysis

The Manhattan plots of the genetic variance and phenotypic variance explained by each 1-Mb window on different chromosomes for all traits are shown in **Figures 4A,B**, respectively. The genetic variance and phenotypic variance explained by the top 1-Mb windows for all traits are listed in **Table 3**. Among a total of 2014 windows, the top window explained 0.95–4.41% of the genetic variance and 0.29–1.13% of the phenotypic variance of the weight traits. For the carcass composition traits, the top window explained 2.00–3.00% of the genetic variance and 0.39–0.45% of the phenotypic variance. The top windows for all traits overlapped with the significant regions obtained from the GWAS.

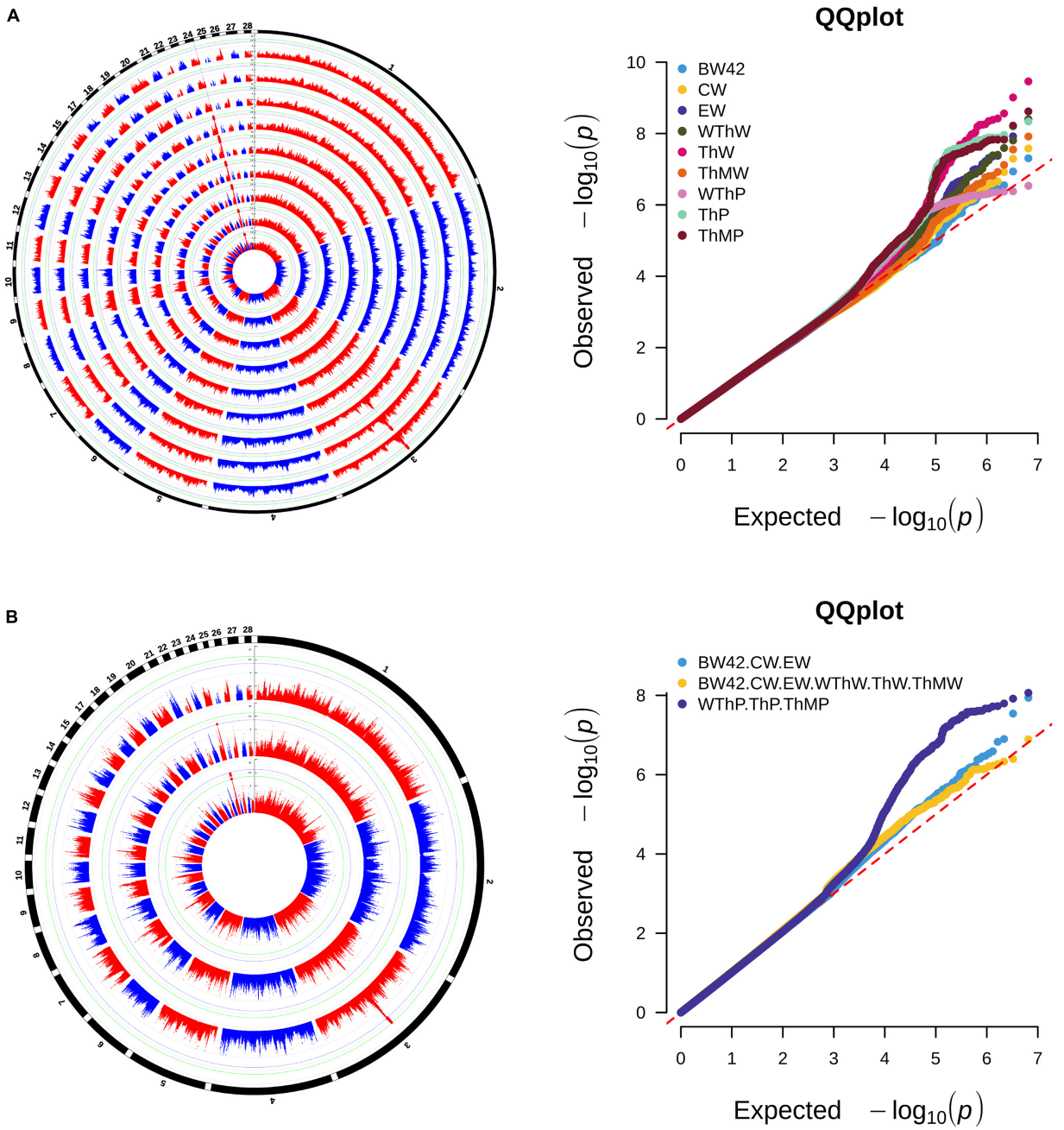

FIGURE 2 | Manhattan and quantile-quantile plots of univariate GWAS **(A)** and multivariate GWAS **(B)** for carcass traits. Each dot represents an SNP in the dataset. The circle Manhattan plots from the inside to the outside are BW42, CW, EW, WThW, ThW, ThMW, WThP, ThP, and ThMP. The horizontal blue and green lines indicate the thresholds for genome-wide significance (*P*-value = 7.64e-9) and suggestive significance (*P*-value = 1.53e-7), respectively.

Estimation of SNP Effect Size

The Manhattan plots of the SNP effect size are shown in **Figure 5**. The SNPs highlighted in red were suggestively significant in GWAS. The SNP effect size of the significant SNPs in GWAS of the six weight traits was between 0.008 and 0.155. The SNP effect size of the significant SNPs in GWAS of the three carcass composition traits was between −3.01e−4 and −2.06e−4. The SNP effect size of the significant SNPs in GWAS of the weight traits and carcass composition traits reached the top 0.24 and 0.26% of the whole-genome SNPs, respectively.

Phylogenetic Analysis

The phylogenetic analysis showed that the Line B chickens were separated from the Chahua and Daweishan mini chickens based on whole-genome SNPs, the SNPs within *DRD2,* or the SNPs within *ADGRG6*, respectively (**Figures 6A–C**). In contrast,

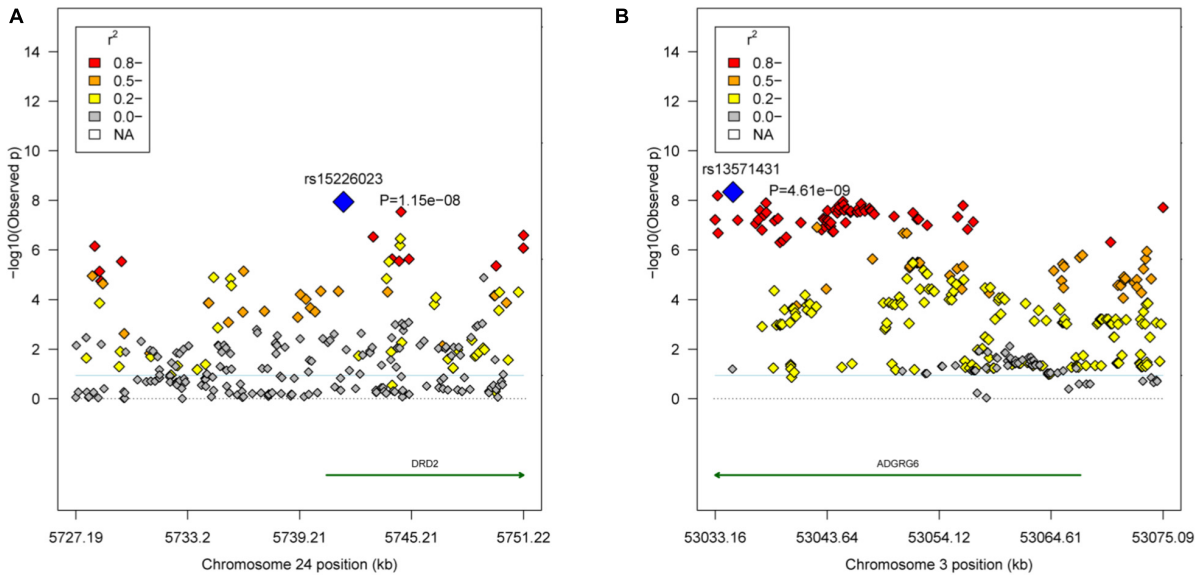

FIGURE 3 | Regional association plots of the candidate areas. **(A)** The QTL region GGA24: 5.73–5.75 Mb associated with the weight traits. The blue dot represents the lead SNP rs15226023. **(B)** The QTL region GGA3: 53.03–53.08 Mb associated with the carcass composition traits. The blue dot represents the lead SNP rs13571431. Different levels of linkage disequilibrium (LD) between the lead SNP and the surrounding SNPs are shown in different colors (red: $r^2 \geq 0.8$; orange: $0.5 \leq r^2 < 0.8$; yellow: $0.2 \leq r^2 < 0.5$; and gray: $r^2 < 0.2$). The gene annotations were obtained from the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu).

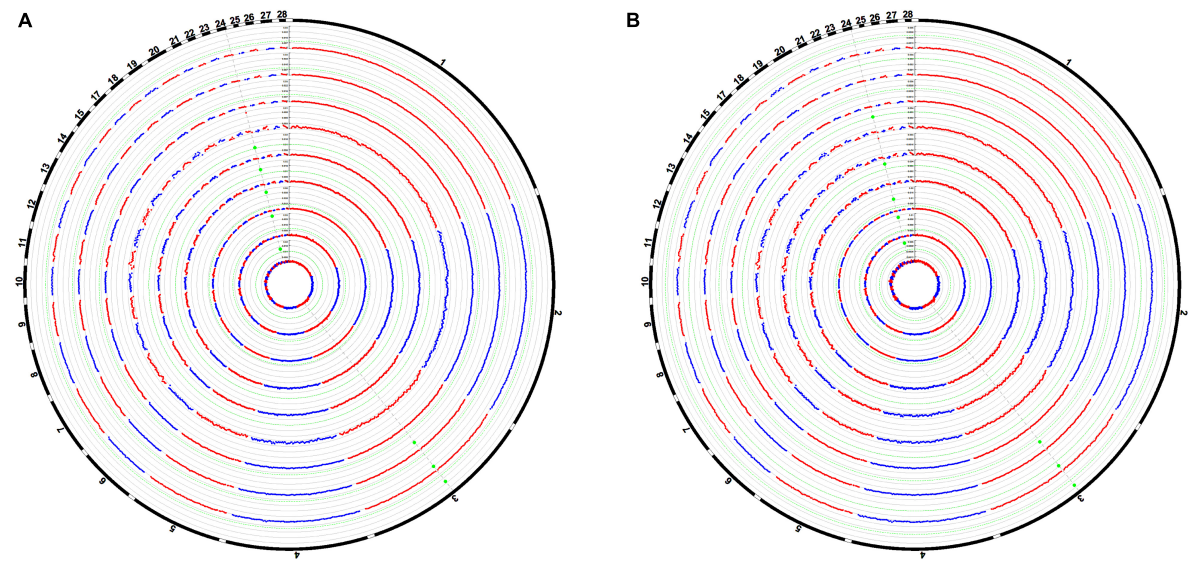

FIGURE 4 | Manhattan plots of the average proportion of genetic variance **(A)** and phenotypic variance **(B)** explained by the 1-Mb window for all traits. Each dot represents a 1-Mb window. The circle Manhattan plots from the inside to the outside are BW42, CW, EW, WThW, ThW, ThMW, WThP, ThP, and ThMP. The dotted green lines indicate the 1.00% threshold of genetic variance and the 0.30% threshold of the phenotypic variance explained by the windows.

the Line B chickens were not separated from the Chahua and Daweishan mini chickens when five random-selected 100 Kb regions were used (**Figures 6D–H**).

Selective Sweep Analysis

A genome-wide selective sweep analysis was performed with the fast-growing Line B chickens, the Chahua chickens, and the Daweishan mini chickens (**Figure 7**). The F_{ST} values in *DRD2* were between 0.11 and 0.21, and the F_{ST} values in *ADGRG6* were between 0.17 and 0.40. Only the values in *ADGRG6* reached the top 5% threshold (0.38) of the whole genome. The $\log_2(\pi\ ratio)$ values in *DRD2* were between 0.07 and 0.20, and those in *ADGRG6* were between -0.04 and 0.27, reaching the top 10% threshold (-0.04) of the whole genome. The zHp values in *DRD2*

TABLE 3 | genetic variance and phenotypic variance explained by top 1-Mb windows for all traits using Bayes C_{π}.

Top window[1]	Traits[2]	GVE (%)[3]	PVE (%)[4]
Chr 24 (5.05–5.97 Mb)	BW42	1.31	0.47
	CW	2.92	0.91
	EW	4.41	1.13
	WThW	1.43	0.39
	ThW	1.08	0.29
	ThMW	0.95	0.31
Chr 3 (53.02–54.00 Mb)	WThP	2.51	0.40
	ThP	3.00	0.39
	ThMP	2.00	0.45

[1]*Window position in galGal6.*
[2]*Body weight at 42 days of age (BW42), carcass weight (CW), eviscerated weight (EW), whole thigh weight (WThW), thigh weight (ThW), thigh muscle weight (ThMW), whole thigh percentage (WThP), thigh percentage (ThP), thigh muscle percentage (ThMP), and feed intake from 28 to 42 days of age (FI).*
[3]*Average proportion of genetic variance explained by 1-Mb window.*
[4]*Average proportion of phenotypic variance explained by 1-Mb window.*

were between 0.18 and 0.45, and those in *ADGRG6* were between -1.37 and 0.19, reaching the top 10% threshold (-1.37) of the whole genome. The XP-EHH values in *DRD2* were between -0.12 and 0.30, and those in *ADGRG6* were between 0.64 and 1.02. Overall, only the *ADGRG6* region showed some degree of genetic and nucleotide differentiation between large and small breeds.

DISCUSSION

Genotype imputation has been widely used in GWAS to boost power (Spencer et al., 2009). This method can aid in identifying many novel SNPs and QTLs associated with phenotypes of interest. In previous GWAS, imputation from low-density SNP chip genotypes to the WGS level was implemented in chickens (Huang et al., 2018; Li D. et al., 2020), pigs (Yan et al., 2018), and cattle (Höglund et al., 2015). Imputed genotypes with sufficiently high imputation accuracy are necessary to obtain reliable results in follow-up analyses, such as GWAS. Ni et al. (2017) reported that the post-imputation filtering criterion should be 0.80 to ensure the high accuracy of the imputed WGS data. In the current study, the average allelic R^2 value and the genotype concordance rate between the imputed and true genotypes were 0.89 and 0.91, respectively.

The heritability estimates for the six weight traits were moderate to high (0.30–0.39), showing high consistency with previous reports (Demeure et al., 2013). However, the estimates were lower than 0.56 in medium-growing broilers at 44 days of age (Xu et al., 2016). Our heritability estimate for ThP was 0.22, which was slightly lower than 0.37, as reported by Demeure et al. (2013). The likely reason is the genetic background difference in the chicken lines.

A 24.08 Kb QTL on GGA24 (GGA24: 5.73–5.75 Mb) was identified for the six weight traits. According to the Animal QTL Database[8], this region has been recorded as a QTL (GGA24: 4.3–6.0 Mb) for BW08 in an F2 population (Taiwan local chicken line L2 × experimental Rhode Island Red line; Lien et al., 2017). This region contains only one candidate gene (*DRD2*), which encodes the D2 subtype of the dopamine receptor. The gene is highly expressed in the basal ganglia (Missale et al., 1998), which is a control center for movement. Neurotransmission mediated by *DRD2* is known to have a key role in the control of movement. *DRD2* TaqIA polymorphisms were correlated with body mass in previous studies (Spitz et al., 2000; Thomas et al., 2001). The

[8]https://www.animalgenome.org/cgi-bin/QTLdb/GG/index

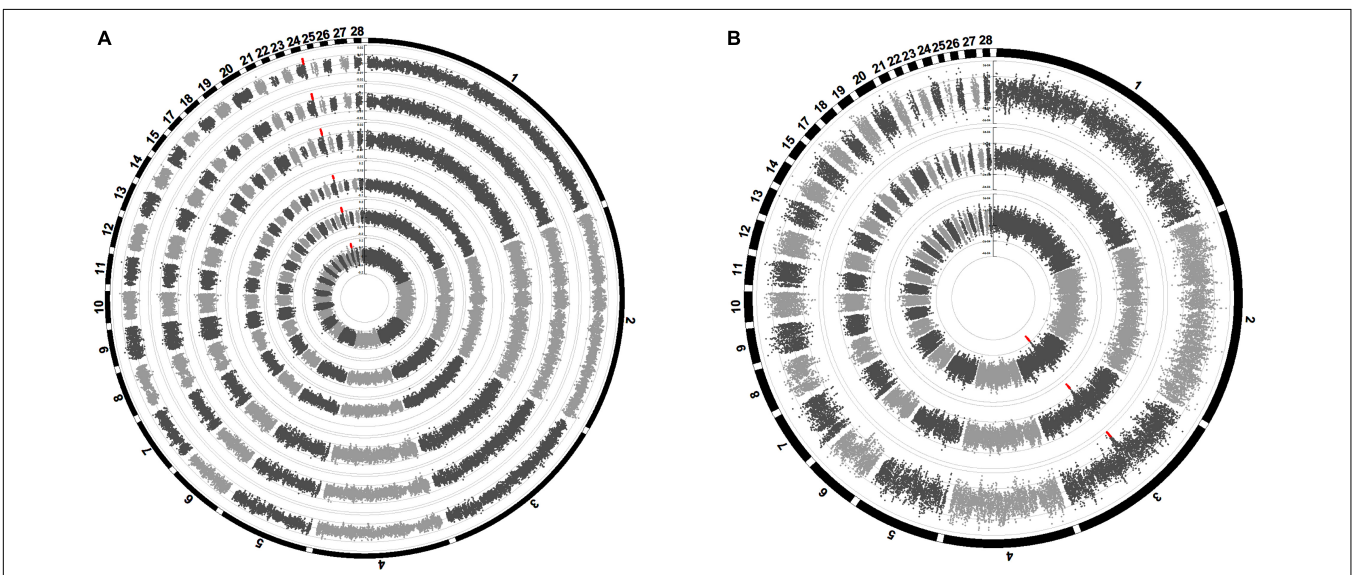

FIGURE 5 | Manhattan plots of SNP effect sizes for the carcass traits **(A)** and carcass composition traits **(B)**. Each dot represents an SNP. The SNP effect sizes are shown on the *y*-axis. The circle Manhattan plots from the inside to the outside are BW42, CW, EW, WThW, ThW, ThMW, WThP, ThP, and ThMP. The SNPs highlighted in red are suggestively significant in GWAS.

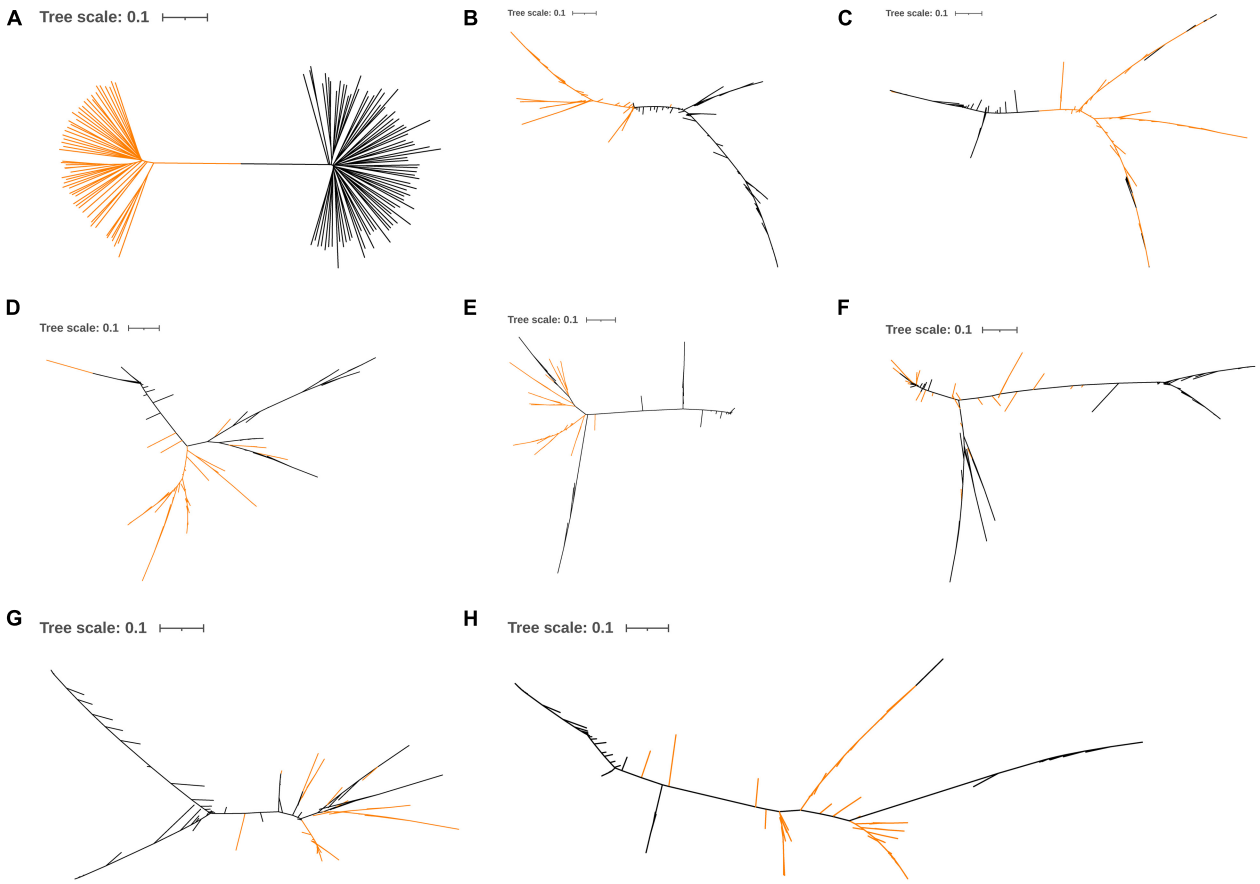

FIGURE 6 | Neighbor-joining tree based on the whole-genome SNPs **(A)**, candidate genes **(B,C)**, and 5 random 100-Kb regions **(D–H)**. The yellow line represents the Chahua chickens and Daweishan mini chickens, and the black line represents Line B.

density of the *DRD2* receptors can affect food reinforcement to influence energy intake (Epstein et al., 2007). Down-regulation of *DRD2* receptors leads to increased food intake and weight gain (Epstein et al., 2007). In the current study, significant differences in weight traits and feed intake were observed between birds with different genotypes. The mRNA expression of *DRD2* in the thigh muscle was not detected by quantitative polymerase chain reaction (Q-PCR; data not shown). The Galbase data shows that *DRD2* has high mRNA expression in chickens' brain tissue[9]. Therefore, brain tissue should be obtained and tested in a future study.

We focused specifically on the thigh traits in the carcass composition traits in the current study. A 42.52 kb genomic region on GGA3 (GGA3: 53.03–53.08 Mb) was identified for WThP and ThP. This QTL was not reported previously. This region contained only one candidate gene (*ADGRG6*), which encodes the G-protein-coupled receptor 126. Ravenscroft et al. (2015) proved that *ADGRG6* is critical for myelination of peripheral nerves in humans and mutations of *ADGRG6* are responsible for severe arthrogryposis multiplex congenita. Soranzo et al. (2009) found that mutations in *ADGRG6* are related to trunk length, hip axis length, and height. Numerous studies have shown that *ADGRG6* is associated with adult height and pediatric stature (Hirschhorn et al., 2001; Xu et al., 2002; Gudbjartsson et al., 2008; Lettre et al., 2008; Sovio et al., 2009; Liu et al., 2010; Zhao et al., 2010). Kou et al. (2013) found that *ADGRG6* was highly expressed in cartilage, and the knockdown of *ADGRG6* in zebrafish caused delayed ossification of the developing spine. It is believed that *ADGRG6* may affect both adolescent idiopathic scoliosis susceptibility and height through abnormal spinal development and/or growth. Karner et al. (2015) also proved that the loss of *ADGRG6* in osteochondroprogenitor cells alters cartilage biology and spinal column development.

Regarding weight traits, the significant region of BW42, CW, and EW detected by multivariate GWAS was consistent with that detected by univariate GWAS. The significance levels of BW42, CW, EW, WThW, ThW, and ThMW were lower in the multivariate GWAS than the univariate GWAS. Therefore, it was believed that the significant region on the GGA24 had a greater impact on the overall body weight of the chicken than on the thigh weight. Regarding carcass composition traits, the significant region of WThP and ThP detected by multivariate GWAS was consistent with that detected by univariate GWAS. Therefore,

[9]http://animal.nwsuaf.edu.cn/code/index.php/ChickenVar

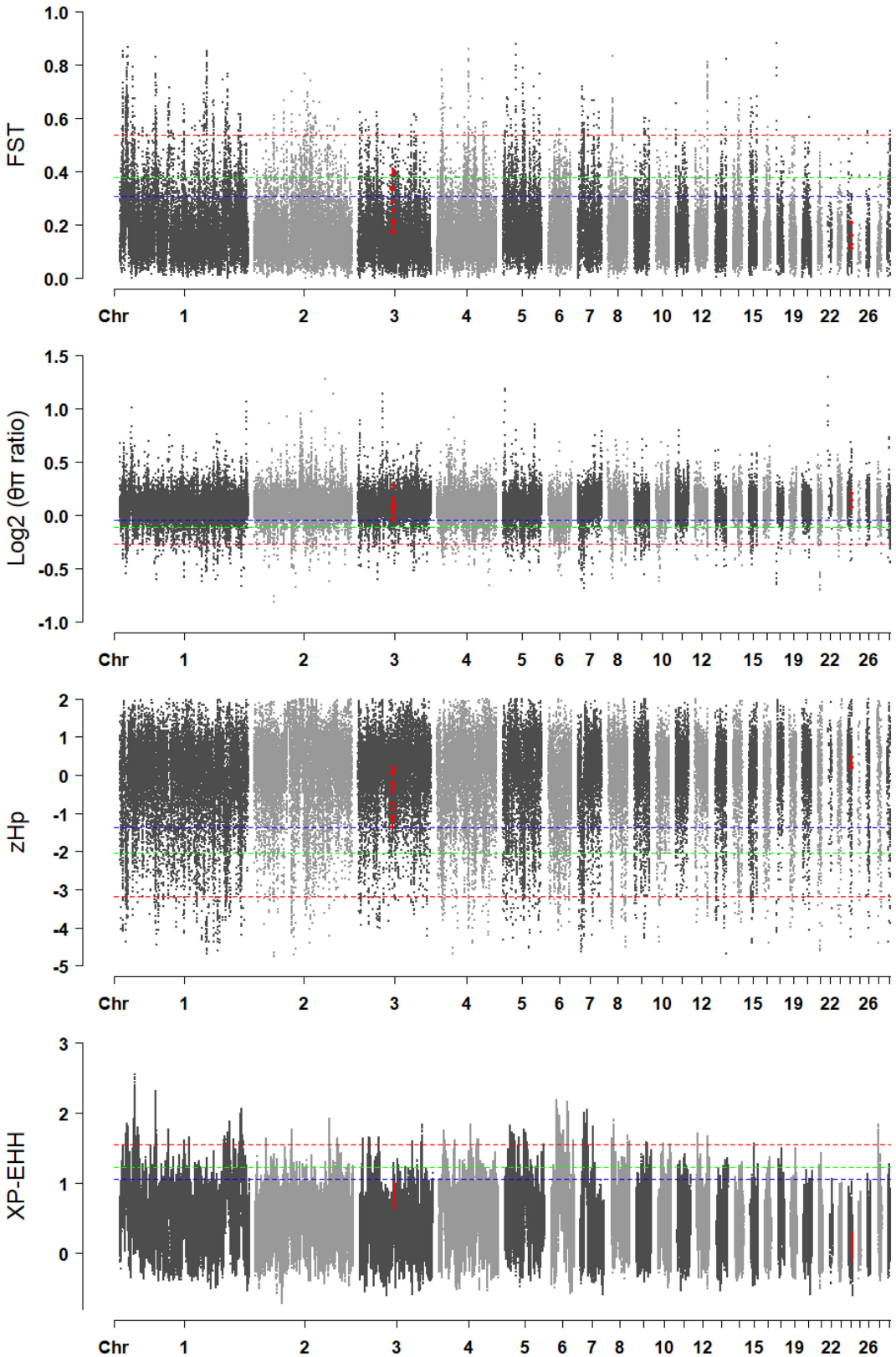

FIGURE 7 | The visualization of the selective sweep analysis. The horizontal blue, green, and red lines indicate the thresholds for 10% genome-wide significance, 5% genome-wide significance, and 1% genome-wide significance, respectively.

it was believed that the significant region on the GGA3 had an impact on WThP, ThP, and ThMP.

It was interesting that significant differences in thigh composition traits were observed between birds with different genotypes of the associated variants. The flavor of thigh meat is usually better than that of the breast muscle because its intramuscular fat content is three times higher than that of the latter (Zhao et al., 2007; Fu et al., 2014). The significant SNPs could be used in genomic selection programs to improve the percentage of thigh meat (Zhang et al., 2015).

In the current study, the results of the Bayesian analysis and the estimation of the SNP effect size verified the GWAS results. The SNP effect sizes of the significant SNPs in GWAS of the weight and carcass composition traits reached the top 0.24 and 0.26% of the whole genome SNPs, respectively. Among all traits, the top 1-Mb sliding windows, which overlapped with the significant regions in GWAS, explained 0.95–4.41% of genetic variation and 0.29–1.13% of phenotypic variation. In contrast, the remaining 1-Mb sliding window explained no more than 0.24% of genetic variation and no more than 0.08% of phenotypic variation.

In the study by Yoshida et al. (2017), the genes located in the top ten 1-Mb windows were identified as strong functional candidate genes, and the top window for the growth traits explained 3.71 and 3.61% of genetic variance, respectively. In a study of a pure line of broilers, the window with the largest effect for the bodyweight, breast meat, and leg score explained 2.5, 1.14, and 1.12% of the genetic variation, respectively (Fragomeni Bde et al., 2014).

In the current study, the phylogenetic tree indicated genetic differentiation in *DRD2* and *ADGRG6* between the Line B chickens and the small-sized breeds because the former could not be completely separated from the Chahua and Daweishan mini chickens when five random-selected 100 Kb regions were used. In the selective sweep analysis of chickens, windows with thresholds of 1 or 5% outliers were typically identified as candidate regions (Yin et al., 2019; Li W. et al., 2020; Luo et al., 2020; Weng et al., 2020). The F_{ST} values in *ADGRG6* reached the top 5% threshold, whereas the $\log_2(\pi$ ratio) values and zHp values reached the top 10% threshold of the whole genome. *ADGRG6* showed some degree of genetic and nucleotide differentiation between the fast-growing Line B chickens and the Chahua and Daweishan mini chickens. However, the *DRD2* showed no related signals. Since the significant SNPs were located in the introns and the upstream and intergenic regions of the candidate genes, it is challenging to investigate causative mutations, which will be performed in a future study.

CONCLUSION

In summary, the genomic heritability estimates of nine chicken carcass traits ranged from moderate to high (0.21 to 0.39). Twelve genome-wide significant SNPs and 118 suggestively significant SNPs were detected. One 24.08 Kb region (GGA24: 5.73–5.75 Mb) for six weight traits and one 42.52 Kb region (GGA3: 53.03–53.08 Mb) for three thigh-related carcass traits were identified. The significant SNPs could be used in genomic selection programs to improve the weight traits and thigh composition traits. In the QTL regions, *DRD2* was the only major-effect candidate gene for weight traits, and *ADGRG6* was the only major-effect candidate gene for carcass composition traits. Some degree of genetic differentiation in *ADGRG6* between large-sized and small-sized breeds was observed. Our results supply essential information for causative mutation identification of carcass traits in broilers.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository and accession numbers can be found below: https://bigd.big.ac.cn/gsa, CRA002454 and CRA004023.

ETHICS STATEMENT

All experimental procedures with broilers were performed according to the Guidelines for Experimental Animals established by the Ministry of Science and Technology (Beijing, China). Ethical approval on animal survival was given by the animal welfare and ethics committee of the Institute of Animal Sciences (IAS) and the Chinese Academy of Agricultural Sciences (CAAS, Beijing, China) with the following reference number: IAS2019–44.

AUTHOR CONTRIBUTIONS

XY contributed to data collection, data analysis and interpretation, and wrote the manuscript. JS contributed to data collection and manuscript revision. GZ, MZ, and JW designed the research and revised manuscript. WL, XT, FF, and DL contributed to the chicken raising, sampling, and data collection. RL designed the research and contributed to data collection, data analysis and interpretation, and wrote the manuscript. All authors submitted comments on the draft and approved the final manuscript.

FUNDING

This research was supported by grants from the National Natural Science Foundation of China (31772591), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202005), and the joint research project of white feather broiler breeding of the Ministry of Agriculture and Rural Affairs of the People Republic of China (19200133).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.645107/full#supplementary-material

REFERENCES

Barbato, M., Orozco-Terwengel, P., Tapio, M., and Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* 6:109. doi: 10.3389/fgene.2015.00109

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

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

Canela-Xandri, O., Law, A., Gray, A., Woolliams, J. A., and Tenesa, A. (2015). A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. *Nat. Commun.* 6:10162.

Claire D'Andre, H., Paul, W., Shen, X., Jia, X., Zhang, R., Sun, L., et al. (2013). Identification and characterization of genes that control fat deposition in chickens. *J. Anim. Sci. Biotechnol.* 4:43.

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

Demeure, O., Duclos, M. J., Bacciu, N., Le Mignon, G., Filangi, O., Pitel, F., et al. (2013). Genome-wide interval mapping using SNPs identifies new QTL for growth, body composition and several physiological variables in an F 2 intercross between fat and lean chicken lines. *Genet. Sel. Evol.* 45:36.

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

Epstein, L. H., Temple, J. L., Neaderhiser, B. J., Salis, R. J., Erbe, R. W., and Leddy, J. J. (2007). Food reinforcement, the dopamine D2 receptor genotype, and energy intake in obese and nonobese humans. *Behav. Neurosci.* 121, 877–886. doi: 10.1037/0735-7044.121.5.877

Flisar, T., Malovrh, Š, Terèiè, D., Holcman, A., and Kovaè, M. (2014). Thirty-four generations of divergent selection for 8-week body weight in chickens. *Poultry Sci.* 93, 16–23. doi: 10.3382/ps.2013-03464

Fragomeni Bde, O., Misztal, I., Lourenco, D. L., Aguilar, I., Okimoto, R., and Muir, W. M. (2014). Changes in variance explained by top SNP windows over generations for three traits in broiler chicken. *Front. Genet.* 5:332. doi: 10.3389/fgene.2014.00332

Fu, R. Q., Liu, R. R., Zhao, G. P., Zheng, M. Q., Chen, J. L., and Wen, J. (2014). Expression profiles of key transcription factors involved in lipid metabolism in Beijing-You chickens. *Gene* 537, 120–125. doi: 10.1016/j.gene.2013.07.109

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Thompson, R., et al. (2009). *ASReml User Guide Release 3.0*. Hemel Hempstead: VSN International Ltd.

Gu, X., Feng, C., Ma, L., Song, C., Wang, Y., Da, Y., et al. (2011). Genome-Wide association study of body weight in chicken F2 resource population. *PLoS One* 6:e21872. doi: 10.1371/journal.pone.0021872

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., et al. (2008). Many sequence variants affecting diversity of adult human height. *Nat. Genet.* 40, 609–615. doi: 10.1038/ng.122

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

Hall, S. J. G. (2016). Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal* 10, 1778–1785. doi: 10.1017/s1751731116000914

Hirschhorn, J. N., Lindgren, C. M., Daly, M. J., Kirby, A., Schaffner, S. F., Burtt, N. P., et al. (2001). Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *Am. J. Hum. Genet.* 69, 106–116. doi: 10.1086/321287

Höglund, J. K., Buitenhuis, B., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2015). Genome-wide association study for female fertility in Nordic Red cattle. *BMC Genetics* 16:110. doi: 10.1186/s12863-015-0269-x

Huang, S., He, Y., Ye, S., Wang, J., Yuan, X., Zhang, H., et al. (2018). Genome-wide association study on chicken carcass traits using sequence data imputed from SNP array. *J. Appl. Genet.* 59, 335–344. doi: 10.1007/s13353-018-0448-3

Karner, C. M., Long, F., Solnica-Krezel, L., Monk, K. R., and Gray, R. S. (2015). Gpr126/Adgrg6 deletion in cartilage models idiopathic scoliosis and pectus excavatum in mice. *Hum. Mol. Genet.* 24, 4365–4373. doi: 10.1093/hmg/ddv170

Kou, I., Takahashi, Y., Johnson, T. A., Takahashi, A., Guo, L., Dai, J., et al. (2013). Genetic variants in GPR126 are associated with adolescent idiopathic scoliosis. *Nat. Genet.* 45:676.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* 40, 584–591. doi: 10.1038/ng.125

Li, D., Sun, G., Zhang, M., Cao, Y., Zhang, C., Fu, Y., et al. (2020). Breeding history and candidate genes responsible for black skin of Xichuan black-bone chicken. *BMC Genomics* 21:511. doi: 10.1186/s12864-020-06900-8

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [preprint] arXiv:1303.3997

Li, W., Liu, R., Zheng, M., Feng, F., Liu, D., Guo, Y., et al. (2020). New insights into the associations among feed efficiency, metabolizable efficiency traits and related QTL regions in broiler chickens. *J. Anim. Sci. Biotechnol.* 11:65.

Li, W. J., Zhao, G. P., Chen, J. L., Zheng, M. Q., and Wen, J. (2009). Influence of dietary vitamin E supplementation on meat quality traits and gene expression related to lipid metabolism in the Beijing-you chicken. *Br. Poultry Sci.* 50, 188–198. doi: 10.1080/00071660902755409

Lien, C.-Y., Tixier-Boichard, M., Wu, S.-W., Wang, W.-F., Ng, C. S., and Chen, C.-F. (2017). Detection of QTL for traits related to adaptation to sub-optimal climatic conditions in chickens. *Genet. Sel. Evol.* 49:39.

Liu, J. Z., Medland, S. E., Wright, M. J., Henders, A. K., Heath, A. C., Madden, P. a. F., et al. (2010). Genome-wide association study of height and body mass index in Australian twin families. *Twin Res. Hum. Genet.* 13, 179–193. doi: 10.1375/twin.13.2.179

Liu, R., Sun, Y., Zhao, G., Wang, F., Wu, D., Zheng, M., et al. (2013). Genome-Wide association study identifies loci and candidate genes for body composition and meat quality traits in Beijing-You chickens. *PLoS One* 8:e61172. doi: 10.1371/journal.pone.0061172

Liu, R., Sun, Y., Zhao, G., Wang, H., Zheng, M., Li, P., et al. (2015). Identification of loci and genes for growth related traits from a genome-wide association study in a slow- × fast-growing broiler chicken cross. *Genes Genom.* 37, 829–836. doi: 10.1007/s13258-015-0314-1

Liu, R., Xing, S., Wang, J., Zheng, M., Cui, H., Crooijmans, R. P. M. A., et al. (2019). A new chicken 55K SNP genotyping array. *BMC Genomics* 20:410. doi: 10.1186/s12864-019-5736-8

Luo, W., Luo, C., Wang, M., Guo, L., Chen, X., Li, Z., et al. (2020). Genome diversity of Chinese indigenous chicken and the selective signatures in Chinese gamecock chicken. *Sci. Rep.* 10:14532.

Missale, C., Nash, S. R., Robinson, S. W., Jaber, M., and Caron, M. G. (1998). Dopamine receptors: from structure to function. *Physiol. Rev.* 78, 189–225.

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

Olszewski, P. K., Rozman, J., Jacobsson, J. A., Rathkolb, B., Strömberg, S., Hans, W., et al. (2012). Neurobeachin, a regulator of synaptic protein targeting, is associated with body fat mass and feeding behavior in mice and body-mass index in humans. *PLoS Genet.* 8:e1002568. doi: 10.1371/journal.pgen.1002568

Plassais, J., Rimbault, M., Williams, F. J., Davis, B. W., Schoenebeck, J. J., and Ostrander, E. A. (2017). Analysis of large versus small dogs reveals three genes on the canine X chromosome associated with body weight, muscling and back fat thickness. *PLoS Genet.* 13:e1006661. doi: 10.1371/journal.pgen.1006661

Ravenscroft, G., Nolent, F., Rajagopalan, S., Meireles, A. M., Paavola, K. J., Gaillard, D., et al. (2015). Mutations of GPR126 are responsible for severe arthrogryposis multiplex congenita. *Am. J. Hum. Genet.* 96, 955–961.

Rong, H., Chen, X., Xiong, B., Wan, Q., Cao, Z., Chen, B., et al. (2011). Study on the germplasm characteristic of Yunnan Daweishan mini chickens. *J. Yunnan Agric. Univ.* 26, 48–63.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J. B., Hammond, N., et al. (2009). Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genet.* 5:e1000445. doi: 10.1371/journal.pgen.1000445

Sovio, U., Bennett, A. J., Millwood, I. Y., Molitor, J., O'reilly, P. F., Timpson, N. J., et al. (2009). Genetic determinants of height growth assessed longitudinally from infancy to adulthood in the northern Finland birth cohort 1966. *PLoS Genet.* 5:e1000409. doi: 10.1371/journal.pgen.1000409

Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* 5:e1000477. doi: 10.1371/journal.pgen.1000477

Spitz, M. R., Detry, M. A., Pillow, P., Hu, Y., Amos, C. I., Hong, W. K., et al. (2000). Variant alleles of the D2 dopamine receptor gene and obesity. *Nutr. Res.* 20, 371–380. doi: 10.1016/s0271-5317(00)00130-5

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

Takasuga, A., Sato, K., Nakamura, R., Saito, Y., Sasaki, S., Tsuji, T., et al. (2015). Non-synonymous FGD3 variant as positional candidate for disproportional tall stature accounting for a carcass weight QTL (CW-3) and skeletal dysplasia in Japanese black cattle. *PLoS Genet.* 11:e1005433. doi: 10.1371/journal.pgen.1005433

Thomas, G. N., Critchley, J. a. J. H., Tomlinson, B., Cockram, C. S., and Chan, J. C. N. (2001). Relationships between the taqI polymorphism of the dopamine D2 receptor and blood pressure in hyperglycaemic and normoglycaemic Chinese subjects. *Clin. Endocrinol.* 55, 605–611. doi: 10.1046/j.1365-2265.2001.01404.x

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

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.11–11.10.33.

Van Laere, A.-S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 425, 832–836. doi: 10.1038/nature02064

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Wang, Y., Cao, X., Luo, C., Sheng, Z., Zhang, C., Bian, C., et al. (2020). Multiple ancestral haplotypes harboring regulatory mutations cumulatively contribute to a QTL affecting chicken growth traits. *Commun. Biol.* 3:472.

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

Weng, Z., Xu, Y., Li, W., Chen, J., Zhong, M., Zhong, F., et al. (2020). Genomic variations and signatures of selection in Wuhua yellow chicken. *PLoS One* 15:e0241137. doi: 10.1371/journal.pone.0241137

Xu, J., Bleecker, E. R., Jongepier, H., Howard, T. D., Koppelman, G. H., Postma, D. S., et al. (2002). Major recessive gene (s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. *Am. J. Hum. Genet.* 71, 646–650. doi: 10.1086/342216

Xu, Z., Ji, C., Zhang, Y., Zhang, Z., Nie, Q., Xu, J., et al. (2016). Combination analysis of genome-wide association and transcriptome sequencing of residual feed intake in quality chickens. *BMC Genomics* 17:594. doi: 10.1186/s12864-016-2861-5

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

Yin, H., Li, D., Wang, Y., and Zhu, Q. (2019). Whole-genome resequencing analysis of Pengxian Yellow Chicken to identify genome-wide SNPs and signatures of selection. *3 Biotech* 9:383.

Yoshida, G. M., Lhorente, J. P., Carvalheiro, R., and Yáñez, J. M. (2017). Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (*Salmo salar* L.). *Anim. Genet.* 48, 698–703. doi: 10.1111/age.12621

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., et al. (2015). Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. *G3 Genes Genom. Genet.* 5, 615–627. doi: 10.1534/g3.114.016261

Zhao, G. P., Chen, J. L., Zheng, M. Q., Wen, J., and Zhang, Y. (2007). Correlated responses to selection for increased intramuscular fat in a Chinese quality chicken line. *Poultry Sci.* 86, 2309–2314. doi: 10.1093/ps/86.11.2309

Zhao, J., Li, M., Bradfield, J. P., Zhang, H., Mentch, F. D., Wang, K., et al. (2010). The role of height-associated loci identified in genome wide association studies in the determination of pediatric stature. *BMC Med. Genet.* 11:96. doi: 10.1186/1471-2350-11-96

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest: FF and DL were employed by the Foshan Gaoming Xinguang Agricultural and Animal Industrials Corporation.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 April 2021
doi: 10.3389/fgene.2021.628205

Genomic Prediction Using Bayesian Regression Models With Global–Local Prior

Shaolei Shi[1], Xiujin Li[2], Lingzhao Fang[3], Aoxing Liu[4], Guosheng Su[4], Yi Zhang[1], Basang Luobu[5], Xiangdong Ding[1] and Shengli Zhang[1]**

[1] National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China, [2] Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, China, [3] Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, United Kingdom, [4] Center for Quantitative Genetics and Genomics, Aarhus University, Tjele, Denmark, [5] Shannan Animal Husbandry and Veterinary Master Station, Shannan, China

OPEN ACCESS

Edited by:
Fabyano Fonseca Silva, Universidade Federal de Viçosa, Brazil

Reviewed by:
Hao Cheng, University of California, Davis, United States
Miriam Piles, Institute of Agrifood Research and Technology (IRTA), Spain
Idalmo Pereira, Federal University of Minas Gerais, Brazil

***Correspondence:**
Xiangdong Ding
xding@cau.edu.cn
Shengli Zhang
zhangslcau@cau.edu.cn

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

Received: *11 November 2020*
Accepted: *02 March 2021*
Published: *15 April 2021*

Citation:
Shi S, Li X, Fang L, Liu A, Su G, Zhang Y, Luobu B, Ding X and Zhang S (2021) Genomic Prediction Using Bayesian Regression Models With Global–Local Prior. Front. Genet. 12:628205. doi: 10.3389/fgene.2021.628205

Bayesian regression models are widely used in genomic prediction for various species. By introducing the global parameter τ, which can shrink marker effects to zero, and the local parameter λ_k, which can allow markers with large effects to escape from the shrinkage, we developed two novel Bayesian models, named BayesHP and BayesHE. The BayesHP model uses Horseshoe+ prior, whereas the BayesHE model assumes local parameter λ_k, after a half-t distribution with an unknown degree of freedom. The performances of BayesHP and BayesHE models were compared with three classical prediction models, including GBLUP, BayesA, and BayesB, and BayesU, which also applied global–local prior (Horseshoe prior). To assess model performances for traits with various genetic architectures, simulated data and real data in cattle (milk production, health, and type traits) and mice (type and growth traits) were analyzed. The results of simulation data analysis indicated that models based on global–local priors, including BayesU, BayesHP, and BayesHE, performed better in traits with higher heritability and fewer quantitative trait locus. The results of real data analysis showed that BayesHE was optimal or suboptimal for all traits, whereas BayesHP was not superior to other classical models. For BayesHE, its flexibility to estimate hyperparameter automatically allows the model to be more adaptable to a wider range of traits. The BayesHP model, however, tended to be suitable for traits having major/large quantitative trait locus, given its nature of the "U" type-like shrinkage pattern. Our results suggested that auto-estimate the degree of freedom (e.g., BayesHE) would be a better choice other than increasing the local parameter layers (e.g., BayesHP). In this study, we introduced the global–local prior with unknown hyperparameter to Bayesian regression models for genomic prediction, which can trigger further investigations on model development.

Keywords: half-Cauchy, half-t distribution, Horseshoe+ prior, hyperparameter estimating, Horseshoe

INTRODUCTION

The genomic prediction has been widely applied in animal and plant breeding. The statistical method being used is one of the critical factors for the accuracy of genomic estimated breeding value and consequently impacting the genetic gain achieved by genomic prediction. Models commonly used for predicting genomic estimated breeding value can be divided into two categories: (i) methods based on the framework of best linear unbiased prediction (BLUP), such as GBLUP (Habier et al., 2007; VanRaden, 2008); (ii) a set of Bayesian regression models – such as BayesA and BayesB (Meuwissen et al., 2001), which are also called as "Bayesian alphabet" models (Gianola et al., 2009). GBLUP assumes that the effects of all genetic markers are normally distributed and share the same variance, thus fitting well for traits with polygenic inheritance. In Bayesian methods, the marker effects are relevant to the choice of the prior probability distribution. By giving different priors, the Bayesian models can fit well for traits with different genetic architectures. For example, the widely used BayesA and BayesB models allow effects of genetic markers to follow a heavy-tailed distribution and therefore in line with the real distribution of marker effects for traits with large quantitative trait locus (QTL).

Bayesian regression models can be further classified into a one-group model and a two-group model from the perspective of the number of groups of genetic markers being used when estimating marker effects. The one-group model is generally a variable shrinkage model that shrinks the effects of some markers toward zero, such as BayesA (Meuwissen et al., 2001). The two-group model (or spike-and-slab model), can also be a multigroup model, is generally a variable selection model that only selects a subset of markers to be included in the model and assumes the remaining markers to have zero effects, such as BayesB (Meuwissen et al., 2001) and BayesC (Habier et al., 2011). The shrinkage process and the selection process can also be combined in a Bayesian regression model. In such a model, a subset of markers was selected to be included in the model, and then, the effects of some selected markers were further shrunk toward zero. BayesCπ (Habier et al., 2011) is an example of this type of model. Compared with BayesB and BayesC, BayesCπ can estimate the proportion of genetic markers with a non-zero effect based on the data. However, BayesCπ could be challenged by problems such as poor convergence and mixing in some situations (Wolc et al., 2011).

A Bayesian regression model with "global–local" shrinkage prior could be a good alternative for genomic prediction. With the global–local prior, the variances of marker effects can be shaped by global and local parameters simultaneously. The global parameter, τ, can shrink the marker effects to approach zero, whereas the local parameter, λ_k, allows a marker to escape from the shrinkage when the marker has a big effect (Piironen and Vehtari, 2017). Horseshoe prior (Carvalho et al., 2010) is one of the most popular estimators of global–local prior. In Horseshoe prior, the local parameter follows a positive half-Cauchy distribution, which is a special case of half-t distribution where the degree of freedom is one. The Horseshoe prior has a similar form as the one-group model, but its prior can lead to a "pseudo-posterior," which shows the same pattern as the "two-group" model (Bhadra et al., 2017).

Horseshoe prior has already been applied to many scenarios, such as genomic prediction (Pong-Wong and Woolliams, 2014), genome-wide association study (Johndrow et al., 2017), and eQTL mapping (Li et al., 2019). Until now, BayesU (Pong-Wong and Woolliams, 2014) is the only model that uses the Horseshoe prior, and BayesU had similar performance with BayesA and BayesB tested with simulation data. However, the performance of BayesU has not yet been tested with real data. To better separate signals and noise, an extension of the Horseshoe estimator, named Horseshoe+ prior (Bhadra et al., 2017), was proposed. Horseshoe+ introduces one more local parameter with a positive half-Cauchy distribution, which leads a heavier tail than using standard Horseshoe prior. The investigation of using Horseshoe+ prior in Bayesian regression models for genomic prediction could be interesting but has not yet been explored previously. Besides, in variable shrinkage and selection models, hyperparameters is a challenge, and many previous studies have tried to estimate hyperparameter to improve prediction accuracy (Habier et al., 2011; Zhu et al., 2016).

This study's objectives were to (i) develop two Bayesian methods for genomic prediction based on the global–local prior, which have the flexibility in estimating hyperparameters, and (ii) to test the model performance with simulated and real data for traits with various genetic architectures.

MATERIALS AND METHODS

Statistical Models

In Bayesian regression models, the differences in different models were the prior assumptions on the effects of single-nucleotide polymorphisms (SNPs). All Bayesian multiple regression model can be described as follows:

$$\boldsymbol{y} = \boldsymbol{\mu} + \sum_{k=1}^{m} \boldsymbol{x_k}\beta_k + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of pre-corrected phenotypes, $\boldsymbol{\mu}$ is the overall mean, $\boldsymbol{x_k}$ is the vector of genotypes for the kth SNP, m is the number of SNPs, β_k is the effect of the kth SNP, and e is a vector of random residuals. The assumptions of the residuals are $\boldsymbol{e} \sim N(0, \boldsymbol{D}\sigma_e^2)$, where σ_e^2 is the random residual variance. The $\boldsymbol{D}$ is an identity matrix when using pre-corrected phenotypes other than de-regressed proofs (DRPs). DRPs were derived from an official estimated breeding value (EBV) with the method that Jairath et al. (1998) suggested. When using DRP as $\boldsymbol{y}$, $\boldsymbol{D}$ is a diagonal matrix with diagonal elements calculated as $d_{ii} = \frac{1-r_i^2}{r_i^2}$, to account for heterogeneities in σ_e^2 due to differences in reliability (r_i^2) of DRP.

In Bayesian inference, a total of 50,000 Markov chain Monte Carlo samples were generated, with the first 20,000 samples discarded as burn-in and every 50th sample of the remaining 30,000 samples saved for inferring posterior statistics. All analyses with Bayesian regression models were conducted using in-house scripts written in Fortran 95 by the first author.

BayesU

The BayesU (Pong-Wong and Woolliams, 2014) was developed based on Horseshoe prior (HS). To make it comparable with other methods, the global prior τ was set as a flat prior. A detailed description is given as:

$$\beta_k \sim N(0, \lambda_k^2\tau^2),\ \lambda_k \sim C^+(0,1)\,, \textit{and}\, \tau \sim \textit{flat}$$

where λ_k and τ are local and global parameters, respectively. The local parameter λ_k follows a positive half-Cauchy distribution, which is a special case of student-t distribution where the degree of freedom equals to 1.

BayesHP

Compared with HS, its modified version Horseshoe+ (HS+) can form a heavier tailed prior distribution by introducing an additional local parameter with a positive half-Cauchy distribution. Regarding the performance, HS+ can better distinguish the signals and noise than the standard HS (Bhadra et al., 2017). However, Horseshoe+ prior (HS+) (Bhadra et al., 2017) has not yet been applied in Bayesian regression models for genomic prediction. In this study, we proposed a novel BayesHP model based on HS+:

$$\beta_k \sim N(0, \lambda_k^2\tau^2),\ \lambda_k \sim C^+(0,\eta_k),\ \eta_k \sim C^+(0,1),$$
$$\textit{and}\ \tau \sim C^+(0, N^{-1})$$

where λ_k and η_k are local parameters, and τ is a global parameter following a positive half-Cauchy distribution with scale parameter equals to N^{-1}. The N is the size of training data, as Bhadra et al. (2017) suggested. Compared with HS, HS+ introduces one more layer of local parameter η_k. As described by Makalic and Schmidt (2016) and Makalic et al. (2016) (**Appendix**), the half-Cauchy distribution can be modeled as a scale mixture of inverse gamma distributions: if $x^2 \sim IG\left(\frac{1}{2}, \frac{1}{a}\right)$ and $a \sim IG\left(\frac{1}{2}, \frac{1}{A^2}\right)$, then $x \sim C^+(0, A)$. Finally, the distribution of parameters for the revised Horseshoe+ hierarchy is as follows:

$$\beta_k \sim N\left(0, \lambda_k^2\tau^2\right), \lambda_k^2 \sim IG\left(\frac{1}{2}, \frac{1}{\theta_k}\right), \theta_k \sim IG\left(\frac{1}{2}, \frac{1}{\eta_k^2}\right), \eta_k^2$$
$$\sim IG\left(\frac{1}{2}, \frac{1}{\nu_k}\right), \nu_k \sim IG(\frac{1}{2}, 1), \tau^2 \sim IG\left(\frac{1}{2}, \frac{1}{\xi}\right),$$
$$\textit{and}\ \ \xi \sim IG\left(\frac{1}{2}, N^2\right).$$

The conditional posterior distributions of λ_k and η_k parameters are inverse-gamma distributions, which makes Gibbs sampling straightforward.

BayesHE

In BayesU and BayesHP, the prior of local parameter λ_k followed a positive half-Cauchy distribution. Both of these two models used a fixed value as the degree of freedom. To increase the flexibility and the suitability of the prediction model, we proposed a new model, named BayesHE, which assumed the local parameter λ_k to follow a half-t distribution with an unknown degree of freedom υ:

$$\beta_k \sim N(0, \lambda_k^2\tau^2),\ \lambda_k \sim \textit{half} - t^+(\upsilon, 1),\ \tau \sim C^+(0, N^{-1}),$$
$$\textit{and}\ \upsilon \sim G\left(a, b\right)$$

By introducing auxiliary variables (Wand et al., 2011), the revised hierarchy is as follows:

$$\beta_k \sim N\left(0, \lambda_k^2\tau^2\right),\ \lambda_k^2 \sim IG\left(\frac{\upsilon}{2}, \frac{\upsilon}{\theta_k}\right),\ \theta_k \sim IG\left(\frac{1}{2}, 1\right),$$
$$\tau^2 \sim IG\left(\frac{1}{2},\ \frac{1}{\xi}\right),\ \xi \sim IG(\frac{1}{2}, N^2),\ \textit{and}\ \upsilon \sim G\left(a, b\right).$$

All parameters, including β_k, λ_k^2, θ_k, τ^2, and ξ, had a standard form, except υ. The full conditional distribution of the hyperparameter υ is described as follows:

$$f\left(\upsilon|.\right) \propto f\left(\lambda_k^2|\upsilon\right) * f(\upsilon) \propto \prod_{k\ =1}^{m} \frac{\left(\frac{\upsilon}{\theta_k}\right)^{\left(\frac{\upsilon}{2}\right)}}{\Gamma_{\left(\frac{\upsilon}{2}\right)}} \lambda_k^{2-}\left(\frac{\upsilon}{2}+1\right) \exp$$
$$\left(-\frac{\frac{\upsilon}{\theta_k}}{\lambda_k^2}\right) * \upsilon^{(a-1)} \exp\left(-b\upsilon\right) \propto \upsilon^{(\frac{\upsilon * m}{2}+a-1)} * \Gamma\left(\frac{\upsilon}{2}\right)^{-m}$$
$$* \exp\left(-\upsilon\left(\frac{1}{2}\sum_{k\ =1}^{m} \ln(\theta_k\lambda_k^2) + \sum_{k\ =1}^{m} \frac{1}{\theta_k\lambda_k^2} + b\right)\right)$$

where m is the number of SNPs, $\Gamma_{(.)}$ is the gamma function, a is the shape parameter in gamma distribution, and b is the scale parameter of the gamma distribution for υ. In this study, we compared two models with the same b (b equals to 1) but with different a, including BayesHE1 with a equals to 4 and BayesHE2 with a equals to 5. The hyperparameter was inferred by applying a univariate Metropolis–Hastings sampling (DFMH) process (Yang et al., 2015). The random walk M-H step worked with $\zeta = \log(\upsilon)$ because υ was inherently positive. The corresponding full conditional distribution of ζ is as follows:

$$f\left(\zeta|ELSE\right) \propto \exp\left(\zeta\right)^{\left(\frac{\exp(\zeta) * m}{2}+a-1\right)} * \Gamma\left(\frac{\exp\left(\zeta\right)}{2}\right)^{-m}$$
$$* \exp\left(-\exp\left(\zeta\right)\left[\frac{1}{2}\sum_{k\ =1}^{m} \ln(\theta_k\lambda_k^2) + \sum_{k\ =1}^{m} \frac{1}{\theta_k\lambda_k^2} + b\right]\right) * \exp(\zeta)$$
$$\propto \exp\left(\zeta\right)^{\frac{\exp(\zeta) * m+2a}{2}} * \Gamma\left(\frac{\exp\left(\zeta\right)}{2}\right)^{-m} *$$
$$\exp\left(-\exp\left(\zeta\right)\left[\frac{1}{2}\sum_{k\ =1}^{m} \ln(\theta_k\lambda_k^2) + \sum_{k\ =1}^{m} \frac{1}{\theta_k\lambda_k^2} + b\right]\right)$$

where $\exp\left(\zeta\right)$ is the Jacobian from υ to ζ.

The performance of three Bayesian regression models applying global–local priors, including BayesU, BayesHP, and BayesHE, were further compared with three widely used genomic prediction models, including GBLUP, BayesA, and BayesB.

BayesA/BayesB

In BayesB, the prior distribution of β_k is as follows:

$$\beta_k|S_\beta^2, \upsilon, \pi \sim IID \begin{cases} 0 & with\ probabilitty\pi \\ t(0, S_\beta^2, \upsilon\pi) & with\ probabilitty\ 1-\pi \end{cases}$$

The BayesA model can be considered as a specific case of BayesB, where $\pi = 0$. In this study, we set $\pi = 0.95$ for BayesB.

GBLUP

The GBLUP model is described as follows:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Zg} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of pre-corrected phenotypes, μ is the overall mean, $\boldsymbol{Z}$ is the design matrix linking genetic value ($\boldsymbol{g}$) to $\boldsymbol{y}$, and $\boldsymbol{e}$ is a vector of random residuals. It was assumed that,

$$\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)\ and\ \boldsymbol{e} \sim N(0, \boldsymbol{D}\sigma_e^2)$$

where σ_g^2 is the additive genetic variance and σ_e^2 is the random residual variance. The genomic relationship matrix ($\boldsymbol{G}$) (VanRaden, 2008) was calculated with SNPs:

$$\boldsymbol{G} = \frac{(\boldsymbol{M}-\boldsymbol{P})(\boldsymbol{M}-\boldsymbol{P})'}{2\sum_{k=1}^{m} p_k(1-p_k)},$$

where $\boldsymbol{M}$ is a $n \times m$ matrix with n for the number of individuals and m for the number of SNPs, p_k is the MAF of ith SNP, and $\boldsymbol{P}$ is the matrix in which the kth column elements are $2p_k$. In this study, GBLUP was implemented using the DMU software (Madsen and Jensen, 2012).

Datasets

Quantitative Trait Locus-Marker-Assisted Selection Data

We used the simulated data from the 15th QTL-marker-assisted selection (MAS) workshop (Elsen et al., 2012) to test model performances. The founder animals consist of 20 sires and 200 dams. For each generation, one sire was mated to 10 dams, and each dam produced 15 offspring. Eight QTLs were simulated across the five chromosomes, with one QTL being quadri-allelic, two linked in phase, two linked in repulsion, one imprinted, and two epistatic. Random residual effects were added to achieve a realized heritability of 0.3. After removing loci without polymorphisms, 7,121 SNPs were retained for analysis. Details on the simulated dataset are in Li et al. (2018).

In each full-sib family, 10 individuals had marker genotypes and phenotype, and the remaining five individuals only had marker genotypes. In total, 2,000 individuals had both genotype and phenotype information, and 1,000 individuals only had genotype information. In this study, only 2,000 individuals with genotypes and phenotypes were used for cross-validation.

Cattle Data

For real data analysis in dairy cattle, we collected phenotypic and genomic data from Chinese Holsteins. In total, 7,052 individuals were available for analyses on three milk production traits, including milk yield (MY), fat yield (FY), and protein yield (PY), and on one health traits (somatic cell score, SCS), and 3,530 individuals were available for three type traits including conformation (CONF), feet lag (FL), and mammary system (MS). DRP derived from the official EBV were used as pseudo-phenotypes for genomic prediction. The reliability of DRP for each individual was estimated as $r_{DRP}^2 = ERC_i/(ERC_i + \lambda)$, with $\lambda = \frac{1-h^2}{h^2}$, where ERC_i refers to the effective record contribution and h^2 refers to the estimated heritability of the trait. On note, effective record contribution (ERC_i) was relevant to the reliability (REL_i) of the EBV of animal i (Přibyl et al., 2013), $ERC_i = \lambda * REL/(1 - RELi$. Animals were genotyped by the Illumina 50K chip. Missing genotypes were imputed with Beagle version 3 (Browning and Browning, 2011). We further removed the SNPs with a minor allele frequency below 0.01 and significantly deviated from Hardy–Weinberg equilibrium ($p < 10^{-6}$) and the individuals with call rates lower than 0.90. After quality control, 43,447 SNPs remained for subsequent analyses.

Mice Data

For real data analysis in mice, we used the heterogeneous stock mice dataset generated by the Wellcome Trust Centre for Human Genetics[1]. As described by Legarra et al. (2008), the extent of linkage disequilibrium in this population is strong, with an average r_{LD}^2 among adjacent SNPs being 0.62. To compare the performance of different methods, we selected three traits: growth rate between 6 and 10 weeks of age (GSL), body mass index (BMI), and body length (BL). There were 1,821, 1,814, and 1,901 individuals available for analysis on BL, BMI, and GSL, respectively. In total, 9,098 SNPs were available. A detailed description of the population can be found in Li et al. (2018).

Cross-Validation and Prediction Accuracy

To assess the prediction accuracy, a 5×6 cross-validation (six-fold cross-validation repeated five times) procedure was used, and the results are shown as the mean and standard error for replicates. The performances of all methods were evaluated by examining the accuracy of direct genomic value (DGV) in test data. For QTL-MAS and mice data, Pearson correlation of DGV and phenotype/pre-corrected phenotype was used; for cattle data, the prediction accuracy was further corrected by the average accuracy (square root of reliability) of DRP in test data:

$$acc = \frac{cor(DRP, DGV)}{\bar{r}}$$

where $cor(DRP, DGV)$ is the Pearson correlation of DRP and DGV of the validation data, and $\bar{r}$ is the average of the square root of the testing data DRP reliabilities.

In addition, the regression of DRP on phenotype, y, was used to evaluate the unbiasedness of prediction for all three datasets. The closer the regression coefficient to one, the more unbiased the prediction result.

[1] http://gscan.well.ox.ac.uk/

RESULTS

In this study, the performance of our newly proposed Bayesian models with global–local priors was compared with GBLUP, BayesA, BayesB ($\pi = 0.95$), and BayesU, using the simulated data generated by QTL-MAS, and real data in cattle and mice.

To assess the convergence of Markov chain Monte Carlo, trace plots of the overall mean ($\boldsymbol{\mu}$) and additive variance [$V_g = \text{var}(\sum_{k=1}^{m} x_k \beta_k)$] are shown in **Figure 1**. Also, the trace plots suggested that parameters mixed well. However, additive variance from BayesHE (**Figures 1E,F**) converges faster than BayesHP (**Figure 1D**).

Quantitative Trait Locus-Marker-Assisted Selection Data

Table 1 shows the prediction accuracies and bias for all models based on the 15th QTL-MAS workshop dataset. Regarding the prediction accuracy, Bayesian regression models with global–local priors, such as BayesU (0.506), BayesHP (0.505), BayesHE1 (0.505), and BayesHE2 (0.505), outperformed all other methods. The prediction biases of the seven methods were similar and close to one.

Cattle Data

The prediction accuracies of seven traits in the Chinese Holstein population that the mean r^2_{LD} of adjacent SNP pairs ranged from 0.16 to 0.24 (Zhou et al., 2013) are shown in **Table 2**. Generally, BayesHE with two modalities (e.g., BayesHE1 and BayesHE2) on hyperparameters achieved optimal or suboptimal prediction accuracy for all of the seven traits.

For milk production traits, Bayesian regression models with global–local priors had a better performance compared with GBLUP, BayesA, or BayesB, especially for MY. For example, the prediction accuracy of BayesHE1 was 0.473, which increased approximately 2.2% than GBLUP. Also, BayesHE1 had similar prediction accuracy than BayesHE2. However,

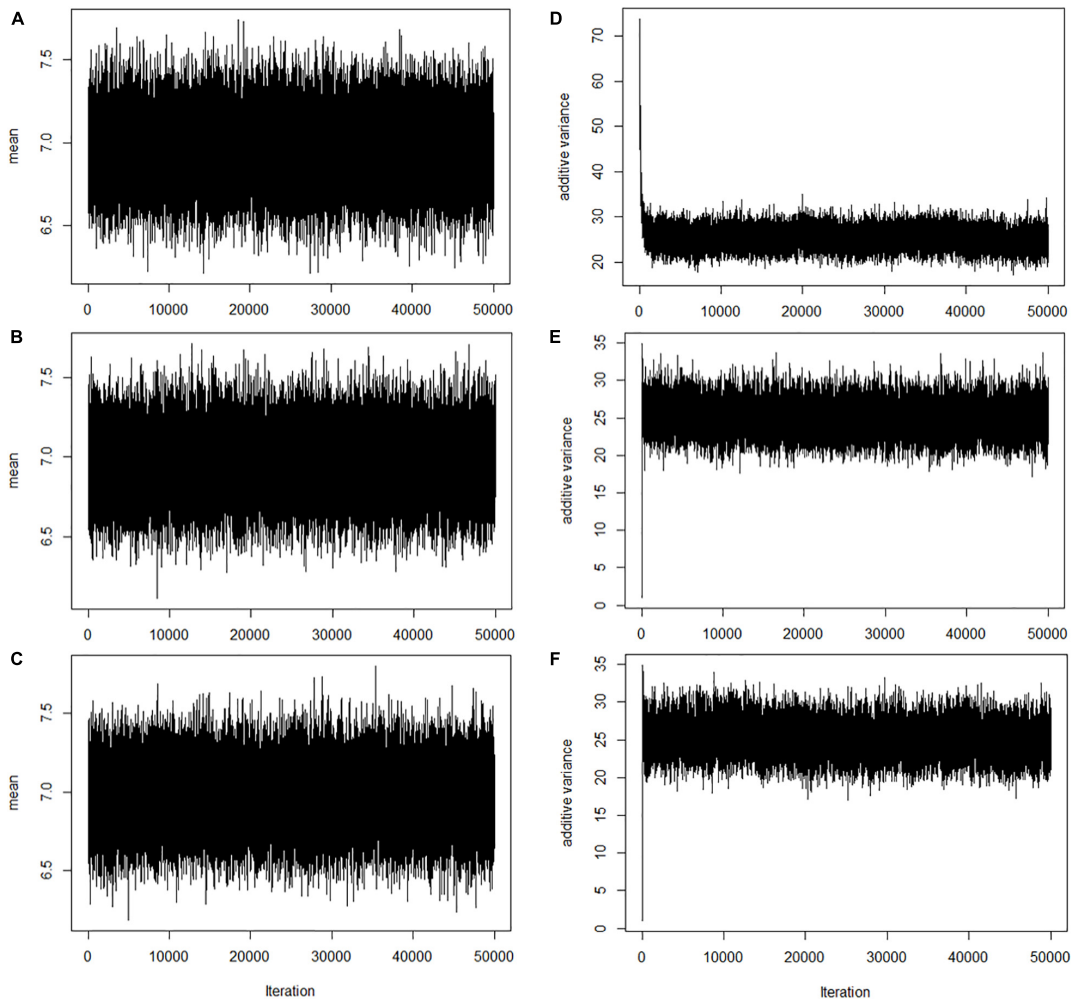

FIGURE 1 | Trace plots of overall mean and additive variance for BayesHP and BayesHE. **(A–C)** Trace plots of overall mean for BayesHP, BayesHE1, and BayesHE2; **(D–F)** Trace plots of additive variance for BayesHP, BayesHE1, and BayesHE2.

TABLE 1 | Prediction accuracies and biases of DGVs of test dataset from 15th QTL-MAS data using six-fold cross-validation with five replications.

	Accuracy	Bias
GBLUP	0.456±0.002	1.010±0.004
BayesA	0.474±0.009	0.924±0.008
BayesB	0.475±0.013	0.925± 0.004
BayesU	0.506±0.002	0.996±0.004
BayesHP	0.505±0.002	1.000±0.005
BayesHE1	0.505±0.002	1.005±0.006
BayesHE2	0.505±0.002	1.003±0.005

DGVs, direct genomic value; and the mean and standard errors are shown in the table.

for SCS, BayesHP achieved the lowest prediction accuracy (0.341), and BayesHE and BayesA had similar prediction accuracy (0.365).

For type traits, BayesB, BayesU, and BayesHP did not perform well, and GBLUP, BayesA, and BayesHE had similar prediction accuracy. Notably, BayesHE2 performed similarly to BayesHE1. Although BayesHE did not perform the best, the prediction accuracy was very close to that of the best model. For example, the prediction accuracy for MS from BayesHE2 was 0.427, which was only slightly lower than the highest prediction accuracy achieved by GBLUP (0.428).

The biases of prediction for the seven traits are shown in **Table 3**. For MY, FY, PY, and SCS, BayesHE1 achieved the least bias of prediction. The performance of BayesHE2, regarding bias, was very close to that of BayesHE1. For FL, BayesA was the most unbiased model, and BayesHE2 was the second best.

Mice Data

The prediction accuracies of three mice traits with different methods are shown in **Table 4**. In the analysis of mice data, there were two kinds of traits: one growth trait (GSL) and two type traits (BL and BMI). For all three traits, BayesHP performed the worst. For example, the prediction accuracy of BL from BayesHP was 0. 253, but accuracies from other methods were greater than 0.260. However, GBLUP, BayesA, and BayesHE had similar prediction accuracies.

Table 5 shows the prediction bias. The regression coefficients were close to the unity for all traits using all models, which indicated unbiasedness of the predictions. Nevertheless, there were still some slight differences. For example, the unbiasedness of BayesB was slightly lower than other models for all traits.

DISCUSSION

In Bayesian regression models, the differences among methods are the assumptions on the genetic marker effects. Because of the flexibility of the Bayesian method, it has attracted increasing attention. In classical one-group models, both signals and noises were assumed to follow one single continuous prior distribution, where the effects of some markers were shrunk toward zero, relying on the posterior distribution. For the two-group model or the spike-and-slab model, the

TABLE 2 | Prediction accuracies of seven traits of dairy cattle using six-fold cross-validation with five replications.

	MY	FY	PY	SCS	CONF	FL	MS
GBLUP	0.451±0.002	0.410±0.002	0.435±0.001	0.356±0.002	0.480±0.003	0.676±0.004	0.428±0.006
BayesA	0.467±0.002	0.425±0.002	0.433±0.001	0.365±0.002	0.478±0.003	0.677±0.004	0.425±0.006
BayesB	0.455±0.003	0.401±0.002	0.421±0.002	0.345±0.002	0.380±0.037	0.656±0.005	0.399±0.006
BayesU	0.463±0.003	0.415±0.002	0.420±0.002	0.346±0.003	0.447±0.007	0.664±0.007	0.404±0.008
BayesHP	0.459±0.003	0.410±0.002	0.414±0.003	0.341±0.003	0.440±0.009	0.660±0.007	0.401±0.008
BayesHE1	0.473±0.003	0.427± 0.002	0.435±0.001	0.365±0.002	0.478±0.003	0.674±0.004	0.426±0.006
BayesHE2	0.473±0.003	0.427±0.002	0.434±0.001	0.365±0.002	0.478± 0.003	0.674±0.004	0.427± 0.005

MY, milk yield; FY, fat yield; PY, protein yield; SCS, somatic cell score, CONF, conformation; FL, feet lag; MS, mammary system; and the mean and standard errors are shown in the table.

TABLE 3 | Prediction biases of seven traits of dairy cattle data using six-fold cross-validation with five replications.

	MY	FY	PY	SCS	CONF	FL	MS
GBLUP	0.865±0.004	0.817± 0.003	0.814±0.002	0.807±0.006	0.803±0.006	0.829±0.007	0.826±0.014
BayesA	0.877±0.005	0.807±0.003	0.807±0.002	0.812±0.005	0.798±0.006	0.837± 0.007	0.809±0.014
BayesB	0.824±0.006	0.755±0.003	0.750±0.004	0.756±0.006	0.763±0.021	0.775±0.006	0.740±0.014
BayesU	0.887±0.007	0.828±0.004	0.816±0.007	0.812± 0.008	0.749±0.019	0.812±0.011	0.786±0.020
BayesHP	0.889± 0.007	0.829±0.005	0.816±0.007	0.815±0.007	0.735± 0.022	0.805±0.011	0.772±0.019
BayesHE1	0.906±0.007	0.835±0.003	0.821±0.002	0.821±0.006	0.805±0.007	0.833±0.007	0.831±0.015
BayesHE2	0.905±0.008	0.834±0.003	0.819±0.002	0.820±0.006	0.807±0.007	0.834±0.007	0.833±0.014

MY, milk yield; FY, fat yield; PY, protein yield; SCS, somatic cell score, CONF, conformation; FL, feet lag; MS, mammary system; and the mean and standard errors are shown in the table.

TABLE 4 | Prediction accuracies of mice data using six-fold cross-validation with five replications.

	BL	BMI	GSL
GBLUP	0.272±0.002	0.226±0.002	0.386±0.003
BayesA	0.275±0.002	0.227±0.002	0.385±0.003
BayesB	0.268±0.001	0.217±0.002	0.374±0.003
BayesU	0.261±0.002	0.220±0.003	0.374±0.003
BayesHP	0.253±0.003	0.214±0.003	0.368±0.004
BayesHE1	0.274±0.002	0.229±0.002	0.386±0.003
BayesHE2	0.272±0.001	0.227±0.002	0.386±0.003

BL, body length; BMI, body mass index; GSL, growth slope; and the mean and standard errors are shown in the table.

TABLE 5 | Prediction biases of mice data using six-fold cross-validation with five replications.

	BL	BMI	GSL
GBLUP	0.988±0.012	1.023±0.022	1.004±0.010
BayesA	0.995±0.009	0.988±0.015	0.988±0.008
BayesB	0.948±0.004	0.929±0.017	0.959±0.007
BayesU	0.972±0.007	0.999±0.036	0.996±0.012
BayesHP	0.981±0.011	1.032± 0.039	1.002±0.015
BayesHE1	1.006±0.011	1.024±0.021	1.004±0.009
BayesHE2	1.003±0.009	1.016±0.021	1.005±0.008

BL, body length; BMI, body mass index; GSL, growth slope; and the mean and standard errors are shown in the table.

prior regarding the proportion of genetic markers being signal usually impact its performance in genomic prediction. Some two-group models, such as BayesCπ (Habier et al., 2011), have been developed to estimate the proportion of non-zero effect markers based on both prior and the analyzed data. However, there is a poor convergence and mixing in some situations. The global–local prior, which can shrink signals and noises through local and global parameters, seems to be a good alternative, theoretically. Global–local priors is a kind of continuous shrinkage prior, which can adaptively shrink noise to zero while leaving the large data-supported signal unshrunk (Ge et al., 2019).

The model's performance depended on the genetic architecture of the trait. The results of simulation indicated that models based on global–local priors, e.g., BayesU, BayesHP, and BayesHE, performed better in traits with higher heritability (i.e., in this study, heritability is 0.3) and fewer QTL. In real data, BayesHE can achieve optimal or suboptimal performance; however, BayesHP performed better only for production traits. Our results suggested that auto-estimate the degree of freedom (e.g., BayesHE) would be a better choice other than increasing the layers of the local parameter (e.g., BayesHP).

The Bayesian models with the assumption more in line with the real distribution of marker effects will result in more accurate predictions. The Bayesian model shrinks the effect of noise markers toward zero and thus increases the prediction accuracy. However, in genomic prediction, markers are not simply signal or noise due to the existence of linkage disequilibrium. It is reasonable that for some traits, GBLUP will achieve better prediction accuracy. For example, GBLUP performed better for type traits (e.g., CONF, FL, and MS) than BayesB, BayesU, and BayesHP, as shown in our results. Notably, in genome-wide association study, regardless of dairy cattle (Wu et al., 2013) or beef cattle (Vallée et al., 2016), there are few significant signals for type traits, suggesting that most genetic variants have similar medium or small effects on the traits. Therefore, it is reasonable why GBLUP had a better performance for type traits.

Many previous studies have suggested that using hyperparameters is likely to improve classical methods (Habier et al., 2011; Yang and Tempelman, 2012; Zhu et al., 2016). In our study, we assumed that the local parameter, λ_k, followed a half-t distribution (BayesHE) with an unknown degree of freedom instead of half-Cauchy (BayesU). By introducing auxiliary variables (Wand et al., 2011), half-t distribution was translated into a scale mixture of the inverse gamma distribution. In BayesHE, λ_k^2 was assumed to follow an inverse gamma distribution $IG\left(\frac{\upsilon}{2}, \frac{\upsilon}{\theta_k}\right)$, which led to an assumption of student-t distribution for marker effects (Wand et al., 2011). The use of unknown shape parameter is similar to the study of Zhu et al. (2016), but the difference is that there is a global parameter τ^2 in BayesHE model. Besides, in studies of Habier et al. (2011) and Zhu et al. (2016), they set a gamma distribution $G(1, 1)$ for the scale parameter. In our study, the scale parameter θ_k was assumed to follow an inverse gamma distribution, $\theta_k \sim IG(\frac{1}{2}, 1)$. This inspired the authors that the shape parameter of inverse gamma distribution that θ_k followed can also be set as a variable other than a constant.

Horseshoe-like prior with "U" type shrinkage pattern means strong distinguishment of single and noise. According to the results of QTL detection (Wu et al., 2013; Vallée et al., 2016), Horseshoe-like prior with "U" type may be suitable for genomic prediction of the traits affected by many QTLs with large effect. In our study, we assumed an unknown hyperparameter for the distribution of local parameters, which increased the model flexibility and, therefore, more adaptable to traits with different genetic architectures. The possibility to fit a suitable hyperparameter for the global parameter has been proposed by Armagan et al. (2011), where they assumed global parameter followed a gamma distribution with different shape parameter or just set as a constant value. Their study suggested that the changes of hyperparameters of distributions that local parameters followed and the value of global parameter led to different shrinkage patterns on covariates. In the study of Piironen and Vehtari (2016), the global parameter τ was set as a constant value or followed a normal or half-Cauchy distribution, and they recommended τ half-Cauchy distribution, $\tau \sim C^+(0, \tau_0^2)$, where the scale parameter τ_0^2 is relevant to the effective number of variables with non-zero effects. In the global–local prior method, the marker variances were shaped by global and local parameters simultaneously. The global parameter, τ, usually causes the marker effect to approach zero, whereas the local parameter, λ_k, allows marker variance to escape the shrinkage

when that marker has a large effect. In future research, more investigation on choosing the type of distribution for global parameters could be interesting.

The limitation of our study is that it mostly focused on statistical perspectives and lack of consideration of the biological information. With time, an increasing amount of biological information affecting complex traits will be detected. It is reasonable to integrate these genomic features into the prediction model, and then, how to effectively utilize these genomic features is worth exploring. The BayesRC model proposed by MacLeod et al. (2016) divides the genome into three major categories: trait-associated genes, regular regions, and other variations. However, there are some challenges in utilizing biological information because of the dynamics in biological processes.

CONCLUSION

Our results showed that BayesHE could achieve optimal or suboptimal performance. Compared with other methods, such as GBLUP and BayesA, BayesHP did not perform better. With the automatic estimation of hyperparameters, BayesHE was more flexible than BayesU and BayesHP for the adaptation to a wider range of traits. This suggested that auto-estimate the degree of freedom (e.g., BayesHE) would be a better choice other than increasing the layers of a local parameter (e.g., BayesHP).

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: All information supporting the results is included in the text and tables. The dataset of cattle is not publicly available due to commercial restrictions. Requests to access these datasets should be directed to the corresponding author (SZ).

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional of Animal Care and Use Committee (IACUC), China Agricultural University. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

SS conceived the study, wrote the program, analyzed the data, and drafted the manuscript. SZ and XD conceived the study and supervised the project. XL and LF participated in the design and helped to draft the manuscript. AL helped to analyze the data and revised the manuscript. GS revised the manuscript. YZ and BL helped to draft the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the China Agriculture Research System (CARS-36 and CARS-35), Ningxia Province High Quality and High Yield Breeding Project [(2019)21-1], Tianjin Science and Technology Project (19ZXZYSN00130), the National Key Research and Development Project (2019YFE0106800), and the Pearl River S&T Nova Program of Guangzhou (201906010040).

ACKNOWLEDGMENTS

We thank all contributors to the present study. We also thank Dr. Emre Karaman and Dr. Ganjun Tu for their assistance in learning professional knowledge.

REFERENCES

Armagan, A., Dunson, D. B., and Clyde, M. (2011). Generalized beta mixtures of gaussians. *Adv. Neural. Inf. Process. Syst.* 24, 523–531.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* 12, 1105–1131. doi: 10.1214/16-ba1028

Browning, B. L., and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 88, 173–182. doi: 10.1016/j.ajhg.2011.01.010

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–480. doi: 10.1093/biomet/asq017

Elsen, J.-M., Tesseydre, S., Filangi, O., Le Roy, P., and Demeure, O. (2012). XVth QTLMAS: simulated dataset. *BMC Proc.* 6:S1. doi: 10.1186/1753-6561-6-S2-S1

Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* 10:1776. doi: 10.1038/s41467-019-09718-5

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363. doi: 10.1534/genetics.109.103952

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

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

Jairath, L., Dekkers, J. C. M., Schaeffer, L. R., Liu, Z., Burnside, E. B., and Kolstad, B. (1998). Genetic evaluation for herd life in Canada. *J. Dairy Sci.* 81, 550–562. doi: 10.3168/jds.S0022-0302(98)75607-3

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2017). Bayes shrinkage at GWAS scale: convergence and approximation theory of a scalable MCMC algorithm for the horseshoe prior. *arXiv* [Preprint]. arXiv:170500841.

Legarra, A., Robert-Granie, C., Manfredi, E., and Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics* 180, 611–618. doi: 10.1534/genetics.108.088575

Li, X., Liu, X., and Chen, Y. (2018). The influence of a first-order antedependence model and hyperparameters in BayesCπ for genomic prediction. *Asian-Australas. J. Anim. Sci.* 31, 1863–1870. doi: 10.5713/ajas.18.0102

Li, Y., Datta, J., Craig, B. A., and Bhadra, A. (2019). Joint mean-covariance estimation via the horseshoe with an application in genomic data analysis. *arXiv* [Preprint]. arXiv:190306768.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6

Madsen, P., and Jensen, J. (2012). *A User's Guide to DMU. A Package for Analysing Multivariate Mixed Models. Version 6, Release 5.1. Tjele, Denmark; 2012*. Available online at: http://dmu.agrsci.dk/DMU/Doc/Previous/dmuv6_guide.5.1.pdf (accessed December 1, 2012).

Makalic, E., and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* 23, 179–182. doi: 10.1109/LSP.2015.2503725

Makalic, E., Schmidt, D. F., and Hopper, J. L. (2016). "Bayesian robust regression with the horseshoe+ estimator," in *Proceedings of the 29th Australasian Joint Conference Hobart, TAS, Australia, December 5–8, 2016: AI 2016: Advances in Artificial Intelligence*, eds B. H. Kang and Q. Bai (Cham: Springer International Publishing), 429–440.

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

Piironen, J., and Vehtari, A. (2016). *On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior*. Available online at: https://ui.adsabs.harvard.edu/abs/2016arXiv161005559P (accessed October 01, 2016).

Piironen, J., and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electr. J. Statist.* 11, 5018–5051. doi: 10.1214/17-ejs1337si

Pong-Wong, R., and Woolliams, J. (2014). "Bayes U: a genomic prediction method based on the horseshoe prior," in *Proceedingsof the 10th World Congress of Genetics Applied to Livestock Production*, Vancouver, BC, 679.

Přibyl, J., Madsen, P., Bauer, J., Přibylová, J., Šimečková, M., Vostrý, L., et al. (2013). Contribution of domestic production records, Interbull estimated breeding values, and single nucleotide polymorphism genetic markers to the single-step genomic evaluation of milk production. *J. Dairy Sci.* 96, 1865–1873. doi: 10.3168/jds.2012-6157

Vallée, A., Daures, J., van Arendonk, J. A. M., and Bovenhuis, H. (2016). Genome-wide association study for behavior, type traits, and muscular development in Charolais beef cattle1. *J. Anim. Sci.* 94, 2307–2316. doi: 10.2527/jas.2016-0319

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

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Anal.* 6, 847–900. doi: 10.1214/11-ba631

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., et al. (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genet. Sel. Evol.* 43:5. doi: 10.1186/1297-9686-43-5

Wu, X., Fang, M., Liu, L., Wang, S., Liu, J., Ding, X., et al. (2013). Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. *BMC Genomics* 14:897. doi: 10.1186/1471-2164-14-897

Yang, W., Chen, C., and Tempelman, R. J. (2015). Improving the computational efficiency of fully Bayes inference and assessing the effect of misspecification of hyperparameters in whole-genome prediction models. *Genet. Sel. Evol.* 47:13. doi: 10.1186/s12711-015-0092-x

Yang, W., and Tempelman, R. J. (2012). A Bayesian antedependence model for whole genome prediction. *Genetics* 190, 1491–1501. doi: 10.1534/genetics.111.131540

Zhou, L., Ding, X., Zhang, Q., Wang, Y., Lund, M. S., and Su, G. (2013). Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. *Genet. Sel. Evol.* 45:7. doi: 10.1186/1297-9686-45-7

Zhu, B., Zhu, M., Jiang, J., Niu, H., Wang, Y., Wu, Y., et al. (2016). The impact of variable degrees of freedom and scale parameters in Bayesian methods for genomic prediction in Chinese Simmental beef cattle. *PLoS One* 11:e0154118. doi: 10.1371/journal.pone.0154118

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.

APPENDIX

Appendix A: Gibbs Sampler for SNP Effect β_k

The full conditional distribution of β_k for BayesHP and BayesHE can be written as,

$$f(\beta_k|ELSE) \propto f\left(y|\mu,\beta,\sigma_e^2\right) * f\left(\beta_k|\tau^2,\lambda_k^2\right) \propto \exp\left[-\frac{\left(\boldsymbol{y}-\boldsymbol{\mu}-\sum_{k=1}^{m}\boldsymbol{x_k}\beta_k\right)^{'}\boldsymbol{R}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}-\sum_{k=1}^{m}\boldsymbol{x_k}\beta_k\right)}{2\sigma_e^2}\right] * \exp\left[-\frac{\beta_k^2}{2\lambda_k^2\tau^2}\right]$$

$$\propto \exp\left[-\frac{1}{2}\frac{\left(\beta_k-\hat{\beta}_k\right)^2}{\sigma_e^2/c_k}\right] \propto \mathrm{N}(\hat{\beta}_k, \sigma_e^2/c_k)$$

where $\hat{\beta}_k = \frac{x_k^{'}R^{-1}w}{c_k}$, $w = \boldsymbol{y}-\boldsymbol{\mu}-\sum_{j\neq k}^{m}\boldsymbol{x}_j\beta_j$ and $c_k = (x_k^{'}R^{-1}x_k + \frac{\sigma_e^2}{\lambda_k^2\tau^2})$.

Appendix B: Gibbs Sampler for λ_k^2

The full conditional distribution of λ_k^2 can be written as Makalic et al. (2016),

$$f\left(\lambda_k^2|ELSE\right) \propto f\left(\beta_k|\tau^2,\lambda_k^2\right) * f\left(\lambda_k^2|\theta_k,\nu\right) \propto \left(2\pi\lambda_k^2\tau^2\right)^{-\frac{1}{2}}\exp\left[-\frac{\beta_k^2}{2\lambda_k^2\tau^2}\right] * \lambda_k^{2-\left(\frac{\nu}{2}+1\right)}\exp\left(-\frac{\frac{\nu}{\theta_k}}{\lambda_k^2}\right) \propto \lambda_k^{2-\left(\frac{\nu}{2}+\frac{1}{2}+1\right)}$$

$$\exp\left[-\frac{1}{\lambda_k^2}\left(\frac{\beta_k^2}{2\tau^2}+\frac{\nu}{\theta_k}\right)\right] \propto IG\left(\frac{\nu}{2}+\frac{1}{2}, \frac{\beta_k^2}{2\tau^2}+\frac{\nu}{\theta_k}\right)$$

Notably, in BayesHP, ν equals to one.

Appendix C: Gibbs Sampler for θ_k

The full conditional distribution of θ_k can be written as Makalic et al. (2016),

$$f(\theta_k|ELSE) \propto f\left(\lambda_k^2|\theta_k,\nu\right) * f(\theta_k) \propto \left(\frac{p}{\theta_k}\right)^{\left(\frac{p}{2}\right)} * \exp\left(-\frac{\frac{\nu}{\theta_k}}{\lambda_k^2}\right) * \theta_k^{-\left(\frac{1}{2}+1\right)}$$

$$\exp\left(-\frac{1}{\theta_k}\right) \propto \theta_k^{-\left(\frac{\nu}{2}+\frac{1}{2}+1\right)}\exp\left(-\frac{1}{\theta_k}\left(1+\frac{\nu}{\lambda_k^2}\right)\right) \propto IG\left(\frac{\nu}{2}+\frac{1}{2}, 1+\frac{\nu}{\lambda_k^2}\right)$$

Similarly, in BayesHP, ν equals to one.

Appendix D: Gibbs Sampler for η_k^2

The full conditional distribution of η_k^2 can be written as Makalic et al. (2016),

$$f\left(\eta_k^2|ELSE\right) \propto f\left(\theta_k|\eta_k^2\right) * f\left(\eta_k^2|\nu_k\right) \propto \left(\eta_k^2\right)^{-\frac{1}{2}}\exp\left(-\frac{(\eta_k^2)^{-1}}{\theta_k}\right) * \left(\eta_k^2\right)^{-\left(\frac{1}{2}+1\right)}\exp\left(-\frac{\nu_k^{-1}}{\eta_k^2}\right) \propto \eta_k^{2-(1+1)}$$

$$* \exp\left(-\frac{1}{\eta_k^2}\left(\frac{1}{\nu_k}+\frac{1}{\theta_k}\right)\right) \propto IG\left(1, \frac{1}{\nu_k}+\frac{1}{\theta_k}\right)$$

Appendix E: Gibbs Sampler for ν_k

The full conditional distribution of ν_k can be written as Makalic et al. (2016),

$$f(\nu_k|ELSE) \propto f\left(\eta_k^2|\nu_k\right) * f(\nu_k) \propto \nu_k^{-\frac{1}{2}}\exp\left(-\frac{\nu_k^{-1}}{\eta_k^2}\right) * \nu_k^{-\left(\frac{1}{2}+1\right)}\exp\left(-\frac{1}{\nu_k}\right) \propto \nu_k^{-(1+1)}\exp\left(-\frac{1}{\nu_k}\left(1+\frac{1}{\eta_k^2}\right)\right) \propto IG\left(1, 1+\frac{1}{\left(\eta_k^2\right)}\right) \quad (1)$$

Appendix F: Gibbs Sampler for τ^2

The full conditional distribution of τ^2 can be written as Makalic et al. (2016),

$$f\left(\tau^2|ELSE\right) \propto f\left(\tau^2|\xi\right) * \prod_{k=1}^{m} f\left(\beta_k|\tau^2, \lambda_k^2\right) \propto \left(\tau^2\right)^{-\left(\frac{1}{2}+1\right)} \exp\left(-\frac{\xi^{-1}}{\tau^2}\right) * \prod_{k=1}^{m} \left(2\pi\lambda_k^2\tau^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\beta_k^2}{2\lambda_k^2\tau^2}\right) \propto \tau^{2-\left(\frac{k+1}{2}+1\right)}$$

$$\exp\left[-\frac{1}{\tau^2}\left(\frac{1}{\xi} + \sum \frac{\beta_k^2}{2\lambda_k^2}\right)\right] \propto IG\left(\frac{k+1}{2}, \frac{1}{\xi} + \sum \frac{\beta_k^2}{2\lambda_k^2}\right)$$

Appendix G: Gibbs Sampler for ξ

The full conditional distribution of ξ can be written as Makalic et al. (2016),

$$f\left(\xi|ELSE\right) \propto f\left(\tau^2|\xi\right) * f(\xi) \propto \xi^{-\frac{1}{2}} \exp\left(-\frac{\xi^{-1}}{\tau^2}\right) * \xi^{-\left(\frac{1}{2}+1\right)} \exp\left(-\frac{N^2}{\xi}\right) \propto \xi^{-(1+1)} * \exp\left[-\left(\frac{1}{\tau^2} + N^2\right)\frac{1}{\xi}\right] \propto IG\left(1, \frac{1}{\tau^2} + N^2\right)$$

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fgene.2021.642552

The GWAS Analysis of Body Size and Population Verification of Related SNPs in Hu Sheep

Junfang Jiang[1], Yuhao Cao[2], Huili Shan[1], Jianliang Wu[1], Xuemei Song[2]* and Yongqing Jiang[1]**

[1] Institute of Animal Husbandry and Veterinary, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [2] Zhejiang Key Laboratory of Pathophysiology, Department of Biochemistry and Molecular Biology, Medical School of Ningbo University, Ningbo, China

Body size is an important indicator of growth and health in sheep. In the present study, we performed Genome-Wide Association Studies (GWAS) to detect significant single-nucleotide polymorphisms (SNPs) associated with Hu sheep's body size. After genotyping parental (G1) and offspring (G2) generation of the nucleus herd for meat production of Hu sheep and conducting GWAS on the body height, chest circumference, body length, tail length, and tail width of the two groups, 5 SNPs associated with body height and 4 SNPs correlated with chest circumference were identified at the chromosomal significance level. No SNPs were significantly correlated to body length, tail length, and width. Four out of the 9 SNPs were found to be located within the 4 genes. *KITLG* and *CADM2* are considered as candidate functional genes related to body height; *MCTP1* and *COL4A6* are candidate functional genes related to chest circumference. The 9 SNPs found in GWAS were verified using the G3 generation of the nucleus herd for meat production. Nine products were amplified around the 9 sites, and 29 SNPs were found; 3 mutation sites, G > C mutation at 134 bp downstream of s554331, T > G mutation at 19 bp upstream of s26859.1, and A > G mutation at 81 bp downstream of s26859.1, were significantly correlated to the body height. Dual-luciferase reporter gene experiments showed that the 3 SNPs could significantly impact dual-luciferase and gene transcription activity.

Keywords: Hu sheep, body size traits, genome wide association studies, SNPs, transcription activity, population verification

OPEN ACCESS

Edited by:
Zhe Zhang,
South China Agricultural University, China

Reviewed by:
Ricardo Zanella,
The University of Passo Fundo, Brazil
Ran Di,
Institute of Animal Sciences (CAAS), China

****Correspondence:***
Junfang Jiang
jiangjunfang1031@sina.cn
Xuemei Song
songxuemei@nbu.edu.cn
Yongqing Jiang
jyq61@sohu.com

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

Received: *16 December 2020*
Accepted: *20 April 2021*
Published: *20 May 2021*

Citation:
Jiang J, Cao Y, Shan H, Wu J, Song X and Jiang Y (2021) The GWAS Analysis of Body Size and Population Verification of Related SNPs in Hu Sheep. Front. Genet. 12:642552. doi: 10.3389/fgene.2021.642552

INTRODUCTION

In sheep, body size has been widely recognized as an important indicator of growth and health (Kemper et al., 2012), which impacts animal feeding and management as well as adaptation to the environment. Mature body size has been extensively studied in humans, cattle, and other domestic animals but not in sheep (Posbergh and Huson, 2021). In sheep, the mature body size is more polygenic than in other domesticated animals, which suggests that the development of genomic trait selection might be the optimal option for evaluating body size in sheep (Posbergh and Huson, 2021).

GWAS (Genome-wide association study) is a method that uses millions of single nucleotide polymorphism (SNP) in genomes as molecular genetic markers to conduct control analysis or correlation analysis at the whole genome level so as to investigate the genetic mutation of complex traits. This technique has been applied to screen the SNPs of agricultural animals' major traits.

Eight common gene candidates, i.e., *GRID1*, *ALOX12*, *SLC16A13*, *SLC16A11*, *TP53*, *STX8*, *NTN1*, and *ZNF521*, were identified from GWAS for body size traits in crossbreeding sheep between Frizarta sheep and East Friesian sheep (Kominakis et al., 2017). In Hulun Buir sheep, 13 candidate genes, including *SMURF2*, *FBF1*, *DTNBP1*, *SETD7,* and *RBM11,* have been associated with fat metabolism, and *SMARCA5* and *GAB1* were associated with body size (Zhang et al., 2019). In addition, *MARCA5* and *GAB1* have been found to be related to sheep's body size. Height has been associated with 12 SNPs across six chromosomes. Ear length was associated with a single locus on chromosome 3.

Hu sheep, which are mainly housed all year round, are a special type of sheep that is only found in China. Hu sheep are characterized by early sexual maturity, high fecundity, and rapid growth. It is famous for its beautiful lamb skin. Hu sheep also have good meat quality, strong resistance to stress, and resistance to rough feeding (Yue, 1996). Until now, no GWAS study on the body size traits of Hu sheep has been reported.

In this study, GWAS were applied to screen and select candidate SNPs for traits and body size of meat-type Hu sheep. Moreover, the candidate SNPs associated with Hu sheep's body size were verified among meat-type Hu sheep' offspring of G3 generation.

MATERIALS AND METHODS

Animals

The GWAS study included 240 Hu sheep from G1 and G2 generation of meat-type Hu sheep nucleus herd in Huzhou Taihu Lake Culture Cooperative by semi-open nucleus breeding. The SNP herd verification included 202 Hu sheep from the G3 generation of Hu sheep nucleus herd in Hangzhou Pangda Agricultural Development Co., Ltd. The breeding and management of the Hu sheep included in the study were conducted according to the standard methods of breeding and management of Hu sheep.

Determination of Body Size Traits and Genomic DNA Extraction

Selected body size traits included body height, chest circumference, body length, tail length, and tail width. The body height and body length were measured using a measuring stick (Zhengzhou Zhimuren Machinery Equipment Co., Ltd.) with an accuracy of 1 mm; chest circumference, tail length, and tail width were measured using a tape measure with an accuracy of 1 mm (Zhengzhou Zhimuren Machinery Equipment Co., Ltd.). Sheep were on horizontal ground, quiet and relaxed. The measurement methods were: (1) body length: the straight-line length from the front edge of the shoulder and foot bones to the back edge of the ischial tuberosity; (2) body height: the vertical length from the highest point of the bun to the ground; (3) chest circumference: the length of the circumference around the back edge of scapula; (4) tail length: the length from root to the top of the tail; (5) tail width: the widest range length of the tail. All measurements were performed by the same worker to minimize measurement errors caused by artificial reasons. Each sheep was measured at least 2 times, and the average was taken as the final measurement result. In addition, a total of 10 ml of blood was collected from each sheep's jugular vein and then placed into EDTA anticoagulant tubes. DNA was extracted with phenol/chloroform extraction method and kept at −20°C.

Genotyping and Quality Control

Ovine SNP50 BeadChip was applied to genotype individual SNPs. The chip was co-developed by Illumina and experts from International Sheep Genomics Consortium. Plink 1.09 software was applied to conduct quality control on genotypes, phenotypic data and samples, analyze SNPs, and estimate genotypes and phenotypic value.

Population Structure Analysis and Genome-Wide Association Study

Population structure analysis was performed using admixture v1.3. A heat map of the values in the kinship matrix was created for the kinship plot. After quality control was performed on genotype data, GWAS on SNP was performed using the mixed linear model (MLM) of TASSEL5.0 software to identify SNPs related to the body size traits of the nucleus herd for meat production of Hu sheep. MLM model was adjusted according to 3 confounding factors, i.e., sex, herd structure, and genetic relationship. The following concrete model was used:

$$Y = X\beta + S\alpha + Qv + Zu + e$$

where Y is the phenotypic value of Hu sheep's body size traits; β stands for fixed effects apart from SNP and herd structure; α stands for SNP effect; v represents herd structure effect; u stands for polygenic background effect; e represents residual effect, and X, S, Q, Z represent the incidence matrix of β, α, v, u, respectively.

When performing correlation analysis on the body size traits of Hu sheep, if errors were found in multiple hypothesis tests, a *p*-value was analyzed and adjusted. MLM was used to calculate F and *p* values, after which the results were verified. The following formula was applied:

$$Ps = \alpha/N,$$

where α stands for the level of significance and *N* for the number of independent SNPs used in the analysis. If the *p*-value at the SNP site was less than α, this SNP site was considered as significantly correlated to body size traits.

After SNPs sites were obtained with the performance of GWAS, base sequence 500 bp upstream and downstream of the significantly correlated SNP sites were downloaded. Next, BLAST research for the sequence was then performed with NCBI

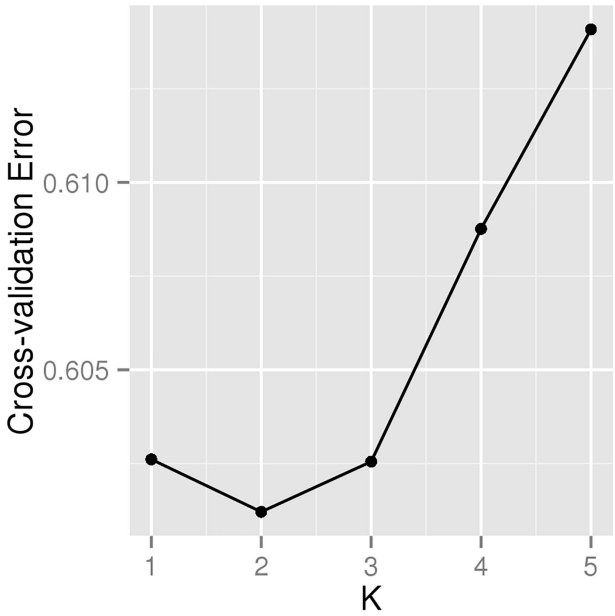

FIGURE 1 | Cross-validation plot for determining the best K.

and Ovis aries_v4.0 (UCSC) to confirm the information of the location of SNP and adjacent genes.

Group Verifying of Significantly Correlated SNPs

Two hundred two ewes (from Hangzhou Pangda Agricultural Development Co., Ltd.) were included as subjects. PCR Amplification, product sequencing, and gene sequence analysis were used to perform SNPs detection. Base sequence 500 bp upstream and downstream of the SNP sites were significantly correlated to Hu sheep's body height, and chest circumference was downloaded. Primers are shown in **Supplementary Table 1**. A total of 25 μL PCR reaction system was used for PCR amplification; the reaction procedure included: initial denaturation for 2 min at 94°C; denaturation for 30 s at 95°C, annealing for 30 s at 55°C, an extension for 30 s at 72°C, 35 circulations; extension for 10 min at 72°C, preservation at 4°C after the completion of the reaction. Direct sequencing was performed on the upstream and downstream primers for PCR products for each SNP site of each sample. Mutation Surveyor 5.02 was used to analyze the forward and reverse sequencing diagram of each ewe so as to confirm the mutation sites and mutation methods of the sequencing results of the amplified products at different sites in each sample. PopGen32 was used to calculate the gene frequency and genotype frequency of the SNPs. Hardy-Weinberg equilibrium test was performed to calculate Polymorphism information content (PIC).

The relationship between the different genotypes or haplotypes and body size traits of meat-type Hu sheep was evaluated by fitting a general linear model using the restricted maximum likelihood method in the Statistical Package for the Social Sciences (SPSS; version 20.0; SPSS Inc., Chicago, IL, United States). The general model used for Hu sheep body size traits was:

$$Yij = \mu i + Mj + eij$$

where Yij is the meat type Hu sheep body height or Chest circumference; μi is the least square mean; Mj is the fixed effect of the jth genotype or haplotypes, and eij is the random residual effect of each Hu sheep body height or chest circumference value.

Linkage Disequilibrium Analysis

HaploView version 4.2 was used to perform Linkage disequilibrium (LD) block and Haplotype analyses (Whitehead Institute for Biomedical Research, Cambridge, MA, United States). The D'-value of the lower 95% confidence interval in the analysis was used to define the haplotype block (Brym et al., 2005).

Effects of Candidate Functional SNPs or Haplotypes on Gene Transcriptional Activity

Candidate functional SNP loci significantly associated with body height in the wild-type and homozygous mutant sheep were selected. The different haplotypes amplification product was cloned into the pGL4.10 vector (Promega, United States), expressing a dual-luciferase gene (General Biosystems Corporation, Anhui, China). The vector was then transfected into sheep kidney cells. After 24 h, the luciferase activity was measured on a microplate reader using the Dual-Luciferase® reporter assay system (Promega, United States).

The relationship between the different genotypes or haplotypes and luciferase activities was evaluated by fitting a general linear model using the restricted maximum likelihood method in the Statistical Package for the Social Sciences (SPSS; version 20.0; SPSS Inc., Chicago, IL, United States). The general model used for luciferase activity value was:

$$Yij = \mu i + Mj + eij$$

where Yij is the luciferase activities; μi is the least square mean; Mj is the fixed effect of the jth genotype or haplotypes, and eij is the random residual effect of luciferase activities.

RESULTS

Descriptive Statistics and Quality Control

The descriptive statistical information of the phenotypic values related to the individual body size traits of 240 G1 and G2 generation Hu sheep has is shown in **Supplementary Table 2**. The descriptive statistical information of the phenotypic values related to the individual body size traits of 202 G3 generation Hu sheep is shown in **Supplementary Table 3**.

In the previous work, we focused on Genome-Wide Association Study of body weights in Hu Sheep. Here, we examined the relationship between the same population and body sizes traits. The result of data quality control can be found in the study by Cao et al. (2020).

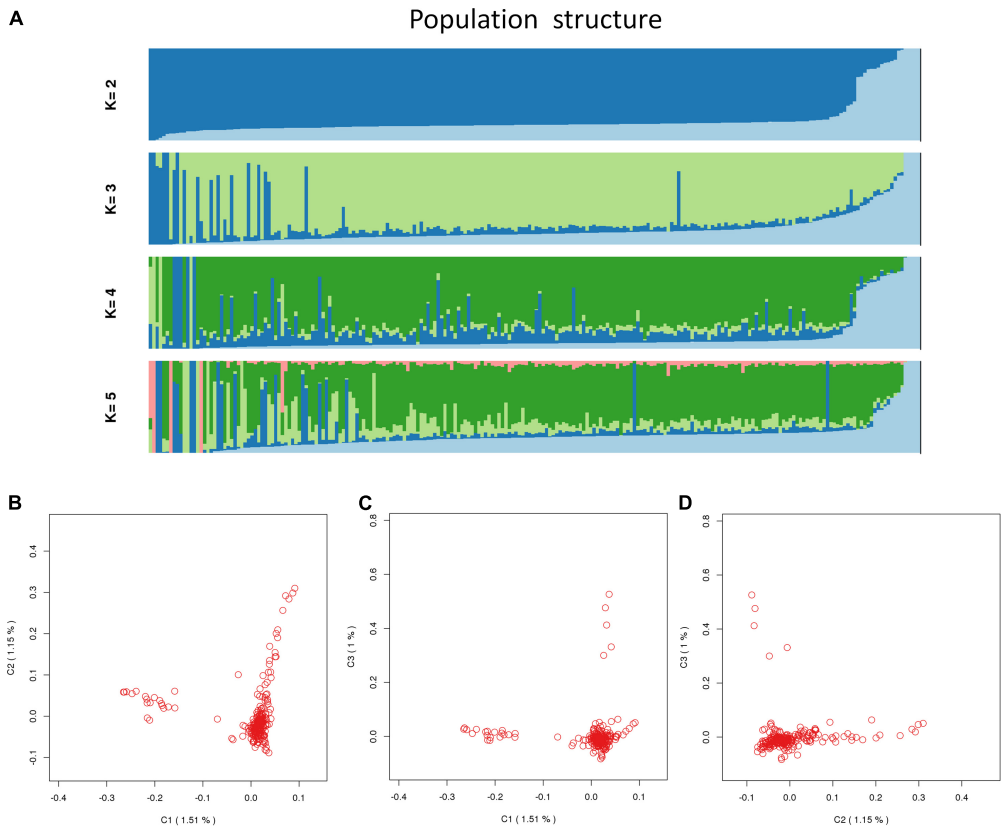

FIGURE 2 | Population_structure_by_Admixture. **(A)** Population structure with K from 2 to 5; **(B)** principal components (PC) plot drawn from the second principal component (C2) against the first principal component (C1); **(C)** principal components (PC) plot drawn from the second principal component (C3) against the first principal component (C1); **(D)** principal components (PC) plot drawn from the second principal component (C3) against the first principal component (C2).

Population Structures and Association Analyses

A total of 226 Hu sheep were randomly selected from the group of Huzhou Taihu Lake Culture Cooperative. According to the population structure, the result was given by admixture v1.3 with K from 1 to 5, where the optimal K was 2 (**Figure 1**). Kinship estimation and Principle component analysis (PCA) of all individuals indicated the effectiveness of sampling (**Figures 2A–D**).

Based on the number of independently effective SNPs, the p-value corresponding to the 1% significance level was 2.83×10^{-7}, and that corresponding to the 5% significance level was 1.41×10^{-6}. SNPs with a p-value lower than this threshold value were considered to be significantly correlated to phenotype. GWAS' results showed that 5 SNPs were significantly correlated to body height (**Figure 3**) in terms of genomic level; OAR23_3237800.1 ($p = 5.53 \times 10^{-8}$) of chromosome 23; OAR6_95218086.1 ($p = 1.52 \times 10^{-8}$) of chromosome 6; OARX_120998827.1 ($p = 1.22 \times 10^{-7}$) of chromosome 27 and OAR3_132833292.1 ($p = 2.30 \times 10^{-7}$) of chromosome 3; OAR1_164254640.1 ($p = 5.08 \times 10^{-7}$) (**Figure 3**) of chromosome 1. Four SNPs, s55433.1 ($p = 3.26 \times 10^{-8}$) and OAR5_99879334.1 ($p = 3.26 \times 10^{-8}$) of chromosome 5, OARX_79209204.1 ($p = 3.26 \times 10^{-8}$) of chromosome 27, and s26859.1 ($p = 1.89 \times 10^{-7}$) of chromosome 1 (**Table 1**), were significantly correlated to chest circumference (**Figure 3**) in terms of genomic level; no SNP was significantly correlated to body length, tail width or tail length.

At the genomic level, annotation information of the 9 SNPs sites that were significantly correlated to body height and chest circumference are shown in **Table 1**. Four of the SNPs were within genes. OAR3_132833292.1 was within gene *KITLG*; OAR1_164254640.1 was within gene *CADM2*; OAR5_99879334.1 was within gene *MCTP1*; OARX_79209204.1 was within gene *COL4A6*. There were 5 other SNPs at intergenic regions: OAR23_3237800.1 located 368,856 bp downstream of *ZNF516*, OAR6_95218086.1 located 114,271 bp upstream of *NPFFR2*, OARX_120998827.1 located 10,161 bp upstream of *PRR32*, s55433.1 located 63,923 bp upstream of LOC101119639, and s26859.1 located 22053bp upstream of *SELENOF* (**Table 1**).

Group Verifying of SNPs Significantly Correlated to Hu Sheep's Body Height

The above 9 SNPs sites found by GWAS were verified by using the G3 generation 202 ewes of the nucleus herd for meat production. Nine products were amplified around nine

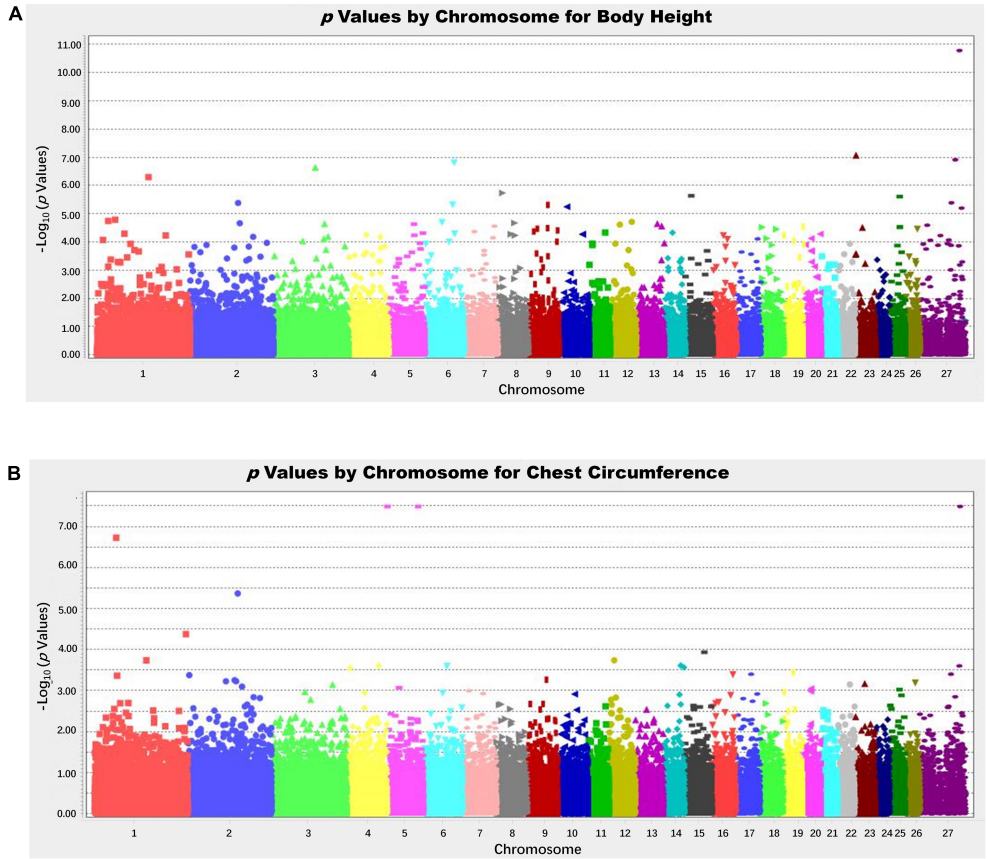

FIGURE 3 | Manhattan plot of analysis results; -$\log_{10}$ (*p*-values) in the studied population of Hu sheep. **(A)** Manhattan plot of the results of the body height analysis. **(B)** Manhattan plot of the results of the chest circumference.

TABLE 1 | Analysis of related SNPs associated with body size traits of Hu sheep.

Trait	Related SNPs	Chr	Position (bp)	*p*-value	add_effect *p*-value	Nearest gene distance[#] (bp)
Body height	OAR23_3237800.1	23	2959015	5.53E-08	9.21E-08	*ZNF516* -368856
	OAR6_95218086.1	6	86778759	1.52E-08	3.06E-08	*NPFFR2* + 114271
	OARX_120998827.1	27	105184061	1.22E-07	1.75E-08	*PRR32* + 10161
	OAR3_132833292.1	3	124516955	2.30E-07	NaN	*KITLG* within
	OAR1_164254640.1	1	152601830	5.08E-07	NaN	*CADM2* within
Chest circumference	s55433.1	5	413256	3.26E-08	NaN	*LOC101119639* + 63923
	OAR5_99879334.1	5	91556623	3.26E-08	NaN	*MCTP1* within
	OARX_79209204.1	27	119635739	3.26E-08	NaN	*COL4A6* within
	s26859.1	1	63255194	1.89E-07	NaN	*SELENOF* + 22053

p-values calculated from the mixed linear model analysis.
[#]*Positive value denotes the gene location downstream of SNP; negative value denotes the gene location upstream of SNP.*

sites (**Figure 4**), and 29 SNPs were found. Two mutation sites were detected in the amplified products at s55433.1, two mutation sites in the amplified products at OAR5_99879334.1, five in the amplified products at OARX_79209204.1, three in the amplified products at OAR23_3237800.1, three in the amplified products at s26859.1, six in the amplified products at OAR3_132833292.1, two in the amplified products at OAR6_95218086.1, two in the amplified products at OARX_120998827.1, and four mutation sites were detected in the amplified products at OAR1_164254640.1 (**Supplementary Table 4**).

Population genetic analysis was performed on 29 sites. All loci were subjected to genotyping, population genetic analysis, and the association analysis between SNPs and body size traits. The value of PIC at 24 sites was < 0.25, which was indicative of low polymorphism. The value of PIC at five sites was between 0.25 and 0.5, which was indicative of intermediate polymorphism (**Supplementary Tables 5**, **6**).

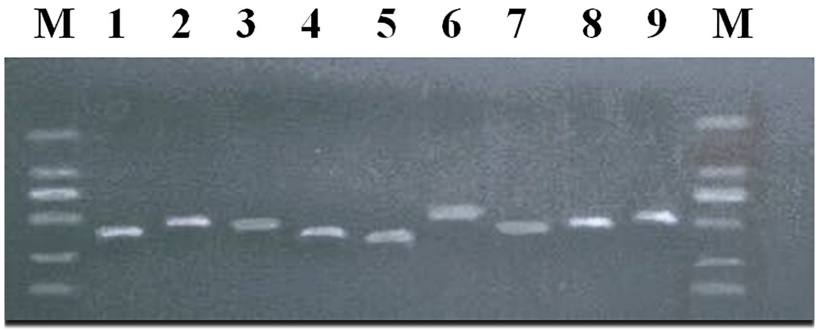

FIGURE 4 | PCR amplification products of body size-related SNPs in meat-type Hu sheep. M:DL2000 plus. (1) *LOC101119639* + 63923; (2) *MCTP1*; (3) *COL4A6*; (4) *ZNF516* -368856; (5) *SELENOF* + 22053; (6) *KITLG*; (7) *NPFFR2* + 114271; (8) *CADM2*; (9) *PRR32* + 10161.

TABLE 2 | Association analysis between SNPs and body height of Hu sheep.

SNP loci	Nearest gene distance# (bp)	Chr	Position (bp)	Genotype	Numbers	Body height (kg)
G > C mutation at 113 bp downstream of s554331	*LOC101119639* + 63789	5	413,369	CC	49	75.22 ± 0.43^{b}
				GG	153	76.49 ± 0.24^{a}
				p-value		0.011
T > G mutation at 19 bp upstream of s26859.1	*SELENOF* + 22072	1	63,255,175	GG	3	72.67 ± 1.74^{b}
				TG	51	75.53 ± 0.42^{b}
				TT	148	76.48 ± 0.25^{a}
				p-value		0.021
A > G mutation at 81 bp downstream of s26859.1	*SELENOF* + 21972	1	63,255,275	AA	154	76.46 ± 0.24^{a}
				AG	46	75.54 ± 0.44^{a}
				GG	2	70.00 ± 2.15^{b}
				p-value		0.003

Values with different superscripts for the same column have significant differences.

Verification results of SNPs showed that three mutation sites were significantly correlated to Hu sheep's body height (**Table 2**): G > C mutation at 134 bp downstream of s554331, T > G mutation at 19 bp upstream of s26859.1, A > G mutation at 81 bp downstream of s26859.1.

Linkage Disequilibrium and Haplotype Block Analyses

Linkage disequilibrium (Linkage Disequilibrium, LD) refers to the non-random co-occurrence of alleles of chromosomes or haplotypes, i.e., there are statistical associations between alleles at different sites, which are different from independent alleles. Usually, D′ and r^2 are used to measure LD. $D' > 0.33$ and $r^2 > 0.1$ represent a meaningful linkage disequilibrium; $D' > 0.8$ and $r^2 > 0.33$ a strong linkage disequilibrium (Long et al., 2004). According to the LD analysis results, the three SNPs showed strong linkage disequilibrium (**Figure 5**).

Three SNPs, g.63255175T/G, g.63255244G/C, and g.63255275A/G, were chosen for haplotype analysis based on linkage disequilibrium evaluation ($D' > 0.8$, $r^2 < 0.05$). Two tag SNPs (g.63255175T/G, g.63255275A/G) represented the genetic variation in the haplotype block. The effects of different haplotypes indicated a significant effect of haplotypes on the body height of Hu sheep (**Table 3**).

Effects of Candidate Functional SNPs on Gene Transcriptional Activity

The results of dual-luciferase reporter gene experiments (**Table 4**) showed that SNPs (G > C mutation at 134 bp downstream of s554331) significantly impacted the activity of dual-luciferase and decreased the activity of dual-luciferase after mutation ($p < 0.05$). Moreover, SNPs (T > G mutation at 19 bp upstream of s26859.1; A > G mutation at 81 bp downstream of s26859.1) significantly decreased the activity of dual-luciferase after mutation (all $p < 0.05$). Haplotypes (T > G mutation at 19 bp upstream of s26859.1, A > G mutation at 81 bp downstream of s26859.1) significantly affected the activity of the reporter gene ($p < 0.05$). These results showed that the above SNPs and haplotypes could significantly impact the activity of gene transcription.

DISCUSSION

Existing GWAS studies on sheep have mainly focused on reproductive traits (Demars et al., 2013; Gholizadeh et al., 2014; Martinez-Royo et al., 2017; Abdoli et al., 2019), body weight, and meat production traits (Zhang et al., 2013; Almamun et al., 2015; Matika et al., 2016), while few investigated body size traits. Thus far, no SNPs significantly correlated to body size traits at the

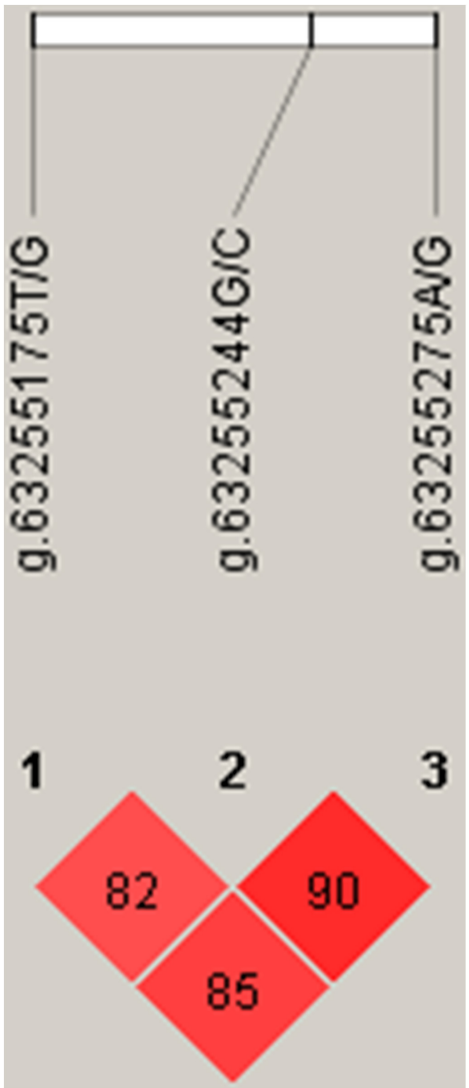

FIGURE 5 | Linkage disequilibrium (LD) analyses of SNPs near s26859.1.

TABLE 3 | Association analysis between haplotypes and body height of Hu sheep.

Loci	Haplotypes	Numbers	Body height (cm)
	GGCCAA	1	78.00^{ab}
	GGCCGG	2	70.00 ± 5.66^{c}
	TGCCAA	8	74.00 ± 3.21^{bc}
T > G mutation at 19 bp upstream of s26859.1	TGCCAG	32	75.50 ± 2.44^{b}
	TGGCAA	2	79.00 ± 0.00^{ab}
G > C mutation at 50 bp downstream of s26859.1	TGGCAG	7	75.57 ± 2.70^{b}
	TGGGA A	1	82.00^{a}
A > G mutation at 81 bp downstream of s26859.1	TGGGAG	1	75.00^{bc}
	TTCCAA	56	76.27 ± 3.22^{ab}
	TTCCAG	6	75.83 ± 3.37^{ab}
	TTGCAA	69	76.71 ± 2.96^{ab}
	TTGGAA	17	76.47 ± 2.79^{ab}

Note: Values with different superscript for the same column have significant differences.

TABLE 4 | Effects of genotypes or haplotypes of candidate functional SNPs on dual-luciferase activities.

SNPs locus	SNPs genotypes or haplotypes	M1/M2	*p*-value
G > C mutation at 134 bp downstream of s554331	GG	4.26 ± 0.44^{a}	0.009
	CC	2.14 ± 0.07^{b}	
T > G mutation at 19 bp upstream of s26859.1	TT	1.91 ± 0.53^{b}	0.000
	GG	4.53 ± 2.21^{a}	
A > G mutation at 81 bp downstream of s26859.1	AA	4.45 ± 2.30^{a}	0.000
	GG	1.99 ± 0.61^{b}	
T > G mutation at 19 bp upstream of s26859.1	TTAA	2.37 ± 0.08^{b}	0.000
	GGAA	6.53 ± 0.49^{a}	
A > G mutation at 81 bp downstream of s26859.1	TTGG	1.45 ± 0.24^{c}	
	GGGG	2.53 ± 0.10^{b}	

Note: Values with different superscript for the same column have significant differences.

genomic level were identified at $p < 0.05$. Eleven chromosome-wide significant SNPs, five for the "width Dimension" factor, four for the "height Dimension" factor, and two for the "length Dimension" factor were confirmed at $p < 0.10$ (Kominakis et al., 2017). One SNP (OAR17_14085599) was found to be significantly correlated to chest circumference. No SNPs were found to be correlated to body length and height (Zhang et al., 2019). In the present study, we found that 5 SNPs were significantly correlated to body height, and 4 SNPs were significantly correlated to chest circumference. Compared to the previous two studies, these SNPs were significantly different.

Our results identified nine significant SNPs at the genomic level. After performing genome annotation on the 9 SNPs, some candidate functional genes correlated to Hu sheep's body height and chest circumference were found: *KITLG* and *CADM2* are candidate functional genes correlated to body height, while *MCTP1* and *COL4A6* are candidate functional genes correlated to chest circumference.

OAR1_164254640.1 is within *CADM2* (Gene ID: 101120371). Cell adhesion molecules (*CADM*) consist of a protein family that maintains cell polarity. Most *CADM* belong to the immunoglobulin superfamily. Previous studies have shown that *CADM* can be used as a tumor inhibitor (He et al., 2013). *CADM2* belongs to the *CADM* family. *CADM2* activates methylation and/or heterogeneity loss by promoting DNA to contain human kidney clear cell carcinoma. The loss of *CADM2* leads to tumor progression (He et al., 2013). Previous genome-wide association meta-analysis confirmed several susceptibility sites to be correlated to BMI, including *CADM2* (Speliotes et al., 2010; Locke et al., 2015). Moreover, obesity and glucose level can be reduced, and insulin sensitivity, sports function, energy expenditure rate, and core temperature can be increased in cadm2-knockout mice, emphasizing its relevance in systematic energy balance (Yan et al., 2018). Moreover, *CADM2* is related to a series of behavioral and metabolic features, including physical

activity, adventure, educational level, and obesity (Morris et al., 2019). It has been proved that *CADM2* gene mutation has a critical role in BMI through the central nervous system (Speakman et al., 2018).

OAR5_99879334.1 is within gene *MCTP1*, a neuronal vesicle/endosome protein. In terms of structure, *MCTP* protein contains 3 C2 domains and 2 transmembrane domains near the C-end (Shin et al., 2005). The mutation or expression of *MCTP1* variants is related to neuro psychosis. Genome-wide analysis shows that *MCTP1* single nucleotide polymorphism (SNP), rs17418283, is related to bipolar affective disorder (Scott et al., 2009). *In vivo* and *in vitro* imaging studies all identified the location of *MCTP1* on the endocrine recovery approach. Moreover, functional tests have shown that *MCTP1* participates in various cell functions, including endocrine, cell migration, and anti-excitement virulence of neuronal cells (Qiu et al., 2015).

OAR3_132833292.1 is within the *KITLG* gene known as mammary gland cell growth factor (MGF) or stem cell factor (SCF). It encodes the ligand of c-Kit, a receptor tyrosine kinase, and participates in many biological processes, including hematopoiesis, gametogenesis, and melanogenesis (Talenti et al., 2018). The *KITLG* gene affects pigmentation in both humans and mice (Guenther et al., 2014). Polymorphisms in the *KITLG* gene have already been associated with litter size in goats (An et al., 2012, 2015). The genomic analysis suggested that *KITLG* is Responsible for a Roan Pattern in two Pakistani Goat Breeds (Talenti et al., 2018). Pigmentation genes *KITLG* have also been shown to have strong selection characteristics on Tibetan Cashmere Goat (Guo et al., 2019).

OARX_79209204.1 is within the *COL4A6* gene (Collagen Type IV Alpha 6 Chain), a protein-coding gene that encodes the alpha-6 chain of type IV collagen basal membranes. The genes *COL4A5* and *COL4A6* are located head-to-head near human chromosome Xq22. *COL4A6* activates transcription with 2 selectable promoters in a particular way of the tissue (Sugimoto et al., 1994). Gene Ontology (GO) annotations related to this gene include structural molecule activity and extracellular matrix structural constituent. Wang et al. (2018) suggested that the two dislocation mutations in *COL4A5* and *COL4A6* could be risk factors for cerebrovascular fibromuscular dysplasia. The dislocation mutation of the *COL4A6* gene causes serious non-syndromic hearing impairment in males (Rost et al., 2014). Downregulation of *COL4A6* may promote prostate cancer progression and invasion (Ma et al., 2020).

The 9 SNPs found in GWAS were verified by using 202 G3 generation ewes of the nucleus herd. Nine products were amplified around the 9 sites, and 29 SNPs were found using direct sequencing. Ovine SNP50 BeadChip developed by Illumina contains 54,241 SNP sites, with a mark on average every 46 kb. However, it is estimated that 1 SNP will appear every 1,000 bp in the human genome. Due to the insufficient SNP density of GWAS chips for commercial use, only 15% of genetic variation could be tested; thus, a large amount of genetic variation is yet to be found. Because of such GWAS defects, follow-up herd verification is essential. Herd verification is not only a supplement for the results of GWAS analysis but may also reveal new sites during the verification.

Our results revealed four new SNPs in follow-up verification that were significantly correlated to Hu sheep's body size traits. The results of dual-luciferase reporter gene experiments showed that the 4 SNPs could significantly impact gene transcription activity. These significant sites can be included in our analysis field because they are close to GWAS positive sites. Therefore, we believe that GWAS is an important tool for candidate functional genes and the screening of functional SNP that can be used as a signpost to guide follow-up verification, thus preventing researchers from being overwhelmed by sequential information. However, a higher-density SNP detection chip may greatly improve the reliability of the results.

DATA AVAILABILITY STATEMENT

Data supporting this study has been deposited in GEO—accession number GSE152717.

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethical Committee of Zhejiang Academy of Agricultural Sciences.

AUTHOR CONTRIBUTIONS

YJ was in charge of the whole trial. XS designed the experiments. JJ completed the majority of the experiments and was a major contributor in writing the manuscript. HS and YC participated in sampling and laboratory analyses. JW for animal feeding and care. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the Zhejiang Provincial Major Science and Technology Projects on Agricultural New Varieties Selection and Breeding (2016C02054-8) and the Zhejiang Agriculture Science and Technology Cooperation Projects (2019SNLF016).

ACKNOWLEDGMENTS

We thank China Hangzhou Giant Agricultural Development Co., Ltd. for support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.642552/full#supplementary-material

REFERENCES

Abdoli, R., Mirhoseini, S. Z., Ghavi Hossein-Zadeh, N., Zamani, P., Ferdosi, M. H., and Gondro, C. (2019). Genome-wide association study of four composite reproductive traits in Iranian fat-tailed sheep. *Reprod. Fertil. Dev.* 31, 1127–1133. doi: 10.1071/RD18282

Almamun, H. A., Kwan, P., Clark, S. A., Ferdosi, M. H., Tellam, R., Gondro, C., et al. (2015). Genome-wide association study of body weight in Australian merino sheep reveals an orthologous region on OAR6 to human and bovine genomic regions affecting height and weight. *Genet. Sel. Evol.* 47:66.

An, X. P., Hou, J. X., Gao, T. Y., Lei, Y. N., Song, Y. X., Wang, J. G., et al. (2015). Association analysis between variants in KITLG gene and litter size in goats. *Gene* 558, 126–130. doi: 10.1016/j.gene.2014.12.058

An, X. P., Hou, J. X., Li, G., Song, Y. X., Wang, J. G., Chen, Q. J., et al. (2012). Polymorphism identification in the goat KITLG gene and association analysis with litter size. *Anim. Genet.* 43, 104–107. doi: 10.1111/j.1365-2052.2011.02219.x

Brym, P., Kaminski, S., and Wojcik, E. (2005). Nucleotide sequence polymorphism within exon 4 of the bovine prolactin gene and its associations with milk performance traits. *J. Appl. Genet.* 46, 179–185.

Cao, Y., Song, X., Shan, H., Jiang, J., Xiong, P., Wu, J., et al. (2020). Genome-wide sssociation study of body weights in Hu Sheep and population verification of related single-nucleotide polymorphisms. *Front. Genet.* 11:588. doi: 10.3389/fgene.2020.00588

Demars, J., Fabre, S., Sarry, J., Rossetti, R., Gilbert, H., Persani, L., et al. (2013). Genome-wide association studies identify two novel BMP15 mutations responsible for an atypical hyperprolificacy phenotype in sheep. *PLoS Genet.* 9:e1003482. doi: 10.1371/journal.pgen.1003482

Gholizadeh, M., Rahimimianji, G., Nejatijavaremi, A., De Koning, D. J., and Jonas, E. (2014). Genomewide association study to detect QTL for twinning rate in baluchi sheep. *J. Genet.* 93, 489–493. doi: 10.1007/s12041-014-0372-1

Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A., and Kingsley, D. M. (2014). A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* 46, 748–752. doi: 10.1038/ng.2991

Guo, J., Zhong, J., Li, L., Zhong, T., Wang, L., Song, T., et al. (2019). Comparative genome analyses reveal the unique genetic composition and selection signals underlying the phenotypic characteristics of three Chinese domestic goat breeds. *Genet. Sel. Evol.* 51:70. doi: 10.1186/s12711-019-0512-4

He, W., Li, X., Xu, S., Ai, J., Gong, Y., Gregg, J. L., et al. (2013). Aberrant methylation and loss of CADM2 tumor suppressor expression is associated with human renal cell carcinoma tumor progression. *Biochem. Biophys. Res. Commun.* 435, 526–532. doi: 10.1016/j.bbrc.2013.04.074

Kemper, K. E., Visscher, P. M., and Goddard, M. E. (2012). Genetic architecture of body size in mammals. *Genome Biol.* 13:244. doi: 10.1186/gb-2012-13-4-244

Kominakis, A., Hager-Theodorides, A. L., Zoidis, E., Saridaki, A., Antonakos, G., and Tsiamis, G. (2017). Combined GWAS and 'guilt by association'-based prioritization analysis identifies functional candidate genes for body size in sheep. *Genet. Sel. Evol.* 49:41. doi: 10.1186/s12711-017-0316-3

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., and Day, F. R. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206.

Long, J. R., Zhao, L. J., and Liu, P. Y. (2004). Patterns of linkage disequilibrium and haplotype distribution in disease candidate genes. *BMC Genetics* 5:11. doi: 10.1186/1471-2156-5-11

Ma, J. B., Bai, J. Y., Zhang, H. B., Gu, L., He, D., and Guo, P. (2020). Downregulation of collagen COL4A6 is associated with prostate cancer progression and metastasis. *Genet. Test Mol. Biomarkers* 24, 399–408. doi: 10.1089/gtmb.2020.0009

Martinez-Royo, A., Alabart, J. L., Sarto, P., Serrano, M., Lahoz, B., Folch, J., et al. (2017). Genome-wide association studies for reproductive seasonality traits in Rasa Aragonesa sheep breed. *Theriogenology* 99, 21–29. doi: 10.1016/j.theriogenology.2017.05.011

Matika, O., Riggio, V., Anselme-Moizan, M., Law, A. S., Pong-Wong, R., Archibald, A. L., et al. (2016). Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. *Genet. Sel. Evol.* 48:11.

Morris, J., Bailey, M. E. S., Baldassarre, D., Cullen, B., de Faire, U., Ferguson, A., et al. (2019). Genetic variation in CADM2 as a link between psychological traits and obesity. *Sci Rep.* 9:7339.

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

Qiu, L., Yu, H., and Liang, F. (2015). Multiple C2 domains transmembrane protein 1 is expressed in CNS neurons and possibly regulates cellular vesicle retrieval and oxidative stress. *J. Neurochem.* 135, 492–507. doi: 10.1111/jnc.13251

Rost, S., Bach, E., Neuner, C., Nanda, I., Dysek, S., Bittner, R. E., et al. (2014). Novel form of X-linked nonsyndromic hearing loss with cochlear malformation caused by a mutation in the type IV collagen gene COL4A6. *Eur. J. Hum. Genet.* 22, 208–215. doi: 10.1038/ejhg.2013.108

Scott, L. J., Muglia, P., Kong, X. Q., Guan, W., Flickinger, M., Upmanyu, R., et al. (2009). Genome−wide association and meta−analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7501–7506.

Shin, O. H., Han, W., Wang, Y., and Südhof, T. C. (2005). Evolutionarily conserved multiple C2 domain proteins with two transmembrane regions (MCTPs) and unusual Ca2 + binding properties. *J. Biol. Chem.* 280, 1641–1651. doi: 10.1074/jbc.m407305200

Speakman, J. R., Loos, R. J. F., O'Rahilly, S., Hirschhorn, J. N., and Allison, D. B. (2018). GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. *Int. J. Obes.* 42, 1524–1531. doi: 10.1038/s41366-018-0147-5

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., and Jackson, A. U. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948.

Sugimoto, M., Oohashi, T., and Ninomiya, Y. (1994). The genes COL4A5 and COL4A6, coding for basement membrane collagen chains alpha 5(IV) and alpha 6(IV), are located head-to-head in close proximity on human chromosome Xq22 and COL4A6 is transcribed from two alternative promoters. *Proc. Natl. Acad. Sci. U.S.A.* 91, 11679–11683. doi: 10.1073/pnas.91.24.11679

Talenti, A., Bertolini, F., Williams, J., Moaeen-Ud-Din, M., Frattini, S., Coizet, B., et al. (2018). Genomic analysis suggests KITLG is responsible for a roan pattern in two pakistani goat breeds. *J. Hered.* 109, 315–319. doi: 10.1093/jhered/esx093

Wang, X., Li, W., Wei, K., Xiao, R., Wang, J., Ma, H., et al. (2018). Missense mutations in COL4A5 or COL4A6 genes may cause cerebrovascular fibromuscular dysplasia: case report and literature review. *Medicine* 97:e11538. doi: 10.1097/MD.0000000000011538

Yan, X., Wang, Z., Schmidt, V., Gauert, A., Willnow, T. E., Heinig, M., et al. (2018). Cadm2 regulates body weight and energy homeostasis in mice. *Mol. Metab.* 8, 180–188. doi: 10.1016/j.molmet.2017.11.010

Yue, G. H. (1996). Reproductive characteristics of Chinese hu sheep. *Anim. Reprod. Sci.* 44, 223–230. doi: 10.1016/0378-4320(96)01562-x

Zhang, L., Liu, J., Zhao, F., Ren, H., Xu, L., Lu, J., et al. (2013). Genome-wide association studies for growth and meat production traits in sheep. *PLoS One* 8:e66569. doi: 10.1371/journal.pone.0066569

Zhang, T., Gao, H., Sahana, G., Zan, Y., Fan, H., Liu, J., et al. (2019). Genome-wide association studies revealed candidate genes for tail fat deposition and body size in the Hulun Buir sheep. *J. Anim. Breed. Genet.* 136, 362–370. doi: 10.1111/jbg.12402

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 July 2021
doi: 10.3389/fgene.2021.667300

Reliabilities of Genomic Prediction for Young Stock Survival Traits Using 54K SNP Chip Augmented With Additional Single-Nucleotide Polymorphisms Selected From Imputed Whole-Genome Sequencing Data

Grum Gebreyesus, Mogens Sandø Lund, Goutam Sahana and Guosheng Su*

Center for Quantitative Genetics and Genomics, Aarhus University, Tjele, Denmark

OPEN ACCESS

Edited by:
Ruidong Xiang,
The University of Melbourne, Australia

Reviewed by:
Christian Maltecca,
North Carolina State University, United States
Renata Veroneze,
Universidade Federal de Viçosa, Brazil

***Correspondence:**
Grum Gebreyesus
grum.gebreyesus@qgg.au.dk

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

Received: *12 February 2021*
Accepted: *23 June 2021*
Published: *19 July 2021*

Citation:
Gebreyesus G, Lund MS, Sahana G and Su G (2021) Reliabilities of Genomic Prediction for Young Stock Survival Traits Using 54K SNP Chip Augmented With Additional Single-Nucleotide Polymorphisms Selected From Imputed Whole-Genome Sequencing Data. Front. Genet. 12:667300. doi: 10.3389/fgene.2021.667300

This study investigated effects of integrating single-nucleotide polymorphisms (SNPs) selected based on previous genome-wide association studies (GWASs), from imputed whole-genome sequencing (WGS) data, in the conventional 54K chip on genomic prediction reliability of young stock survival (YSS) traits in dairy cattle. The WGS SNPs included two groups of SNP sets that were selected based on GWAS in the Danish Holstein for YSS index (YSS_SNPs, n = 98) and SNPs chosen as peaks of quantitative trait loci for the traits of Nordic total merit index in Denmark–Finland–Sweden dairy cattle populations (DFS_SNPs, n = 1,541). Additionally, the study also investigated the possibility of improving genomic prediction reliability for survival traits by modeling the SNPs within recessive lethal haplotypes (LET_SNP, n = 130) detected from the 54K chip in the Nordic Holstein. De-regressed proofs (DRPs) were obtained from 6,558 Danish Holstein bulls genotyped with either 54K chip or customized LD chip that includes SNPs in the standard LD chip and some of the selected WGS SNPs. The chip data were subsequently imputed to 54K SNP together with the selected WGS SNPs. Genomic best linear unbiased prediction (GBLUP) models were implemented to predict breeding values through either pooling the 54K and selected WGS SNPs together as one genetic component (a one-component model) or considering 54K SNPs and selected WGS SNPs as two separate genetic components (a two-component model). Across all the traits, inclusion of each of the selected WGS SNP sets led to negligible improvements in prediction accuracies (0.17 percentage points on average) compared to prediction using only 54K. Similarly, marginal improvement in prediction reliability was obtained when all the selected WGS SNPs were included (0.22 percentage points). No further

improvement in prediction reliability was observed when considering random regression on genotype code of recessive lethal alleles in the model including both groups of the WGS SNPs. Additionally, there was no difference in prediction reliability from integrating the selected WGS SNP sets through the two-component model compared to the one-component GBLUP.

Keywords: young stock survival, genomic prediction, GWAS, whole-genome sequencing, recessive lethal alleles

INTRODUCTION

Young stock mortality represents a major economic loss for dairy farmers due, for instance, to fewer heifers available for replacement in the production system, fewer male calves for slaughter, higher veterinarian cost, and cost related to disposal of dead calf. In the Nordic countries, annual total loss due to dairy calf mortality (including stillbirth) is estimated to be approximately €70 million (Østerårs et al., 2007). In addition, young stock mortality poses a large animal welfare issue and threatens the public perceptions of the dairy industry.

Part of the variation in young stock mortality is genetic with reported heritability estimates ranging from 0.00 to 0.08 (e.g., Hansen et al., 2003; Fuerst-Waltl and Sørensen, 2010; Henderson et al., 2011). In the Nordic countries, young stock survival (YSS) in calves is included in the Nordic total merit (NTM) index (NAV).[1] A challenge in the genetic evaluation for YSS traits is the low heritability leading to low prediction accuracies. Theoretically, there are possibilities to improve the reliability of genomic prediction models by incorporating causative variants (if known) or markers highly correlated with them (de Los Campos et al., 2013).

Genome-wide association studies (GWASs) based on sequence data have shown high power to identify putative causative variants and strong signals of association for various economic traits in cattle (Daetwyler et al., 2014; Sahana et al., 2014; Wu et al., 2017). Studies have shown that genomic prediction models incorporating single-nucleotide polymorphisms (SNPs) selected from whole-genome sequencing (WGS) data based on such GWASs lead to improved accuracy of prediction of breeding values for some traits. Brøndum et al. (2015) added quantitative trait loci (QTLs) from GWAS to genomic prediction models and achieved up to 5 percentage point increase in accuracy for milk production traits. Similarly, Liu et al. (2019) reported gains in prediction reliability for milk production traits in the Danish Jersey by integrating selected WGS variants with the 54K SNP chip. A GWAS by Wu et al. (2017) using WGS data reported interesting genomic regions across the *Bos taurus* autosome (BTA) significantly associated with the YSS index trait in the NTM index. Incorporating such WGS variants from GWASs might enable improvement of genomic prediction reliability for YSS traits. Additionally, the genetic underpinnings of young stock and calf mortality can be partly polygenic and partly due to deleterious effects of recessive lethal alleles (Gebreyesus et al., 2020). Several studies have reported haplotypes with harmful recessive effects on fertility and responsible for early embryonic lethality and stillbirth in various cattle breeds (e.g., VanRaden et al., 2011; Sahana et al., 2016; Hoff et al., 2017; Wu et al., 2019), which might have an important predictive ability for breeding values for YSS traits.

We hypothesize in this study that incorporation of WGS variants selected based on previous GWASs and variants within previously reported deleterious haplotypes might improve the reliability of genomic prediction for YSS traits. The objective of this study was therefore to investigate effects of integrating SNPs selected, based on previous studies, from imputed WGS data in the conventional 54K chip on genomic prediction of YSS traits in the Nordic Holstein cattle. Additionally, we also assessed the possibility of improving genomic prediction reliability for survival traits by considering in the prediction model the effect of SNPs located within recessive lethal haplotypes previously reported in the Nordic Holstein.

MATERIALS AND METHODS

Ethics Approval Statement

All procedures to collect the DNA samples followed the protocols approved by the National Guidelines for Animal Experimentation and the Danish Animal Experimental Ethics Committee, and hence, no specific permission was required.

Animals and Genotypes

A total of 6,558 Nordic Holstein bulls were genotyped with the Illumina Bovine SNP50 chip (54K, Illumina, Inc.). A reference population of 129,000 Holstein cows and bulls was also available for the imputation that were genotyped mostly with the EuroGenomics customized chip (Boichard et al., 2018) that included SNPs in the standard Illumina Bovine LD chip together with SNPs identified as causal mutation, functional annotation, or association with economic traits. The EuroGenomics customized chip that started with the standard LD chip (Boichard et al., 2018) is updated every year with selected variants and currently includes 70K SNPs including most of the variants in the conventional 54K chip along with additional selected SNPs. A total of 1,754 selected WGS SNPs, selected by GWAS in Denmark–Finland–Sweden dairy cattle populations (DFS_SNPs), are included in the EuroGenomics chip. The DFS_SNPs were peaks of QTL detected from imputed WGS data for 16 index traits included in the NTM index, which includes the YSS index. The selection of the DFS SNPs was undertaken within each breed according to *p*-values of a single-marker regression model while considering functional

[1] www.nordicebv.info

annotations and linkage disequilibrium between SNPs (Brøndum et al., 2015). Before the imputation, 54K genotypes were subjected to quality control using the minor allele frequency (MAF) threshold of 0.05. Bulls genotyped with 54K and the custom chips were imputed to 54K + DFS using FImpute software (Sargolzaei et al., 2014). Additionally, another set of WGS SNPs (147 SNPs) were selected from GWAS by Wu et al. (2017) for survival index (YSS_SNPs). The genotypes of these SNPs for the bulls in this study were imputed using the 1,000 bull genome data as reference and using the Minimac3 v.2.0.1 software (Das et al., 2016). The SNP-wise imputation accuracy was measured as the Pearson correlation between observed and imputed genotypes (coded as 0, 1, or 2) and the proportion of correctly imputed genotypes to all imputed genotypes (i.e., concordance). Only SNPs with both correlation and concordance higher than 0.80 were used in genomic prediction. Ultimately, 39,803 SNPs in the 54K chip, 1,541 DFS_SNPs, and 98 YSS_SNPs were kept for genomic prediction, with 22 SNPs overlapped between DFS and YSS_SNPs. The average imputation accuracy for SNPs used in genomic prediction was 0.977 for standard LD chip to 54K, 0.980 for DFS_SNPs, and 0.923 for YSS_SNPs, while concordance was 0.960 for standard LD chip to 54K, 0.962 for DFS_SNPs, and 0.955 for YSS_SNPs.

Of the 39,803 SNPs in the 54K chip used for the genomic prediction, 130 SNPs (LET_SNP) were within recessive lethal haplotypes reported by Wu et al. (2019) in the Nordic Holstein. The study of Wu et al. (2019) reported a total of 11 haplotypes of which nine were completely homozygous-deficient while two had significantly lower homozygotes observed than expected.

Phenotypes

The traits included in the analyses were four different definitions of YSS (sub-traits) and an index trait (YSS index) derived from these four sub-traits. The sub-traits were as follows:

i) Bull period 1 (BP1): Bull calf survival day in the period 1–30 days;
ii) Bull period 2 (BP2): Bull calf survival day in the period 31–183 days;
iii) Heifer period 1 (HP1): Heifer calf survival day in the period 1–30 days;
iv) Heifer period 2 (HP2): Heifer calf survival day in the period 31–458 days.

Calf death and survival during each period were recorded as 0 and 1, respectively. Calves slaughtered, exported, or with missing records were recorded as missing. The YSS index was calculated by combining the estimated breeding values (EBVs) for the sub-traits, i.e., BP1, BP2, HP1, and HP2, by the Nordic Cattle Genetic Evaluation center (NAV, Denmark), which were weighted by their relative economic values and standardized in terms of mean and standard deviation (Pedersen et al., 2015).

De-regressed proof (DRP) derived from official EBV was used as the pseudo phenotype in the genomic prediction. The official EBVs were calculated using linear models by the Nordic Cattle Genetic Evaluation center as described in NAV (Nordic Cattle Genetic Evaluation) (2017). DRPs were derived using the official EBVs based on the standard method described in Jairath et al. (1998) and implemented using the mix99 program (Strandén, 2015).

The reliability of DRP was calculated as:

$$r^2_{DRPi} = \frac{EDC_i}{EDC_i + \lambda'} \quad (1)$$

where $\lambda = \frac{4-h^2}{h^2}$. The EDC_i was the effective daughter contribution of i^{th} bull, and h^2 was the heritability for each trait as used in the official Nordic Cattle Genetic Evaluation (Pedersen et al., 2015). The heritability estimates and mean DRP reliability for each trait are given in **Table 1**, and histogram plots showing reliability distributions are presented in **Figure 1**.

Statistical Analysis

Implemented prediction models included linear mixed model using pedigree-based best linear unbiased prediction (PBLUP)- or genomic best linear unbiased prediction (GBLUP)-based relationships. Different scenarios were investigated to study the effect of adding selected WGS SNPs and modeling recessive lethal SNPs on prediction reliability. These include:

(i) Only using 54K;
(ii) 54K plus YSS_SNPs (54K + YSS);
(iii) 54K plus DFS_SNPs (54K + DFS);
(iv) 54K plus YSS_SNPs and DFS_SNPs (54K + YSS + DFS);
(v) Reduced 54K (minus SNPs in recessive lethal haplotypes), plus YSS_SNPs and DFS_SNPS, and the model considered random regression on genotype code of LET_SNPs (54K* + YSS + DFS + LET).

In addition, two approaches of integrating the selected SNPs were assessed. Accordingly, one-component model pooling the selected WGS SNPs together with the 54K SNPs as one genetic component and two-component model considering 54K SNPs and selected WGS SNPs as two separate genetic components were implemented and compared for prediction accuracy.

The PBLUP model fitted was:

$$y = \mathbf{1}\mu + \mathbf{Za} + \mathbf{e} \quad (2)$$

where y is the vector of DRPs; **1** is the vector of ones; μ is the overall mean; **a** is the vector of additive genetic

TABLE 1 | Heritability estimates and mean reliability of DRPs used in the genomic prediction of the young stock survival traits.

Trait	h^2*	Mean DRP reliability
YSS Index	0.014	0.698
BP1	0.007	0.611
BP2	0.027	0.742
HP1	0.009	0.626
HP2	0.011	0.737

**Heritability estimates used in the official Nordic evaluations (Pedersen et al., 2015). BP1, Bull period 1; BP2, Bull period 2; DRP, de-regressed proof; HP1, Heifer period 1; HP2, Heifer period 2; YSS, young stock survival.*

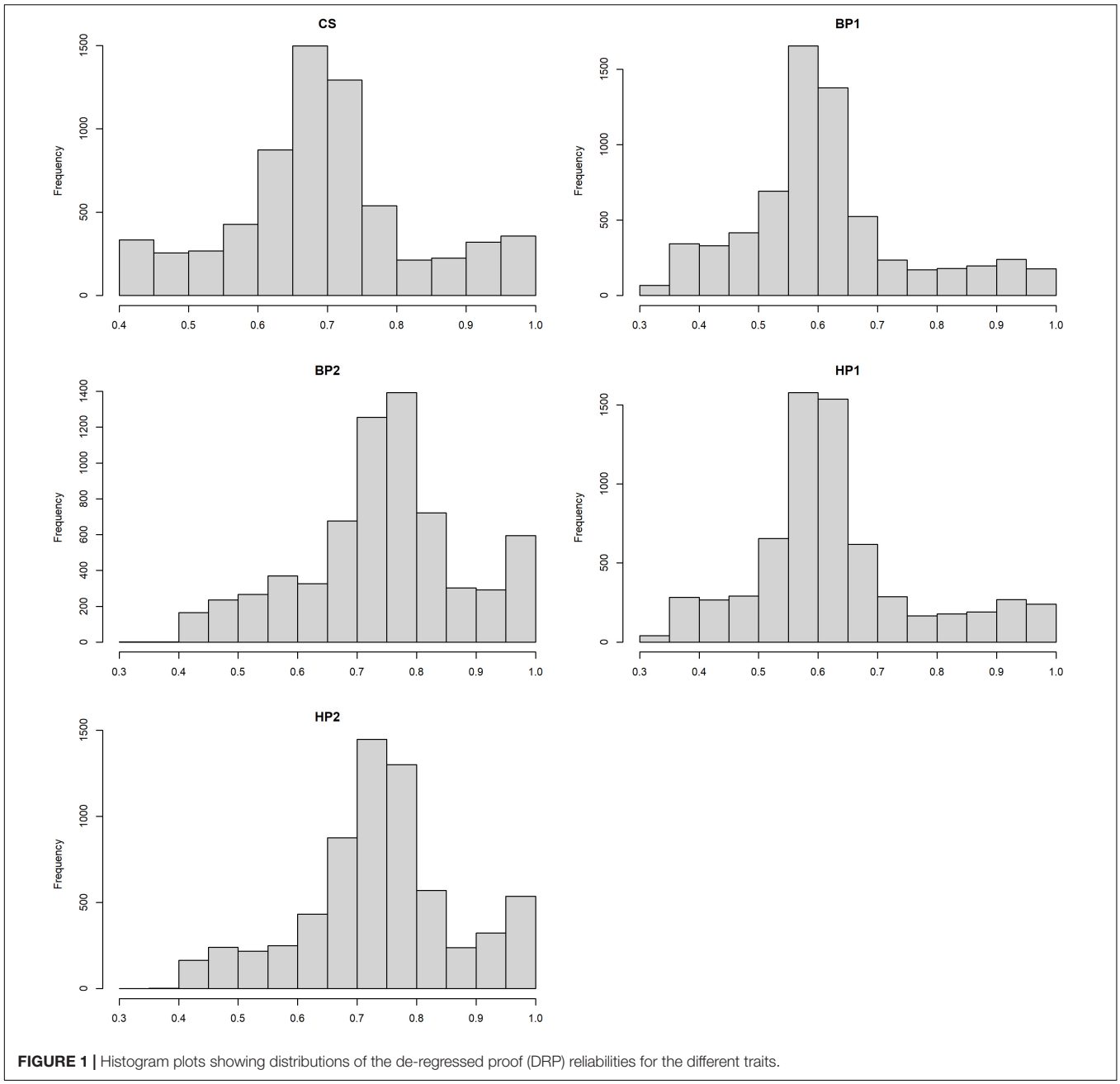

FIGURE 1 | Histogram plots showing distributions of the de-regressed proof (DRP) reliabilities for the different traits.

effects; **Z** is the incidence matrix relating **a** to phenotypes; and **e** is the vector of random residuals. It was assumed that $\mathbf{a} \sim \mathrm{N}\left(0, \mathbf{A}\sigma_a^2\right)$ and $\mathbf{e} \sim \mathrm{N}\left(0, \mathbf{D}\sigma_e^2\right)$. The **A** was the additive relationship matrix constructed from the pedigree that traced genotyped animals five generations back and included a total of 16,763 animals. The **D** is a diagonal matrix with elements $d_i = \frac{1-r^2_{DRPi}}{r^2_{DRPi}}$ for each bull i to account for heterogeneous residual variances due to differences in reliability of DRPs (r^2_{DRPi}) calculated as in Eq. 1.

The following one-component GBLUP models were fitted:

$$y = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e} \quad (3)$$

where **g** is the additive genetic effect with $\mathbf{g} \sim \mathrm{N}\left(0, \mathbf{G}\sigma_a^2\right)$, where **G** is the genomic relationship matrix (GRM) constructed using SNPs described in the different scenarios of adding WGS SNPs (YSS, DFS, or YSS + DFS) on the conventional 54K, while the remaining terms of the model are as described in model 2.

Additionally, a one-component GBLUP model considering random regression on the genotype code of the recessive lethal SNPs was implemented:

$$y = \mathbf{1}\mu + \mathbf{Mb} + \mathbf{Zg}^* + \mathbf{e} \quad (4)$$

where **M** is a matrix of genotype code (0, 1, or 2) for recessive lethal SNPs with dimension of 6,558 (number of individuals)

by 130 (number of recessive lethal SNPs), **b** is the vector of random regression coefficients on genotype code of recessive lethal SNPs (n = 130), and $\mathbf{g}^*$ is the random additive genetic effect based on GRM constructed using all SNPs (54K + YSS + DFS) excluding SNPs within recessive lethal haplotypes. The random regression coefficient b is assumed to be normally distributed:$\mathbf{b} \sim N\left(0, \mathbf{I}\sigma_b^2\right)$, where **I** is an identity matrix and σ_b^2 is the variance of the regression coefficient estimates. In addition to the one-component models, genomic breeding values were also predicted using a two-component GBLUP model that accounted for the difference between effects of the 54K SNPs and effects of selected WGS SNPs. The two-component model for the 54K and WGS data was:

$$y = \mathbf{1}\mu + \mathbf{Zg_{54K}} + \mathbf{Zg_{WGS}} + \mathbf{e} \quad (5)$$

Additionally, a two-component model considering random regression on the genotype code of the recessive lethal SNPs was implemented:

$$y = \mathbf{1}\mu + \mathbf{Mb} + \mathbf{Zg_{54K^*}} + \mathbf{Zg_{WGS}} + \mathbf{e} \quad (6)$$

where **M** and **b** are as described in model 4, $\mathbf{g_{54K^*}}$ is the additive genetic effect based on GRM constructed with 54K SNPs excluding the SNPs within recessive lethal haplotypes, $\mathbf{g_{WGS}}$ is the random genetic effect based on GRM constructed WGS SNPs (either DFS or YSS GWAS SNPs, or both, depending on the considered scenario).

An additional three-component GBLUP model was run to estimate the proportion of genomic variance explained by the SNP sets, i.e., 54K, YSS_SNPs, and DFS_SNPs by extending model 5 as follows:

$$y = \mathbf{1}\mu + \mathbf{Zg}_{54K} + \mathbf{Zg}_{YSS} + \mathbf{Zg}_{DFS} + \mathbf{e} \quad (7)$$

The proportion of the genomic variance explained by each SNP set of the three-component GBLUP model was then computed as:

$$\%var_{SNPset_i} = \frac{\sigma^2_{SNPset_i}}{\sigma^2_{total}} \times 100, \quad (8)$$

where $\sigma^2_{SNPset_i}$ was the additive genetic variance estimated based on the GRM corresponding to each SNP set (54K, DFS, and YSS), and σ^2_{total} was the total genomic variance computed as:

$$\sigma^2_{total} = \sigma^2_{54K} + \sigma^2_{YSS} + \sigma^2_{DFS} \quad (9)$$

All GRMs used for the different scenarios were calculated using the first method presented by VanRaden (2008), and SNP allele frequencies for building GRMs were calculated directly from the SNP data.

All models were implemented using the DMU software (Madsen and Jensen, 2013).

Computation of Prediction Reliabilities

The studies of Wu et al. (2017, 2019) used part of the current dataset (bulls born on or before the year 2009) to detect the WGS markers for YSS and the recessive lethal haplotypes, respectively. Therefore, the validation set in the current study consisted of only bulls born after the year 2010 (n = 1,312), and the rest was used as the training population (n = 5,246).

Reliability of genomic prediction was computed as the squared correlation between estimated breeding values (GEBVs) and DRP divided by the average reliability of DRP for the bulls in the validation population. For the two-component GBLUP models, the total GEBV for each individual was computed by summing together the breeding values from the two components. Bias of prediction was measured as the regression coefficient of DRP on the estimated breeding values for the bulls in the validation population. Reliability and bias were then compared among different models.

For the model considering random regression on genotype codes of recessive lethal alleles, effects of the recessive lethal alleles from the random regression coefficients were added to the GEBVs to calculate the correlation with DRP and subsequently compute the reliability.

In addition, model fit for the different models was assessed and compared using the Akaike information criteria (AIC; Akaike, 1974).

RESULTS

Proportion of the Genetic Variance Explained by the Different Single-Nucleotide Polymorphism Sets

Figure 2 presents the percentages of total genomic variance explained by the different SNP sets, i.e., 54K SNPs, YSS_SNPs, and DFS_SNPs, in the different YSS sub-traits and the index trait. In general, at least 80% of the total genetic variance in all the traits is explained by the SNPs in the standard 54K chip. On average, the YSS_SNPs explained 6% of the genetic variation, while the DFS_SNPs explained 11%. Across the traits, the proportion of total genetic variance explained by YSS_SNPs (4.2%) and DFS_SNPs (9.5%) was lowest for BP2, which was 5% and 10.2% for YSS_SNPs and DFS_SNPs, respectively.

Genomic Prediction Reliabilities and Bias

Table 2 presents genomic prediction accuracies using PBLUP and the GBLUP models that use different SNP sets. In general, across all scenarios, prediction reliability was lowest in the YSS index trait compared to the four sub-traits used to calculate the index trait. Among the sub-traits, prediction accuracies were higher for bull and heifer period 1 (BP1 and HP1) compared to the traits in period 2 (BP2 and HP2). For all the traits, the various GBLUP models resulted in higher prediction accuracies compared to the PBLUP model. An average gain in reliability of 16 percentage points was obtained using relationships derived from the 54K SNPs compared to using relationships derived from pedigree.

Comparison among the GBLUP models using different SNP sets in one- or two-component models indicates no or only marginal improvements in prediction accuracies compared to using only the 54K data. On average over the five traits, the

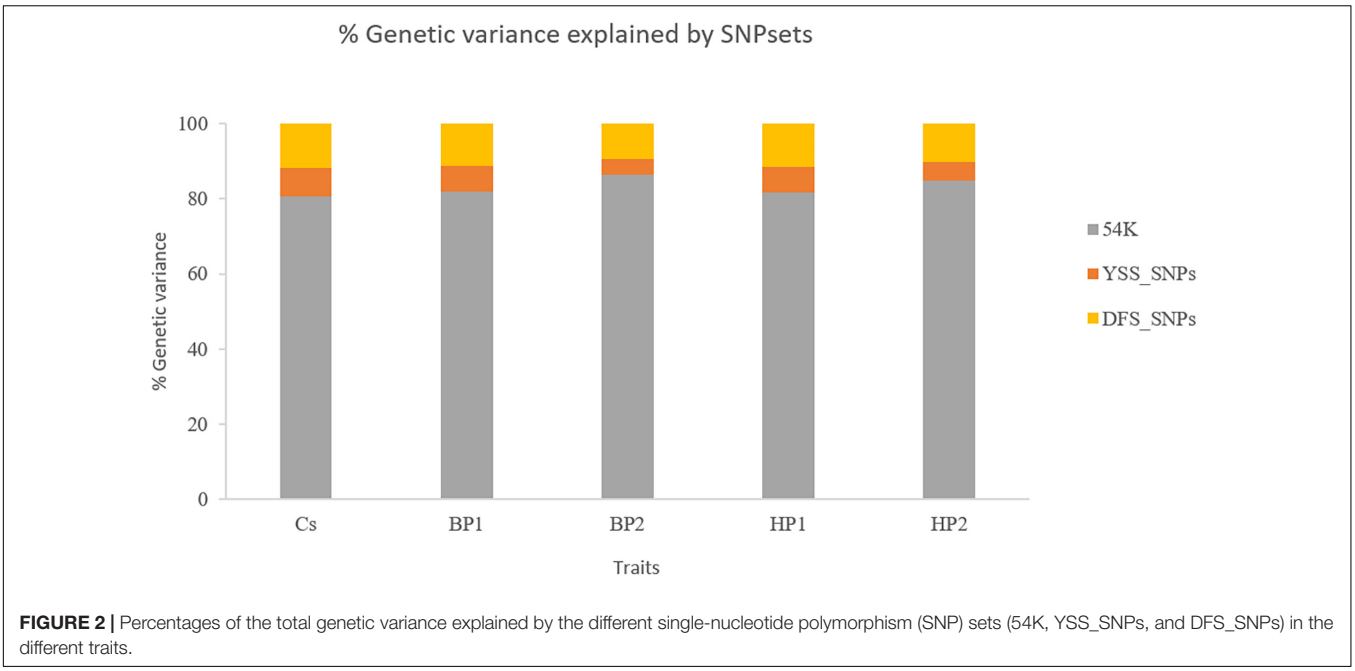

FIGURE 2 | Percentages of the total genetic variance explained by the different single-nucleotide polymorphism (SNP) sets (54K, YSS_SNPs, and DFS_SNPs) in the different traits.

TABLE 2 | Genomic prediction accuracies from PBLUP and GBLUP models.

Trait	PBLUP	GBLUP one-component					GBLUP two-component			
		54K	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET
YSS Index	0.100	0.272	0.274	0.275	0.276	0.276	0.278	0.269	0.271	0.271
BP1	0.236	0.376	0.378	0.379	0.381	0.381	0.388	0.373	0.375	0.375
BP2	0.180	0.332	0.332	0.333	0.334	0.334	0.333	0.330	0.331	0.332
HP1	0.267	0.404	0.406	0.404	0.404	0.404	0.413	0.391	0.393	0.393
HP2	0.140	0.308	0.308	0.307	0.308	0.308	0.309	0.302	0.303	0.303

54K + YSS = Conventional 54K SNPs plus SNPs from GWAS on YSS (YSS_SNPs).
54K + DFS = Conventional 54K SNPs plus SNPs from GWAS on all traits in Nordic total merit index (DFS_SNPs).
54K + YSS + DFS = Conventional 54K SNPs plus YSS_SNPs and DFS_SNPs.
54K + YSS + DFS + LET = Reduced 54K (minus SNPs in recessive lethal haplotypes), plus YSS_SNPs (YSS) and DFS_SNPs and the model considered random regression on genotype code of SNPs in recessive lethal haplotypes (LET_SNPs).*
BP1, Bull period 1; BP2, Bull period 2; GBLUP, genomic best linear unbiased prediction; GWAS, genome-wide association study; HP1, Heifer period 1; HP2, Heifer period 2; PBLUP, pedigree-based best linear unbiased prediction; SNP, single-nucleotide polymorphism; YSS, young stock survival.

improvement in prediction reliability obtained from adding the YSS_SNPs in the one-component model compared to prediction using only the 54K markers was 0.12 percentage points. Similar results were obtained when the 54K marker set was augmented with DFS_SNPs in the one-component model. Fitting both the YSS_SNP sets and DFS_SNPs together with the 54K markers in the one-component model resulted in an average gain in reliability of 0.22 percentage points compared to the prediction using only 54K markers. Additional consideration of random regression on genotype code of recessive lethal alleles in this model did not result in further improvement of prediction reliability. Among the two-component GBLUP models, addition of the YSS_SNPs resulted in an average improvement of 0.58 percentage points compared to the prediction with only 54K SNPs. Addition of the rest of SNP sets (DFS, DFS + YSS) using the two-component GBLUP resulted in slightly lower prediction reliability compared to the model using only 54K.

Table 3 presents the bias in predicting the breeding values across the different models. Regression coefficients were generally close to 1.00 across the different models. Between the different traits, regression coefficient for BP1 and HP1 were generally lower compared to BP2 and HP2 as well as the YSS index trait. For these traits (BP1 and HP1), the one-component GBLUP resulted in slightly less bias compared to the two-component GBLUP model. In addition, model fit for the different scenarios assessed with the AIC is presented in **Table 4**. Generally, the GBLUP models had lower AIC values compared to the PBLUP models across all the traits. Hence, the GBLUP models tend to have better fit to the data compared to the PBLUP models, which is in agreement with the overall performance of the two models in prediction accuracy. Among the different GBLUP

TABLE 3 | Regression coefficients[a] of DRP on prediction.

Trait	PBLUP	GBLUP one-component					GBLUP two-component			
		54K	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET
YSS Index	0.976	1.027	1.026	1.027	1.026	1.022	1.003	1.005	1.000	0.998
BP1	0.976	0.892	0.893	0.891	0.891	0.888	0.891	0.866	0.865	0.863
BP2	1.046	0.953	0.954	0.954	0.955	0.952	0.955	0.954	0.954	0.952
HP1	0.968	0.886	0.887	0.885	0.885	0.883	0.884	0.864	0.863	0.862
HP2	1.045	0.968	0.969	0.967	0.967	0.964	0.965	0.963	0.963	0.960

[a]*Standard errors of regression coefficients across the scenarios = (0.059–0.092).*
54K + YSS = Conventional 54K SNPs plus SNPs from GWAS on young stock survival (YSS_SNPs).
54K + DFS = Conventional 54K SNPs plus SNPs from GWAS on all traits in Nordic total merit index (DFS_SNPs).
54K + YSS + DFS = Conventional 54K SNPs plus YSS_SNPs and DFS_SNPs.
54K + YSS + DFS + LET = Reduced 54K (minus SNPs in recessive lethal haplotypes), plus YSS_SNPs (YSS) and DFS_SNPs and the model considered random regression on genotype code of SNPs in recessive lethal haplotypes (LET_SNPs).*
BP1, Bull period 1; BP2, Bull period 2; DRP, de-regressed proof; GBLUP, genomic best linear unbiased prediction; GWAS, genome-wide association study; HP1, Heifer period 1; HP2, Heifer period 2; PBLUP, pedigree-based best linear unbiased prediction; SNP, single-nucleotide polymorphism; YSS, young stock survival.

TABLE 4 | Akaike information criteria (AIC) for the different models implemented.[a]

Trait	PBLUP	GBLUP one-component					GBLUP two-component			
		54K	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET	54K + YSS	54K + DFS	54K + YSS + DFS	54K* + YSS + DFS + LET
YSS Index	−26.99	−45.18	−45.17	−45.16	−45.16	−45.13	−45.12	−45.14	−45.12	−45.10
BP1	−27.65	−39.86	−39.85	−39.84	−39.84	−39.82	−39.81	−39.82	−39.81	−39.79
BP2	−22.16	−44.98	−44.97	−44.96	−44.96	−44.93	−44.94	−44.95	−44.94	−44.91
HP1	−26.26	−40.75	−40.74	−40.73	−40.72	−40.70	−40.70	−40.70	−40.69	−40.67
HP2	−24.00	−44.96	−44.95	−44.94	−44.94	−44.91	−44.92	−44.93	−44.92	−44.89

[a]*$x10^3$.*
54K + YSS = Conventional 54K SNPs plus SNPs from GWAS on young stock survival (YSS_SNPs).
54K + DFS = Conventional 54K SNPs plus SNPs from GWAS on all traits in Nordic total merit index (DFS_SNPs).
54K + YSS + DFS = Conventional 54K SNPs plus YSS_SNPs and DFS_SNPs.
54K + YSS + DFS + LET = Reduced 54K (minus SNPs in recessive lethal haplotypes), plus YSS_SNPs (YSS) and DFS_SNPs and the model considered random regression on genotype code of SNPs in recessive lethal haplotypes (LET_SNPs).*
BP1, Bull period 1; BP2, Bull period 2; GBLUP, genomic best linear unbiased prediction; GWAS, genome-wide association study; HP1, Heifer period 1; HP2, Heifer period 2; PBLUP, pedigree-based best linear unbiased prediction; SNP, single-nucleotide polymorphism; YSS, young stock survival.

models, the AIC values computed for the different scenarios were quite comparable.

DISCUSSION

Genomic Prediction Accuracies for Young Stock Survival Traits

In general, prediction accuracies for the YSS index trait and the sub-traits were low in our study across scenarios. Our findings are however, in line with reported prediction accuracies in the literature for calf and YSS traits defined in various periods. In a previous study, genomic prediction accuracies ranging between 0.15 and 0.30 were reported for maternal calf survival in different parities for the Canadian Holstein (Abo-Ismail et al., 2017).

Accurate genomic prediction of survival traits in cattle is difficult (van der Heide et al., 2020), as the traits are affected by a combination of environmental factors such as farm management as well as non-additive genetic effects such as recessive lethal gene effects (Gebreyesus et al., 2020).

Across the studied YSS traits, relatively higher prediction accuracies were observed for BP1 and HP1 compared to the YSS index trait and the other two sub-traits. Although the heritability estimates (**Table 1**) for all the traits studied here are among the lowest of the dairy cattle traits (Pedersen et al., 2015), heritability for BP1 and HP1 was even lower compared to the other sub-traits and the index trait. Similarly, DRP reliabilities were slightly lower for BP1 and HP1. Therefore, the slightly higher prediction reliability for BP1 and HP1 was contrary to our expectations. DRP reliability is the function of number of records used to estimate the EBVs and heritability of the traits. Across the studied traits, heritability is quite low and differences in heritability between the traits are small. Therefore, the slight differences in average DRP reliabilities between the studied traits might be due to differences in numbers of observations used to predict the EBVs of the bulls for different traits in the official Nordic cattle evaluations.

Benefits of Incorporation of Selected Variants on Genomic Prediction Reliability

In our study, integration of additional selected WGS SNPs and recessive lethal haplotypes resulted in negligible improvement in genomic prediction reliability for YSS index and the four sub-traits. Previous studies reported some gains in genomic prediction accuracies from additional variants selected from WGS data using GWAS, functional annotation, and pathway analysis, depending on the trait and population studied [e.g., Brøndum et al. (2015), van den Berg et al. (2016), Liu et al. (2019)]. Gains in genomic prediction reliability from integration of additional selected WGS SNPs partly depend on the genetic architecture of the traits and consequently the proportion of variation explained by the selected SNPs (Hayes et al., 2010). In the literature, while additional WGS SNPs improved genomic prediction accuracies for some traits, often marginal improvement is reported for others. Liu et al. (2019) for instance reported increases in prediction accuracies for milk production traits in the Danish Jersey from addition of selected WGS SNPs but lack of improvement in prediction reliability for fertility and only marginal improvement for mastitis. Brøndum et al. (2015) reported increases in prediction reliability of up to 5 percentage points for milk production traits in Nordic Holstein and Red populations, while improvement of reliability was negligible for fertility. Similar results were reported in the study of Veerkamp et al. (2016) where genomic prediction with the addition of a selected set of WGS variants for protein yield (PY), somatic cell score (SCS), and interval from first to last insemination led to negligible improvement in prediction reliability. In the current study, neither of the SNP sets, i.e., DFS_SNPs and YSS_SNPs, led to improvement in prediction reliability of the YSS traits. The DFS_SNPs explained on average 11% of the genomic variance for the studied traits compared to an average of 6% explained by the YSS_SNPs. However, the higher proportion of genomic variance explained by the DFS SNPs in contrast to the YSS SNPs could be merely due to the difference in the number of SNPs in the two sets. The DFS SNPs were selected based on relevance to multiple traits including production, disease, and calving traits. Moreover, the NTM index, which is based on several traits that include the YSS trait, was considered in the selection of the DFS SNPs (Brøndum et al., 2015). However, the main emphasis, in terms of weights, was placed on milk production traits compared to fitness traits such as fertility, mastitis, and other disease traits, as well as the NTM index. On the other hand, the YSS_SNPs reported by Wu et al. (2017) were selected based on GWAS for YSS index specifically; therefore, improvements in prediction reliability were to be expected compared to the DFS SNPs. However, the YSS_SNPs included only 98 SNPs that might make it difficult to explain a sizable proportion of the genetic variation for polygenic traits such as YSS (Wu et al., 2017).

Additionally, the effects of selected variants might be somehow underestimated in this study due to the use DRPs as response variable rather than raw phenotypes for the survival traits. This might specially be of relevant impact to the models that include the effect of recessive lethal alleles rather than those incorporating the selected WGS SNPs, as these were selected based on GWASs using DRPs as response variable (Brøndum et al., 2015; Wu et al., 2017).

One-Component vs. Two-Component Genomic Best Linear Unbiased Prediction Models

It has also been shown that the effect of integrating selected variants on the reliability of genomic prediction might depend on whether or not the effects of these variants have been weighted appropriately in the models (Raymond et al., 2018). In the traditional GBLUP model, the contribution of genetic markers to the genomic relationship is the same. In this context, Sørensen et al. (2014) suggested an extension of the GBLUP model to allow differentiation among the markers through a genomic feature BLUP (GFBLUP) approach. In GFBLUP, variants are categorized according to biological information, such as chromosomes, genes, or biological pathways, so that the random genetic effect in the GBLUP model can have more than one component. Implementation of such an approach to integrate selected variants has shown improvement in genomic prediction reliability compared to integrating them using the traditional one-component GBLUP approach. Gebreyesus et al. (2019) reported substantial increases in genomic prediction reliability in different Holstein cattle populations for milk fatty acid composition traits by incorporating selected variants through the three-component GBLUP model compared to pooling all variants in one GRM. Similar improvements using the two-component GBLUP model were reported in pigs (Sarup et al., 2016; Song et al., 2019).

Contrary to these previous findings, there was no difference in prediction reliability from integrating the selected WGS SNP sets through the two-component model compared to the one-component GBLUP in our study. Multiple-component GBLUP model involves simultaneous estimation of more parameters in addition to those estimated in a one-component model. Thus, gains from multiple-component GBLUP, *vis-à-vis* one-component, can only be expected if addition of information from the additional component(s) is substantial enough to offset the extra uncertainty due to more parameters to be estimated in the multiple-component analysis.

CONCLUSION

In this study, we hypothesize that incorporation of WGS variants selected based on GWAS and variants within recessive lethal haplotypes might improve the reliability of genomic prediction for YSS traits. We tested our hypothesis using one- or two-component GBLUP models. Contrary to our hypothesis, the results showed negligible improvements by incorporation of such variants in genomic prediction accuracies for the YSS

index trait and the four sub-traits. The results highlight the difficulty in genetic evaluation for polygenic traits with very low heritability such as the YSS traits and the need for further studies to explore additional information including the genomic information beyond SNP variants to improve the prediction reliability for these traits.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: Phenotypic and genomic data used in this study are property of the industry partners that contributed to the study. Requests to access these datasets should be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because all procedures to collect the DNA samples followed the protocols approved by the National Guidelines for Animal Experimentation and the Danish Animal Experimental Ethics Committee, and hence, no specific permission was required. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

GG processed the data, implemented the analyses, and drafted the manuscript. GSu conceived the study and contributed to the discussion of the results. ML contributed to the interpretation and discussion of the results. GSa acquired funding and contributed to the discussion of the results. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the GUDP project "LiveCalf" (no. 34009-16-1101) from the Ministry of Environment and Food of Denmark (Copenhagen).

REFERENCES

Abo-Ismail, M. K., Brito, L. F., Miller, S. P., Sargolzaei, M., Grossi, D. A., Moore, S. S., et al. (2017). Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet. Sel. Evol.* 49:82. doi: 10.1186/s12711-017-0356-8

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Boichard, D., Boussaha, M., Capitan, A., Rocha, D., Hozé, C., Sanchez, M. P., et al. (2018). "Experience from large scale use of the Euro-Genomics custom SNP chip in cattle," in *Proceedings of the World Congress on Genetics Applied to Livestock Production, Vol, Molecular Genetics* 4, Auckland, 675.

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., et al. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* 98, 4107–4116. doi: 10.3168/jds.2014-9005

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

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

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Fuerst-Waltl, B., and Sørensen, M. K. (2010). Genetic analysis of calf and heifer losses in Danish Holstein. *J. Dairy Sci.* 93, 5436–5442. doi: 10.3168/jds.2010-3227

Gebreyesus, G., Bovenhuis, H., Lund, M. S., Poulsen, N. A., Sun, D., and Buitenhuis, B. (2019). Reliability of genomic prediction for milk fatty acid composition by using a multi-population reference and incorporating GWAS results. *Genet. Sel. Evol.* 51:16. doi: 10.1186/s12711-019-0460-z

Gebreyesus, G., Sahana, G., Christian Sørensen, A., Lund, M. S., and Su, G. (2020). Novel approach to incorporate information about recessive lethal genes increases the accuracy of genomic prediction for mortality traits. *Heredity* 125, 155–166. doi: 10.1038/s41437-020-0329-5

Hansen, M., Madsen, P., Jensen, J., Pedersen, J., and Christensen, L. G. (2003). Genetic parameters of postnatal mortality in Danish Holstein calves. *J. Dairy Sci.* 86, 1807–1817. doi: 10.3168/jds.s0022-0302(03)73766-7

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

Henderson, L., Miglior, F., Sewalem, A., Kelton, D., Robinson, A., and Leslie, K. E. (2011). Estimation of genetic parameters for measures of calf survival in a population of Holstein heifer calves from a heifer-raising facility in New York State. *J. Dairy Sci.* 94, 461–470. doi: 10.3168/jds.2010-3243

Hoff, J. L., Decker, J. E., Schnabel, R. D., and Taylor, J. F. (2017). Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC Genomics* 18:799. doi: 10.1186/s12864-017-4196-2

Jairath, L., Dekkers, J. C. M., Schaeffer, L. R., Liu, Z., Burnside, E. B., and Kolstad, B. (1998). Genetic evaluation for herd life in Canada. *J. Dairy Sci.* 81, 550–562. doi: 10.3168/jds.s0022-0302(98)75607-3

Liu, A., Lund, M. S., Boichard, D., Karaman, E., Fritz, S., Aamand, G. P., et al. (2019). Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. *Heredity* 124, 37–49. doi: 10.1038/s41437-019-0246-7

Madsen, P., and Jensen, J. (2013). *A user's Guide to DMU. Version 6, release5.2*. Tjele: Aarhus University Foulum.

NAV (Nordic Cattle Genetic Evaluation) (2017). *NAV Routine Genetic Evaluation of Dairy Cattle – Data and Genetic Models*, 4th Edn. Available online at: http://www.nordicebv.info/wp-content/uploads/2017/03/NAV-routine-genetic-evaluation-122016_FINAL.pdf (accessed June 19, 2021).

Østerårs, O., Gjestvang, M. S., Vatn, S., and Sølverød, L. (2007). Perinatal death in production animals in the Nordic countries–incidence and costs. *Acta Vet. Scand.* 49:14.

Pedersen, J., Kargo, M., Fogh, A., Pösö, J., Eriksson, J. A., Nielsen, U. S., et al. (2015). *Note on Economic Value of Young Stock Survival*. 1–11. Available online at: http://www.nordicebv.info/wp-content/uploads/2015/10/Economic-value-of-Young-Stock-Survival.pdf (accessed Feb 17, 2020).

Raymond, B., Bouwman, A. C., Wientjes, Y. C. J., Schrooten, C., Houwing-Duistermaat, J., and Veerkamp, R. F. (2018). Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genet. Sel. Evol.* 50:49.

Sahana, G., Guldbrandtsen, B., Thomsen, B., Holm, L. E., Panitz, F., Brøndum, R. F., et al. (2014). Genome-wide association study using high-density single

nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *J. Dairy Sci.* 97, 7258–7275. doi: 10.3168/jds.2014-8141

Sahana, G., Iso-Touru, T., Wu, X., Nielsen, U. S., de Koning, D. J., Lund, M. S., et al. (2016). A 0.5-Mbp deletion on bovine chromosome 23 is astrong candidate for stillbirth in Nordic Red cattle. *Genet. Sel. Evol.* 48:35.

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

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sørensen, P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet.* 17:11. doi: 10.1186/s12863-015-0322-9

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

Sørensen, P., Edwards, S. M., and Jensen, P. (2014). "Genomic feature models," in *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*, Vancouver, BC.

Strandén, I. (2015). *Command Language Interface for MiX99. Release VIII/2015*. Jokioinen: Natural Resources Institute Finland (Luke),

van den Berg, I., Boichard, D., and Lund, M. S. (2016). Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genet. Sel. Evol.* 48:83.

van der Heide, E. M. M., Veerkamp, R. F., van Pelt, M. L., Kamphuis, C., and Ducro, B. J. (2020). Predicting survival in dairy cattle by combining genomic breeding values and phenotypic information. *J. Dairy Sci.* 103, 556–571. doi: 10.3168/jds.2019-16626

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

VanRaden, P. M., Olson, K. M., Null, D. J., and Hutchison, J. L. (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J. Dairy Sci.* 94, 6153–6161. doi: 10.3168/jds.2011-4624

Veerkamp, R. F., Bouwman, A. C., Schrooten, C., and Calus, M. P. (2016). Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet. Sel. Evol.* 48:95.

Wu, X., Guldbrandtsen, B., Nielsen, U. S., Lund, M. S., and Sahana, G. (2017). Association analysis for young stock survival index with imputed whole-genome sequence variants in Nordic Holstein cattle. *J. Dairy Sci.* 100, 6356–6370. doi: 10.3168/jds.2017-12688

Wu, X., Mesbah-Uddin, M., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2019). Haplotypes responsible for early embryonic lethality detected in Nordic Holsteins. *J. Dairy Sci.* 102, 11116–11123. doi: 10.3168/jds.2019-16651

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

The handling editor declared a past co-authorship with one of the authors ML.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fgene.2021.713575

Genetic and Genomic Analyses of Service Sire Effect on Female Reproductive Traits in Holstein Cattle

Ziwei Chen[1], Luiz F. Brito[2], Hanpeng Luo[1], Rui Shi[1], Yao Chang[1], Lin Liu[3], Gang Guo[4] and Yachun Wang[1]**

[1] Key Laboratory of Animal Genetics, Breeding and Reproduction, MARA, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China, [2] Department of Animal Sciences, Purdue University, West Lafayette, IN, United States, [3] Beijing Dairy Cattle Center, Beijing, China, [4] Beijing Sunlon Livestock Development Company Limited, Beijing, China

OPEN ACCESS

Edited by:
Lingzhao Fang,
The University of Edinburgh,
United Kingdom

Reviewed by:
Lubos Vostry,
Czech University of Life Sciences Prague, Czechia
Chao Ning,
Shandong Agricultural University, China
Jicai Jiang,
North Carolina State University, United States

***Correspondence:**
Gang Guo
guogang2180@126.com
Yachun Wang
wangyachun@cau.edu.cn

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

Received: *23 May 2021*
Accepted: *03 August 2021*
Published: *03 September 2021*

Citation:
Chen Z, Brito LF, Luo H, Shi R, Chang Y, Liu L, Guo G and Wang Y (2021) Genetic and Genomic Analyses of Service Sire Effect on Female Reproductive Traits in Holstein Cattle. Front. Genet. 12:713575. doi: 10.3389/fgene.2021.713575

Fertility and reproductive performance are key drivers of dairy farm profitability. Hence, reproduction traits have been included in a large majority of worldwide dairy cattle selection indexes. The reproductive traits are lowly heritable but can be improved through direct genetic selection. However, most scientific studies and dairy cattle breeding programs have focused solely on the genetic effects of the dam (GED) on reproductive performance and, therefore, ignored the contribution of the service sire in the phenotypic outcomes. This study aimed to investigate the service sire effects on female reproductive traits in Holstein cattle from a genomic perspective. Genetic parameter estimation and genome-wide association studies (GWAS) were performed for the genetic effect of service sire (GESS) on conception rate (CR), 56-day non-return rate (NRR56), calving ease (CE), stillbirth (SB), and gestation length (GL). Our findings indicate that the additive genetic effects of both sire and dam contribute to the phenotypic variance of reproductive traits measured in females (0.0196 vs. 0.0109, 0.0237 vs. 0.0133, 0.0040 vs. 0.0289, 0.0782 vs. 0.0083, and 0.1024 vs. 0.1020 for GESS and GED heritability estimates for CR, NRR56, CE, SB, and GL, respectively), and these two genetic effects are positively correlated for SB (0.1394) and GL (0.7871). Interestingly, the breeding values for GESS on insemination success traits (CR and NRR56) are unfavorably and significantly correlated with some production, health, and type breeding values (ranging from −0.449 to 0.274), while the GESS values on calving traits (CE, SB, and GL) are usually favorably associated with those traits (ranging from −0.493 to 0.313). One hundred sixty-two significant single-nucleotide polymorphisms (SNPs) and their surrounding protein-coding genes were identified as significantly associated with GESS and GED, respectively. Six genes overlapped between GESS and GED for calving traits and 10 genes overlapped between GESS for success traits and calving traits. Our findings indicate the importance of considering the GESS when genetically evaluating the female reproductive traits in Holstein cattle.

Keywords: dairy cattle, genetic evaluation, genome-wide association study, paternal effect, reproductive traits, service sire

INTRODUCTION

In recent decades, the genetic selection for functional traits, including reproductive performance, has received great emphasis in dairy cattle selection indexes, aiming to achieve more balanced and sustainable breeding goals (Egger-Danner et al., 2015). Female fertility has only been broadly included in national selection indexes of dairy cattle breeding programs over the past five decades (Cole and VanRaden, 2018). Although most dairy cattle breeding programs focus only on the genetic effects of the dam (GED) when genetically evaluating fertility and calving traits, there are evidence that the service sire may also have a genetic influence on female reproductive performance (Barton et al., 1984; Van Tassell et al., 2003; Averill et al., 2004)—for instance, the direct effects of the service sire (e.g., semen quality and viability) on female reproductive performance have been considered as indirect effects (Jansen, 1985). Jaton et al. (2017) reported that the heritability estimates of service sire on embryo quality were lower than the donor (0.02 versus 0.04) but still statistically significant. In 2008, a national evaluation model of sire conception rate (SCR) was established in the United Sates by the Animal Improvement Program Laboratory of the United States Department of Agriculture (Norman et al., 2008). SCR is measured as confirmed pregnancy ratio (in percentage) of each service sire. They also implemented a sire–maternal grandsire (S-MGS) model to estimate the genetic component of service sire in calving performance (Van Tassell et al., 2003; Jiang et al., 2018). Even though low heritability estimates have been reported for indirect indicators of male fertility (Berry et al., 2011; Tiezzi et al., 2013), it can still be improved through genomic selection (Lillehammer et al., 2011). Hence, understanding the genetic mechanisms underlying male fertility and developing accurate genomic prediction models are of great importance but still underexplored.

The determination of the genetic effect of service sire (GESS) on reproductive traits relies on genetic and genomic analyses, including the estimation of genetic parameters and genome-wide association studies (GWAS). Previous studies have identified interesting genomic regions associated with SCR, as reviewed by Taylor et al. (2018). However, there are few reports of the GESS on other reproductive traits at the genomic level (Fang et al., 2019; Jiang et al., 2018). In this context, the main objectives of this study were as follows: (Egger-Danner et al., 2015) to investigate the genetic background of service sire on female reproductive performance, including conception rate (CR), 56-day non-return rate (NRR56), calving ease (CE), stillbirth (SB), and gestation length (GL) in Holstein cattle and (Cole and VanRaden, 2018) to identify genomic regions and candidate genes associated with GESS on female reproduction performance.

Abbreviations: GED, genetic effects of the dam; SCR, sire conception rate; GESS, genetic effect of service sire; GWAS, genome-wide association study; CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length; SNP, single-nucleotide polymorphism; MAF, minor allele frequency; FDR, false discovery rate; QTL, quantitative trait locus.

MATERIALS AND METHODS

Phenotypic and Genomic Data

Phenotypes

Records of birth dates, insemination events, pregnancy diagnoses, and calving information collected in 39 farms (Sunlon Livestock Development Co. Ltd., Beijing, China) from 1987 to 2020 were extracted from the AfiFarm software (AfiFarm[1]) and used in this study. The reproductive traits include CR (1 = pregnant, 0 = non-pregnant), NRR56 (1 = non-return, 0 = return), CE (scores from 1 to 3, in which 1 = unassisted, 2 = easy pull, 3 = hard pull and surgery needed), SB (1 = calf was alive 24 h after birth and 2 = calf was dead), and GL (in days). The last insemination before a positive pregnancy diagnosis in each parity was considered as pregnancy. Furthermore, the cows with calving records but without a positive pregnancy diagnosis were considered pregnant when last inseminated before calving. Insemination records that had the next insemination within 1–17 days or had neither insemination nor calving records after that time period were excluded from the NRR56 calculation. CE and SB were directly coded from raw data after excluding ambiguous records (3%) caused by mis-recording. GL records lower than 260 and greater than 302 days were also deleted. A descriptive summary for each trait after data editing is shown in **Table 1**. In total, 163,818 Holstein cows had phenotypes and were serviced by 1,952 bulls. The average (±SD) number of services per bull was 489 ± 1,947. A pedigree file spanning over 13 generations was provided by the Beijing Dairy Cattle Center (BDCC, Beijing, China) and consisted of 503,118 cows and 151,273 bulls born between 1957 and 2020. The estimated breeding values (EBVs) for six production traits (milk yield, milk protein yield, milk protein ratio, milk fat yield, milk fat ratio, and somatic cell score), two health traits (reproductive diseases and udder diseases), and five type traits (dairy character, milking system, capacity, rump, and overall conformation) traits were also provided by BDCC.

Genotypes

A total of 3,477 Holstein bulls were genotyped with the Illumina BovineSNP50 BeadChip (Illumina, Inc., San Diego, CA, United States) containing 54,609 single-nucleotide polymorphism (SNP) markers. These genotypes were imputed to the Illumina 150K Bovine Beadchip (containing 123,268 SNPs) using the BEAGLE v5.1 software (Browning et al., 2018) and a reference population consisting of 3,119 cows and 81 bulls. The SNP information was updated from an older version of the cattle reference genome (UMD 3.1) assembly to the current one (ARS-UCD 1.2) using the UCSC LiftOver tool[2]. Eight thousand five SNPs with missing position in the latest reference genome were removed from further analyses. The reference population was divided into a sub-reference population (2,725 individuals genotyped with the

[1] www.afimilk.com.cn

[2] http://genome.ucsc.edu/cgi-bin/hgLiftOver

TABLE 1 | Descriptive statistics for the phenotype of reproductive traits in Holstein cattle.

Trait	Mean	*SD*	Minimum	Maximum	*N*
CR	0.43	0.49	0	1	837,655
NRR56	0.50	0.50	0	1	857,821
CE	1.06	0.27	1	3	259,042
SB	1.07	0.25	1	2	273,367
GL	278.36	6.18	260	302	258,611

CR, conception rate (0 or 1); NRR56, 56-day non-return rate (0 or 1); CE, calving ease (1, 2, or 3); SB, stillbirth (0 or 1), GL, gestation length (day); SD, standard deviation; N, number of records.

150K SNP panel) and a sub-validation population (475 individuals genotyped with the 150K SNP panel but masked to only the 50K panel SNPs) for assessing the accuracy of genotype imputation as the concordance rate of imputed SNPs. Only SNPs with imputation accuracy greater than 90% were kept for further analyses. Furthermore, SNPs with minor allele frequency lower than 0.05, unknown chromosome or genome position, and extreme deviation from the Hardy–Weinberg equilibrium (p-value lower than 10^{-6}) were removed. After data editing, 109,274 SNPs located in the autosomes and pseudo-autosomal regions of the X chromosome were retained in the dataset.

Estimation of Genetic Parameters

The variance and covariance components for each reproductive trait were estimated using the AI-REML algorithm implemented in the DMUAI module of the DMU v6 software (Madsen et al., 2006). A previous study evaluating similar traits indicated that the heritability estimates of linear and threshold models tend to be similar (Meijering, 1985; Boichard and Manfredi, 1994). Therefore, the following linear mixed model was fitted:

$$y = X\beta + Z_1 u_{\mathrm{m}} + Z_1 u_{\mathrm{f}} + W_1 pe_{\mathrm{m}} + W_2 pe_{\mathrm{f}} + Z_2 hym + e$$

where y represents the vector of phenotypic observations (i.e., CR, NRR56, CE, SB, or GL), β is the vector of fixed effects included in the model, in which different systematic effects were fitted for each trait [i.e., AI technician, parity, semen type, and number of inseminations for CR and NRR56 and calf sex, parity, and group of calf size (divided according to their birth weights: 30–40, 40–50, and 50–60 kg) for CE, SB, and GL]. All of the fixed effects significantly ($P < 0.05$) influenced the dependent variables (CR, NRR56, CE, SB, and SB) and were identified based on mixed model analysis using the PROC MIXED function implemented in the SAS software (version 9.1.3; SAS Institute Inc., Cary, NC, United States); u_{m} and u_{f} are the vectors of the random animal effects accounted by GESS and GED, respectively; pe_{m} and pe_{f} are the vectors of the random permanent environmental effects of the service sire and dam, respectively; hym is the vector of the random herd–year–month effects; e is the vector of the random residual effects; and, X, Z_1, W_1, W_2 and Z_2 are the corresponding incidence matrices. We assumed that:

$$\begin{pmatrix} u_{\mathrm{m}} \\ u_{\mathrm{f}} \\ pe_{\mathrm{m}} \\ pe_{\mathrm{f}} \\ h \\ e \end{pmatrix} \sim N(O, V)$$

with:

$$V = \begin{pmatrix} A \otimes \begin{pmatrix} \sigma^2_{\mathrm{m}} & \sigma_{\mathrm{m,f}} \\ \sigma_{\mathrm{m,f}} & \sigma^2_{\mathrm{f}} \end{pmatrix} & 0 & 0 & 0 & 0 \\ 0 & I \otimes \sigma^2_{pef} & 0 & 0 & 0 \\ 0 & 0 & I \otimes \sigma^2_{pem} & 0 & 0 \\ 0 & 0 & 0 & I \otimes \sigma^2_{hym} & 0 \\ 0 & 0 & 0 & 0 & I \otimes \sigma^2_{\mathrm{e}} \end{pmatrix}$$

where σ^2_{m} and σ^2_{f} are the additive genetic variances of service sire and dam, respectively; $\sigma_{\mathrm{m,f}}$ is the genetic covariance of service sire and dam; σ^2_{pem} and σ^2_{pef} are the permanent environmental variances of service sire and dam, respectively; σ^2_{hym} is the herd–year–month variance; σ^2_{e} is the residual variance; ⊗ is the Kronecker product function; A is the additive genetic relationship matrix among the animals; and I is an identity matrix. For calving traits, the model could be considered as an improved sire–dam model that assumes that the GESS has a genetic covariance with the GED. Compared to the traditional evaluation models of calving traits that only consider the animal effect, the current model includes both service sire and dam effects in the female reproductive performance phenotypes through GESS and GED, respectively. Differently from direct and maternal (paternal) effects usually assumed to be correlated, this was not the case for the genetic components of service sire and dam in the current study. Therefore, the GESS heritability estimates were calculated as follows:

$$h^2_{\mathrm{m}} = \sigma^2_{\mathrm{m}} / \left(\sigma^2_{\mathrm{m}} + \sigma^2_{\mathrm{f}} + 2 \times \sigma_{\mathrm{m,f}} + \sigma^2_{\mathrm{pem}} + \sigma^2_{\mathrm{pef}} + \sigma^2_{\mathrm{hym}} + \sigma^2_{\mathrm{e}}\right)$$

and the repeatability estimates were calculated as follows:

$$re_{\mathrm{m}} = \left(\sigma^2_{\mathrm{m}} + \sigma^2_{\mathrm{pem}}\right) / \left(\sigma^2_{\mathrm{m}} + \sigma^2_{\mathrm{f}} + 2 \times \sigma_{\mathrm{m,f}} + \sigma^2_{\mathrm{pem}} + \sigma^2_{\mathrm{pef}} + \sigma^2_{\mathrm{hym}} + \sigma^2_{\mathrm{e}}\right)$$

The GED repeatability was estimated in the same way but replacing σ^2_{m} and σ^2_{pem} in the numerator by $\sigma^2_{\mathrm{f}} + \sigma^2_{\mathrm{pef}}$.

The standard error of the heritability and repeatability estimates, respectively, were calculated using the Delta method (Su et al., 2007). A Wald test was carried out to determine the statistical difference between the genetic parameter estimates and zero. In addition, correlations between the genomic breeding

values for GESS of the reproduction traits, as well as production, health, and type traits, were estimated using the method proposed by Calo et al. (1973) and bull EBVs (filtered based on EBV reliability, as described below). Standard errors (SE) of the approximate genetic correlations were calculated based on the formula proposed by Sokal and Rohlf (1981). The predicted transmitting ability (PTA) of six production traits (milk yield, milk protein yield, milk protein ratio, milk fat yield, milk fat ratio, and somatic cell score), two health traits (reproductive disease and udder disease), and five type traits (dairy character, milking system, capacity, rump, and overall conformation) with reliabilities higher than 20, 20, and 30%, respectively, were provided by BDCC. The statistical models used for the genetic evaluation of these traits are described in Miglior et al. (2009) and Wu et al. (2013). Individuals with PTA reliabilities for GESS on reproductive traits above 40% were used for the calculation of correlation of breeding values.

Genome-Wide Association Study and Functional Enrichment Analyses

The "Fixed and random model Circulating Probability Unification" (FarmCPU) R package (Liu X. et al., 2016) was used to perform single-SNP regression analyses. FarmCPU is a multi-locus model that incorporates multiple markers simultaneously as covariates to partially remove the confounding effect between testing markers and kinship (Liu X. et al., 2016). De-regressed proofs (DRP) of the GESS and GED for female reproductive traits were derived following the procedures suggested by Wiggans et al. (2011). Individuals with accuracies greater than 10% were used as dependent variables in the GWAS model. The obtained *p*-values were corrected for multiple testing by calculating the false discovery rate (FDR) (Benjamini et al., 2001) at the 5% genome-wise level. Quantile–quantile (Q–Q) plots and the inflation factor λ (Devlin and Roeder, 1999) were used to investigate population stratification by comparing the observed and expected distributions of –log(*p*-value).

Positional genes located at up to 200 kb upstream and downstream of the significant SNPs were identified using the BiomaRt package (Durinck et al., 2005). This 200-kb window was defined based on the linkage disequilibrium level of the studied population (**Supplementary Figure 1**). The ClueGO module in Cytoscape (Bindea et al., 2009) was used to identify candidate genes, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms. Furthermore, the Cattle Quantitative Trait Locus (QTL) database (Cattle QTLdb Release 42[3]) was used to identify important trait–QTL associations previously reported in the literature.

RESULTS AND DISCUSSION

Descriptive Statistics of Phenotype

The descriptive statistics of the five reproductive traits evaluated are shown in **Table 1**. For the insemination success traits (CR and NRR56), the mean CR was 43 ± 49% and the mean NRR56 was 50 ± 50% in Chinese Holstein cattle. These estimates are lower than those reported by Guo et al. (2014), which were inferred from the number of services recorded from 2001 to 2010. The average CE, SB, and GL were 1.06 ± 0.27, 1.07 ± 0.25, and 278.36 ± 6.18 days, respectively. The proportion of CE scores higher than 1 was 5.21%, and the SB rate was 6.67%. Vieira-Neto et al. (2017) reported a mean GL of 276 ± 6 days in Holstein cows, which is slightly lower than that of the current study. GL is affected by various environmental effects, such as temperature (McClintock et al., 2003) and cow parity number (King et al., 1985), which may have led to the discrepancy of GL between different populations.

Genetic Parameters for Female Reproductive Traits Considering GESS

Heritability and Repeatability Estimates

The (co)variance components, heritability, and genetic correlations for GESS and GED for all five reproduction traits are shown in **Table 2**. Overall, the heritability estimates for the success traits were low, but the heritability of GESS were significantly higher than those of GED (0.020 ± 0.004 vs. 0.011 ± 0.000 for CR and 0.024 ± 0.005 vs. 0.013 ± 0.001 for NRR56), indicating that service sires actually have greater genetic impact on insemination success than the mating cows. Fertility traits are known to have low heritability (VanRaden et al., 2004), which makes it more difficult to obtain faster genetic progress compared to more heritable traits. Therefore, larger datasets and novel phenotypes (Fleming et al., 2019) should be generated for increasing genetic progress for fertility performance. The small, but significant, genetic contribution of service sires on success traits indicates the possibility of genetically improving the GESS. The repeatability estimates of these reproduction traits were almost equal to their heritabilities, except for the GED on NRR56 (0.031 ± 0.001 vs. 0.013 ± 0.001), indicating the inconsistency among these records. Hence, more repeated records are needed for greater EBV accuracies. Another reason for the low repeatability might be the ignorance of non-additive genetic effects, which could lead to the underestimation of genetic parameters. For most traits, the genetic variance of the GESS was comparable to that of the GED, demonstrating that service sires have considerable genetic contribution to female reproductive outcomes. The heritability estimates for female CR in other studies range from 0.01 to 0.03 (Azzam et al., 1988; Bagnato and Oltenacu, 1993; Muuttoranta et al., 2019). These low estimates support the estimates of GED on CR in this study. The heritability of GED on NRR56 was consistent with the value (0.01) reported by Sun and Su (2010) but different from that of Hoekstra et al. (1994) (0.04) and Eghbalsaied (2011) (0.002–0.003). These discrepancies may be attributed to the differences in the populations evaluated and the statistical models used since GESS was only included in the current study. Some studies have considered the GESS as a non-genetic random effect apart from the GED, and the heritabilities obtained from such models were shown to be lower than 0.02 for CR and NRR56 (Weigel and Rekaya, 2000; Kuhn and Hutchison, 2008). However, Tiezzi et al. (2011) suggested that the heritabilities

[3]www.animalgenome.org/cgi-bin/QTLdb/BT/index

generated using the approach mentioned above are lower than those fitting GESS. Furthermore, Tiezzi et al. (2013) reported that the heritability and repeatability of the GESS on NRR56 were both 0.01 for Italian Brown Swiss cattle. The characteristics of the two populations evaluated might have caused the discrepancy in the estimates observed. In the study of Tiezzi et al. (2013),

TABLE 2 | Genetic components and parameters for reproductive traits in Holstein cattle.

Parameter	CR	NRR56	CE	SB	GL
σ^2_{ss}	4.50×10^{-3}	5.41×10^{-3}	3.06×10^{-4}	1.96×10^{-3}	3.9844
$\sigma_{ss,d}$	1.47×10^{-4}	-3.80×10^{-4}	8.16×10^{-5}	8.90×10^{-5}	3.1303
σ^2_d	2.49×10^{-3}	3.03×10^{-3}	2.24×10^{-3}	2.08×10^{-4}	3.9699
σ^2_{pess}	4.13×10^{-8}	2.68×10^{-4}	2.22×10^{-9}	1.76×10^{-7}	1.16×10^{-6}
σ^2_{ped}	2.75×10^{-7}	4.09×10^{-3}	4.65×10^{-7}	3.09×10^{-4}	0.5436
σ^2_{hys}	0.0126	0.0152	0.0228	8.59×10^{-3}	3.2778
σ^2_e	0.2099	0.2003	0.0521	0.0140	27.1262
h^2_{ss}	0.0196	0.0237	0.0040	0.0782	0.1024
$SE_{h^2_{ss}}$	4.10×10^{-3}	4.50×10^{-3}	0.0013	0.0101	8.70×10^{-3}
re_{ss}	0.0196	0.0249	0.0040	0.0782	0.1024
$SE_{re_{ss}}$	5.30×10^{-3}	0.0058	0.0017	0.0131	0.0109
h^2_d	0.0109	0.0133	0.0289	0.0083	0.1020
$SE_{h^2_d}$	8.00×10^{-4}	9.00×10^{-4}	0.0019	9.00×10^{-4}	3.40×10^{-3}
re_d	0.0109	0.0312	0.0289	0.0207	0.1160
SE_{re_d}	1.10×10^{-3}	0.0012	0.0029	1.80×10^{-3}	0.0043
r	0.0438	−0.0937	0.0986	0.1394	0.7871
SE_r	0.0662	0.0627	0.0811	0.0705	0.0402

σ^2_{ss}, additive genetic variance of service sire; $\sigma_{ss,d}$, genetic covariance of service sire and dam; σ^2_d, additive genetic variance of dam; σ^2_{pess}, permanent environment variance of service sire; σ^2_{ped}, permanent environment variance of dam; σ^2_{hym}, herd-year-month variance; σ^2_e, residual variance; h^2_{ss}, heritability of service sire; $SE_{h^2_{ss}}$, standard error of heritability of service sire; re_{ss}, repeatability of service sire; $SE_{re_{ss}}$, standard error of repeatability of service sire; h^2_d, heritability of dam; $SE_{h^2_d}$, standard error of heritability of dam; re_d, repeatability of dam; SE_{re_d}, standard error of repeatability of dam; r, genetic correlation between genetic effect of service sire (GESS) and genetic effects of the dam (GED); SE_r, standard error of genetic correlation between GESS and GED; CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length.

TABLE 3 | The correlations of breeding values for the genetic effects of service sire of reproductive traits and production, health, and type traits in Holstein cattle.

Trait[a]		CR	NRR56	CE	SB	GL
Production	Milk yield	−0.218 (0.039)	−0.334 (0.036)	−0.044 (0.034)	−0.191 (0.058)	−0.277 (0.031)
	Milk protein yield	−0.262 (0.038)	−0.402 (0.035)	−0.035 (0.034)	−0.254 (0.057)	−0.348 (0.030)
	Milk protein ratio[b]	−0.150 (0.039)	−0.188 (0.037)	0.041 (0.034)	−0.199 (0.058)	−0.250 (0.031)
	Milk fat yield	−0.277 (0.041)	−0.449 (0.037)	−0.055 (0.039)	−0.256 (0.058)	−0.300 (0.034)
	Milk fat ratio[b]	−0.048 (0.043)	−0.132 (0.041)	0.005 (0.039)	−0.135 (0.060)	0.004 (0.035)
	Somatic cell score	−0.020 (0.049)	0.152 (0.048)	0.059 (0.047)	0.313 (0.060)	0.086 (0.042)
Health	Reproductive disease	0.174 (0.043)	0.203 (0.042)	0.067 (0.042)	0.219 (0.054)	0.127 (0.037)
	Udder disease	0.188 (0.043)	0.274 (0.041)	−0.124 (0.041)	−0.004 (0.055)	0.277 (0.036)
Type	Dairy character	−0.006 (0.055)	−0.126 (0.054)	−0.003 (0.053)	−0.022 (0.065)	−0.067 (0.048)
	Milking system	−0.076 (0.047)	−0.334 (0.043)	0.130 (0.044)	−0.003 (0.059)	−0.399 (0.038)
	Capacity	0.038 (0.046)	−0.075 (0.045)	−0.061 (0.044)	−0.043 (0.059)	−0.262 (0.039)
	Rump	−0.099 (0.047)	−0.283 (0.045)	−0.001 (0.046)	−0.155 (0.059)	−0.493 (0.036)
	Overall conformation	−0.013 (0.048)	−0.236 (0.046)	0.093 (0.046)	−0.018 (0.060)	−0.386 (0.039)

CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length.

[a]The EBVs of 400–500 individuals with reliability greater than 40% for reproductive traits and 20% for other traits were used for calculating the approximate correlations.

[b]The EBV of milk fat ratio and milk protein ratio, respectively, come from the formulas below:

$$EBV_{fat_{ratio}} = \frac{EBV_{fat_{yield}} \times 100 - EBV_{milk_{yield}} \times fat_{ratio}}{EBV_{milk_{yield}} + milk_{yield}}$$

$$EBV_{protein_{ratio}} = \frac{EBV_{protein_{yield}} \times 100 - EBV_{milk_{yield}} \times protein_{ratio}}{EBV_{milk_{yield}} + milk_{yield}}$$

where $EBV_{fat_{ratio}}$, $EBV_{fat_{yield}}$, $EBV_{milk_{yield}}$, $EBV_{protein_{ratio}}$, and $EBV_{protein_{yield}}$ represent the EBV of milk fat ratio, milk fat yield, milk yield, milk protein ratio, and milk protein yield, respectively; fat_{ratio}, $protein_{ratio}$, and $milk_{yield}$ represent the average milk fat ratio, milk protein ratio, and milk yield of cows in their second lactation.

TABLE 4 | Significant single-nucleotide polymorphism (SNP) and nearby genes of the genetic effects of service sire on five reproductive traits carried out by genome-wide association studies in Holstein cattle.

Trait	SNP	BTA	Position (bp)	*P*-value	MAF	Effect	FDR	Genes[a]
CR	rs29011049	1	97,448,665	1.21×10^{-6}	0.38	0.006	0.018	***LRRC34***, *ACTRT3*, *MYNN*, *PHC3, SAMD7*, *SEC62*, *GPR160*
	rs133215257	3	79,272,169	6.96×10^{-6}	0.26	−0.006	0.040	*PDE4B*, *MGC137454*
	rs43370565	3	115,520,121	3.81×10^{-6}	0.10	0.005	0.026	*ASB18*, *GBX2*, *AGAP1*
	rs42882387	4	89,316,398	1.63×10^{-6}	0.24	0.004	0.018	*POT1*
	rs137141507	6	3,362,250	1.32×10^{-6}	0.09	0.006	0.018	*EXOSC9*, *CCNA2*, *BBS7*, *ANXA5*
	rs42487799	7	37,625,102	1.32×10^{-9}	0.32	0.005	<0.001	*COMMD10*, *SEMA6A*
	rs109487947	7	54,827,253	1.07×10^{-6}	0.08	−0.005	0.018	
	rs110233258	7	58,460,338	3.50×10^{-8}	0.15	−0.005	0.001	*STK32A*, ***PPP2R2B***, *DPYSL3*
	rs109461455	9	37,700,129	1.92×10^{-6}	0.15	−0.003	0.018	
	rs42864672	9	63,416,140	4.60×10^{-8}	0.37	−0.004	0.001	
	rs136069526	9	67,796,936	1.04×10^{-5}	0.48	−0.003	0.050	*ARHGAP18*
	rs109859987	11	53,890,637	9.83×10^{-6}	0.49	−0.002	0.050	
	rs109992118	13	21,561,365	1.97×10^{-6}	0.21	−0.003	0.018	*PLXDC2*
	rs109632400	13	58,602,292	1.73×10^{-6}	0.35	−0.003	0.018	***PCK1***, *RBM38*, *RAE1*, *ZBP1*, ***CTCFL***, *PMEPA1*
	rs41751511	15	8,668,231	4.63×10^{-7}	0.49	0.003	0.010	*CNTN5*
	rs43713533	15	28,453,282	2.87×10^{-6}	0.29	−0.003	0.022	*TMPRSS13*, *FXYD2*, *FXYD6*
	rs42402130	16	32,852,344	2.86×10^{-8}	0.23	0.005	0.001	*ADSS2*, *C16H1orf100*
	rs41750173	16	57,750,820	4.74×10^{-6}	0.18	−0.004	0.029	*PAPPA2*
	rs29019796	17	25,645,239	4.84×10^{-6}	0.05	0.007	0.029	
	rs110578750	18	2,907,599	1.05×10^{-5}	0.14	−0.004	0.050	*GABARAPL2*, *CFDP2*, *CFDP1*, *TMEM170A*, *TMEM231*, *ADAT1*, *KARS1*, *TERF2IP*, *BCNT2*, *CHST6*
	rs110228250	20	40,015,842	3.52×10^{-6}	0.05	−0.006	0.026	*ADAMTS12*, *SLC45A2*, ***RXFP3***
	rs41568642	21	12,166,884	2.40×10^{-6}	0.17	0.004	0.020	
	rs110340891	22	37,375,611	9.41×10^{-6}	0.19	0.003	0.050	*ATXN7*, *PSMD6*, *PRICKLE2*
NRR56	rs43587839	1	155,197,244	1.91×10^{-7}	0.09	−0.003	0.003	
	rs134652568	2	74,743,823	4.21×10^{-6}	0.15	0.003	0.024	
	rs110190075	2	103,750,400	7.13×10^{-7}	0.28	0.005	0.009	
	rs43354413	3	92,138,083	7.02×10^{-6}	0.30	0.002	0.036	*HSPB11*, *TMEM59*, *TCEANC2*, *MRPL37*, *SSBP3*, *LRRC42*, *CDCP2*, *CYB5RL*, *LDLRAD1*, *YIPF1*
	rs43371984	3	116,894,699	9.45×10^{-6}	0.42	−0.004	0.043	*MLPH*, *LRRFIP1*, *PRLH*, *RAB17*, *COL6A3*
	rs43386888	4	49,509,741	2.41×10^{-6}	0.12	−0.003	0.021	*NRCAM*, *NME8*, *PNPLA8*
	rs42766762	6	108,074,935	3.04×10^{-6}	0.37	−0.005	0.021	
	rs110658771	8	19,488,876	3.48×10^{-6}	0.17	0.003	0.022	*IZUMO3*
	rs109087862	8	63,729,696	7.17×10^{-8}	0.46	−0.003	0.003	*ANKS6*, *GALNT12*, *GABBR2*
	rs110409952	9	87,035,097	2.88×10^{-6}	0.39	0.002	0.021	*NUP43*, *PCMT1*, *LATS1*, *ULBP17*, *LRP11*, *KATNA1*, *ULBP21*
	rs133014180	10	66,613,339	1.56×10^{-7}	0.11	0.004	0.003	***BMP4***
	rs41667346	11	23,497,697	5.49×10^{-6}	0.33	−0.003	0.030	
	rs110387293	11	73,919,337	3.11×10^{-7}	0.21	0.003	0.004	*POMC*, *DTNB*, *DNMT3A*
	rs41633184	13	13,837,771	4.25×10^{-6}	0.09	0.005	0.024	
	rs43709749	14	28,478,462	1.08×10^{-6}	0.44	0.002	0.011	
	rs41616446	15	13,400,749	9.97×10^{-9}	0.26	−0.004	0.001	
	rs110815341	15	63,461,312	7.26×10^{-6}	0.32	−0.003	0.036	*EIF3M*, *QSER1*, *PRRG4*, *DEPDC7*
	rs42427669	17	66,079,314	3.04×10^{-7}	0.07	0.003	0.004	*SEZ6L*, *TPST2*, *SRRD*, *TFIP11*, *HPS4*, *CRYBB1*, *CRYBA4*, *ASPHD2*
	rs133424642	18	13,506,620	1.02×10^{-6}	0.49	−0.003	0.011	*SLC7A5*, *CA5A*, *BANP*
	rs110863925	24	7,065,050	1.59×10^{-7}	0.45	−0.003	0.003	*RTTN*, *SOCS6*
	rs135757150	24	54,887,194	3.08×10^{-6}	0.45	0.003	0.021	*TCF4*
	rs109715869	25	25,154,075	7.91×10^{-6}	0.44	0.002	0.037	*GSG1L*, *GTF3C1*, *KATNIP*, *IL21R*
	rs136986771	30	133,828,729	2.61×10^{-6}	0.14	0.003	0.021	
	rs134555078	30	137,676,697	3.10×10^{-8}	0.36	−0.004	0.002	
CE	rs29016910	2	41,127,106	4.83×10^{-7}	0.42	0.001	0.007	*KCNJ3*
	rs43715311	3	114,365,299	5.28×10^{-6}	0.44	0.001	0.034	*SH3BP4*
	rs43427376	5	7,184,325	1.41×10^{-6}	0.28	0.001	0.015	*NAV3*
	rs133310180	5	90,520,626	2.21×10^{-6}	0.35	0.001	0.019	*AEBP2*, *PLEKHA5*
	rs133412722	7	68,637,147	7.25×10^{-6}	0.48	0.001	0.044	*HAVCR2*, *MED7*
	rs110471321	9	7,979,325	2.22×10^{-7}	0.24	−0.001	0.005	*ADGRB3*
	rs108984322	10	8,009,471	2.40×10^{-6}	0.18	−0.001	0.019	*IQGAP2*, ***F2RL2***, *F2R*, *S100Z*, ***CRHBP***, *AGGF1*, ***F2RL1***
	rs29017584	10	33,364,541	2.50×10^{-6}	0.49	0.002	0.019	
	rs42568446	11	18,452,325	5.33×10^{-7}	0.25	0.001	0.007	
	rs42583510	11	30,305,738	2.78×10^{-6}	0.22	0.002	0.020	*MSH6*, *FBXO11*

(Continued)

TABLE 4 | Continued

Trait	SNP	BTA	Position (bp)	*P*-value	MAF	Effect	FDR	Genes[a]
	rs41628019	14	28,529,843	3.23×10^{-7}	0.37	0.001	0.006	
	rs109928489	14	81,472,215	5.62×10^{-7}	0.41	0.001	0.007	*COL14A1*, *DSCC1*, *DEPTOR*
	rs41824124	16	70,244,076	3.87×10^{-6}	0.19	0.001	0.026	*RPS6KC1*
	rs41635371	16	78,603,871	1.52×10^{-6}	0.40	−0.001	0.015	
	rs42813960	18	64,728,827	1.32×10^{-7}	0.35	−0.001	0.004	*ZNF550*, *ZNF419*, *ZNF548*
	rs136257872	19	8,172,407	3.10×10^{-10}	0.37	0.002	<0.001	*MSI2*
	rs108970271	22	26,765,132	5.86×10^{-8}	0.30	−0.001	0.002	
	rs42070292	25	29,175,853	8.31×10^{-10}	0.12	0.002	<0.001	*GALNT17*, *CALN1*
SB	rs41594258	1	34,978,818	1.06×10^{-6}	0.09	−0.004	0.019	*VGLL3*
	rs109309140	2	4,981,764	1.42×10^{-6}	0.46	0.002	0.022	*PROC*, *SFT2D3*, *WDR33*, *GPR17*, ***LIMS2***, *IWS1*, ***MYO7B***
	rs137754398	2	101,961,755	2.82×10^{-9}	0.40	−0.005	<0.001	
	rs43353437	3	83,295,309	5.81×10^{-6}	0.25	−0.003	0.039	*KANK4*, *DOCK7*, *USP1*
	rs43386788	4	27,127,668	3.98×10^{-6}	0.09	−0.005	0.036	*HDAC9*
	rs43424011	5	7,320,087	6.04×10^{-6}	0.31	−0.003	0.039	*NAV3*
	rs137802315	6	92,046,444	3.27×10^{-6}	0.45	0.003	0.036	*CCNI*, *SEPTIN11*, ***CCNG2***
	rs41630520	9	19,561,418	6.43×10^{-6}	0.23	−0.002	0.039	*TTK*, *ELOVL4*
	rs109624175	9	56,241,273	4.99×10^{-8}	0.24	0.003	0.001	
	rs109920124	11	44,744,209	4.96×10^{-6}	0.38	−0.003	0.039	*RANBP2*, *CCDC138*, *SULT1C3*, ***LIMS1***, *GCC2*, *EDAR*, *SULT1C4*, *SULT1C2*
	rs135416382	12	24,302,230	3.62×10^{-6}	0.27	−0.003	0.036	*TRPC4*, *POSTN*
	rs134002839	13	71,259,389	3.58×10^{-8}	0.18	−0.004	0.001	*PTPRT*
	rs135791311	14	71,330,255	1.44×10^{-9}	0.09	0.006	<0.001	*TRIQK*
	rs43168517	16	10,976,232	5.53×10^{-6}	0.10	0.004	0.039	
	rs108992403	19	16,143,462	4.99×10^{-6}	0.20	−0.002	0.039	*ASIC2*
	rs134228482	22	13,014,212	3.13×10^{-6}	0.41	−0.002	0.036	*MYRIP*, *EIF1B*
	rs41588424	28	12,833,134	3.37×10^{-8}	0.17	−0.004	0.001	*CHRM3*
	rs42492371	29	1,952,181	1.88×10^{-6}	0.50	−0.003	0.026	*FAT3*, *MTNR1B*
GL	rs42630203	1	18,608,496	1.18×10^{-6}	0.30	−0.184	0.018	*CHODL*, *TMPRSS15*
	rs110604162	1	112,217,694	2.99×10^{-6}	0.42	0.162	0.027	***PLCH1***
	rs43271952	1	134,240,312	4.60×10^{-6}	0.30	0.174	0.033	*EPHB1*
	rs43354413	3	92,138,083	2.94×10^{-6}	0.30	0.179	0.027	*HSPB11*, *TMEM59*, *TCEANC2*, *MRPL37*, ***SSBP3***, *LRRC42*, *CDCP2*, *CYB5RL*, *LDLRAD1*, *YIPF1*
	rs134191168	4	66,013,467	6.18×10^{-8}	0.36	0.175	0.002	*ZNRF2*, *GGCT*, *NOD1*, *MTURN*
	rs108993952	7	48,852,799	1.00×10^{-6}	0.39	−0.203	0.018	*SPOCK1*
	rs136460053	7	61,263,431	4.91×10^{-6}	0.48	0.207	0.033	*HMGXB3*, *CSF1R*, *PPARGC1B*, *PDE6A*, *SLC26A2*
	rs109807989	8	88,740,000	1.92×10^{-6}	0.28	−0.254	0.021	***CKS2***, *SECISBP2*, ***SEMA4D***, *SHC3*
	rs109727604	14	39,463,590	3.57×10^{-6}	0.43	−0.155	0.028	
	rs136577145	16	30,791,818	1.73×10^{-6}	0.37	−0.202	0.021	*TFB2M*, *CNST*, *SCCPDH*, *H3-5*, *SMYD3*
	rs41619483	18	6,140,925	3.40×10^{-6}	0.29	0.152	0.028	***WWOX***
	rs110402487	18	64,444,876	2.27×10^{-8}	0.29	−0.228	0.001	*AURKC*, *ZNF304*, *ZNF805*, *ZNF548*, *ZIM3*
	rs109173977	21	41,990,869	3.80×10^{-7}	0.35	−0.328	0.010	*GPR33*, *HEATR5A*, *NUBPL*
	rs110148531	23	39,148,251	1.09×10^{-6}	0.19	−0.175	0.018	*RNF144B*
	rs137469593	27	19,456,244	1.74×10^{-6}	0.37	−0.170	0.021	*PDGFRL*, *PCM1*, *FGL1*, ***MTUS1***
	rs135655219	28	5,171,795	1.79×10^{-8}	0.26	0.236	0.001	*SIPA1L2*
	rs42178394	29	33,061,591	7.27×10^{-6}	0.17	−0.168	0.047	*JAM3*, *IGSF9B*, *SPATA19*

CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length; MAF, minor allele frequency.
[a]*The important candidate genes for each trait are shown in bold face.*

a smaller and older population than that in the current study was used, and the cattle in their population performed better in NRR56 (0.70 for average), indicating a better management and genetic level of their herds.

For calving traits, the heritability estimates of GESS are lower than the GED for CE (0.004 ± 0.001 vs. 0.029 ± 0.002) and similar for GL (0.102 ± 0.001 vs. 0.102 ± 0.000) but higher for SB (0.078 ± 0.010 vs. 0.008 ± 0.001). Both GESS and GED show

TABLE 5 | Significant single-nucleotide polymorphism (SNP) and near-by genes of the genetic effects of dam on five reproductive traits carried out by genome-wide association studies in Holstein cattle.

Trait	SNP	BTA	Position (bp)	*P*-value	MAF	Effect	FDR	Genes[a]
CR	rs110588394	1	157,046,373	8.55×10^{-9}	0.19	0.008	0.001	*PP2D1*, *EFHB*, *RAB5A*
	rs135745940	2	86,206,807	5.58×10^{-8}	0.23	0.006	0.002	*MARS2*, *COQ10B*, *PLCL1*, *RFTN2*, *HSPD1*, *HSPE1*, *MOB4*, *BOLL*
	rs41625668	3	2,131,858	7.13×10^{-7}	0.44	0.005	0.010	*TADA1*, *POGK, ILDR2*
	rs110992111	3	63,079,015	1.70×10^{-6}	0.45	−0.004	0.017	*ADGRL2*
	rs41657989	7	24,218,169	1.10×10^{-7}	0.40	−0.005	0.003	*CHSY3*, *ADAMTS19*, *MINAR2*
	rs41592654	9	32,237,919	1.38×10^{-7}	0.21	−0.005	0.003	*ASF1A*, *MCM9*, *CEP85L*, *PLN*, *FAM184A*
	rs42557707	12	45,131,104	5.35×10^{-7}	0.08	−0.008	0.008	
	rs110113735	12	75,616,202	2.53×10^{-6}	0.17	0.004	0.021	*FARP1, SLC15A1*, *DOCK9, STK24*
	rs137128437	14	73,073,273	5.01×10^{-7}	0.41	−0.005	0.008	*SLC26A7*
	rs43713566	15	29,560,389	9.42×10^{-7}	0.20	0.006	0.011	*BCL9L*, *VPS11*, *HMBS*, *DPAGT1*, *TRAPPC4*, *SLC37A4*, *DDX6*, *HYOU1*, *ABCG4*, *UPK2*, *FOXR1*, *C2CD2L*, *HINFP*, *CENATAC*, *RPS25*, *NLRX1*, *CXCR5*
	rs136206713	15	81,820,622	2.82×10^{-8}	0.41	0.006	0.002	*CNTF*, *LPXN*
	rs132777210	21	69,731,178	2.35×10^{-6}	0.34	0.005	0.021	*BTBD6*, *BRF1*, *TMEM121*, *PACS2*, *TEDC1*, *CRIP1*, *NUDT14*
	rs109776480	22	57,040,932	1.13×10^{-6}	0.43	−0.005	0.012	*MRPS25*, *RBSN*, *SYN2*, *TIMP4*
NRR56	rs42827552	6	23,199,236	3.74×10^{-6}	0.35	−0.005	0.041	*BANK1*
	rs41575824	9	53,639,845	3.48×10^{-6}	0.12	0.007	0.041	*FUT9*
	rs136204465	14	9,422,558	2.62×10^{-7}	0.15	−0.005	0.005	
	rs134979761	17	9,110,531	2.75×10^{-8}	0.41	0.006	0.002	
	rs109533406	17	57,716,043	2.82×10^{-7}	0.39	0.006	0.005	*NOS1*, *FBXO21*, *KSR2*
	rs135974611	18	31,152,142	2.34×10^{-6}	0.27	−0.006	0.032	
	rs41606596	18	45,936,015	1.35×10^{-7}	0.07	−0.006	0.004	*FAM187B*, *GRAMD1A*, *SCN1B*, *HPN*, *FFAR2*, *CD22*, *FFAR3*, *LSR*, *USF2*, *HAMP*, *MAG*, *LGI4*, *FXYD1*, *FXYD7*, *FXYD5*, *FFAR1*
	rs110401168	19	57,629,224	3.60×10^{-9}	0.35	0.008	<0.001	
	rs42428874	21	36,656,375	4.58×10^{-6}	0.19	−0.005	0.042	*NOVA1*
	rs136876790	22	34,075,088	4.90×10^{-8}	0.47	0.005	0.002	*SUCLG2*
	rs110534364	24	50,494,672	4.31×10^{-6}	0.11	0.007	0.042	*ELAC1*, ***SMAD4***, *MEX3C*, *ME2*, *MRO*
	rs109783875	26	27,091,833	1.71×10^{-6}	0.36	−0.004	0.027	
CE	rs110752117	7	51,418,311	1.21×10^{-6}	0.08	−0.007	0.019	*PURA*, *NRG2*, ***CYSTM1***, ***HBEGF***, *SLC4A9*, *IGIP*, *PFDN1*
	rs110761813	10	16,708,640	1.63×10^{-6}	0.37	−0.005	0.022	
	rs109942798	12	18,661,493	1.07×10^{-6}	0.08	0.005	0.019	*MLNR*, *FNDC3A*
	rs109156982	13	47,716,129	6.75×10^{-8}	0.28	0.005	0.002	*GPCPD1*, *PROKR2*
	rs110115548	13	67,003,397	3.93×10^{-6}	0.20	0.005	0.048	*TTI1*, *VSTM2L*, *RPRD1B*, *BPI*, ***TGM2***, *KIAA1755*
	rs109493014	15	72,384,357	1.12×10^{-9}	0.43	0.005	<0.001	
	rs41578821	16	30,855,245	9.26×10^{-8}	0.10	−0.005	0.003	***TFB2M***, *CNST*, *H3-5*, ***SMYD3***
	rs42949634	18	39,542,279	4.32×10^{-7}	0.25	0.004	0.009	*CALB2*, *TAT*, *AP1G1*, *MARVELD3*, *PHLPP2*, *CHST4*, *ZNF19*, *ZNF23*
	rs42534666	26	5,084,073	6.45×10^{-9}	0.14	0.008	<0.001	*PCDH15*
SB	rs135087719	4	69,282,296	4.09×10^{-7}	0.18	−0.001	0.004	*SKAP2*
	rs135473218	7	95,890,433	8.29×10^{-7}	0.23	0.001	0.008	*CAST*, *PCSK1*
	rs43546352	8	31,779,503	1.34×10^{-10}	0.25	−0.002	<0.001	*TYRP1*
	rs134655277	9	4,579,294	1.88×10^{-6}	0.45	0.001	0.015	
	rs135323642	9	67,192,772	2.79×10^{-6}	0.38	0.002	0.020	*LAMA2*
	rs41595401	10	38,651,279	9.51×10^{-7}	0.33	0.001	0.008	
	rs110003547	11	47,431,717	4.84×10^{-9}	0.07	0.001	<0.001	*EIF2AK3*, *RPIA*, *TEX37*, *FOXI3*
	rs42337856	11	58,828,979	2.61×10^{-7}	0.27	−0.001	0.003	
	rs133162533	15	3,989,919	5.61×10^{-6}	0.48	0.001	0.036	
	rs41790653	16	9,685,690	1.47×10^{-7}	0.47	0.001	0.003	
	rs133390427	17	60,309,111	3.61×10^{-6}	0.33	0.001	0.025	*TBX5*
	rs110008365	18	5,593,593	2.74×10^{-7}	0.28	−0.001	0.003	***WWOX***
	rs109395549	20	523,765	4.10×10^{-7}	0.42	0.001	0.004	*PANK3*, *SLIT3*
	rs110695662	21	44,838,064	1.22×10^{-9}	0.17	0.001	<0.001	*EAPP*, *SNX6*, *SPTSSA*
	rs42566616	22	40,247,823	1.65×10^{-10}	0.36	0.001	<0.001	

(Continued)

TABLE 5 | Continued

Trait	SNP	BTA	Position (bp)	*P*-value	MAF	Effect	FDR	Genes[a]
	rs137494875	22	47,178,548	1.88×10^{-7}	0.21	−0.001	0.003	*CACNA1D*, *CHDH*, *IL17RB*, *ACTR8*, *SELENOK*
	rs109564594	26	1,419,659	8.41×10^{-8}	0.20	−0.001	0.002	
GL	rs43747887	1	88,621,323	2.52×10^{-6}	0.49	0.183	0.031	
	rs135396670	2	130,043,555	2.25×10^{-6}	0.32	−0.172	0.031	*C1QA*, *C1QC*, ***C1QB***, *EPHA8*, *LACTBL1*, ***EPHB2***, *TEX46*
	rs42760413	6	90,933,892	5.91×10^{-7}	0.17	−0.168	0.013	***CXCL10***, *CXCL11*, *SDAD1*, *ART3*, *NUP54*, *SCARB2*, *PPEF2*, *NAAA*, *CXCL9*
	rs43482393	6	93,053,914	1.09×10^{-9}	0.34	−0.349	<0.001	*FRAS1*
	rs109176316	7	52,159,740	1.67×10^{-6}	0.31	−0.175	0.026	***PCDHA13***, *PCDHB1*, *PCDHAC2*
	rs136804356	14	28,876,699	4.75×10^{-6}	0.33	0.158	0.047	
	rs136577145	16	30,791,818	3.97×10^{-10}	0.39	−0.244	<0.001	*TFB2M*, *CNST*, *SCCPDH*, *H3-5*, *SMYD3*
	rs109102279	20	19,138,824	6.01×10^{-8}	0.12	0.250	0.002	
	rs42406702	20	55,193,435	3.00×10^{-6}	0.12	−0.229	0.033	
	rs110072536	25	40,344,012	8.43×10^{-7}	0.31	−0.170	0.015	*CARD11*
	rs109960049	26	17,360,622	4.05×10^{-8}	0.48	−0.326	0.001	*ENTPD1*, *ZNF518A*, *BLNK*, *CCNJ*

CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length; MAF, minor allele frequency.
[a]*The important candidate genes for each trait are shown in bold face.*

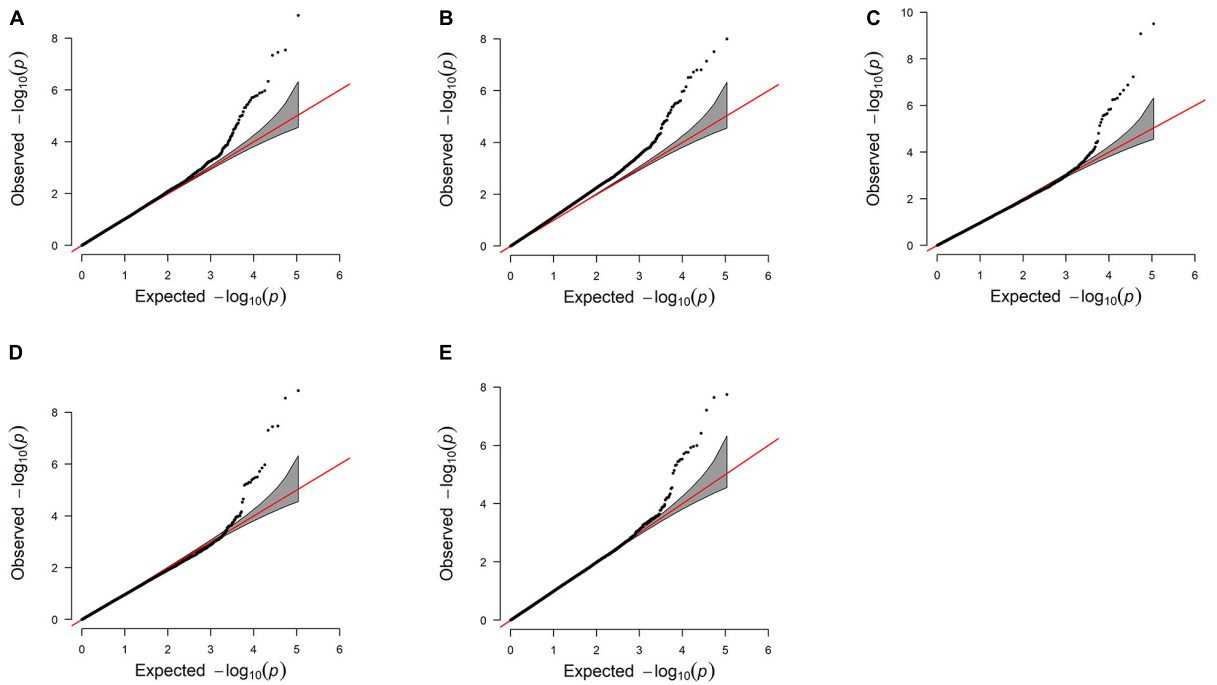

FIGURE 1 | Quantile–quantile plots of the genome-wide association studies for genetic effect of service sire on reproductive traits **(A–E)**. The *x*-axis and the *y*-axis represent the expected and observed $-\log_{10}(P-\text{value})$, respectively. **(A)** Conception rate—the λ value is 1.06. **(B)** 56-day non-return rate—the λ value is 1.15. **(C)** Calving ease—the λ value is 0.96. **(D)** Stillbirth—the λ value is 0.95. **(E)** Gestation length—the λ value is 0.99.

low heritability for SB and CE, though a moderate estimation was observed for GL. Except for the GED for SB (0.021 ± 0.002) and GL (0.116 ± 0.004), the repeatabilities were almost equal to the corresponding heritabilities. Each cow has only one calving record in each parity, which results in poor consistency among repeated data. The low heritability estimates of SB and CE indicate that these traits might benefit more from genomic information. A S-MGS model was used to evaluate calving traits in previous studies, which resulted in low heritability for CE and SB (0.07–0.13) (Van Tassell et al., 2003; Heringstad et al., 2007). The S-MGS model converted the direct and maternal variances to sire and maternal grandsire variances, which was, to some extent, different from those of the present study. Due to the differences in the model used in the current study, here we simply compare our findings with the results of the S-MGS model from previous studies. The maternal heritabilities estimated by other models ranged from 0.01 to 0.13 for GL (Jamrozik et al., 2005; Crews, 2006; Mujibi and Crews, 2009) and 0.02–0.08 for

CE and SB (Luo et al., 2002; Jamrozik et al., 2005; Eaglen et al., 2013). The GED variance obtained in the current study for SB was 0.0002, which is much lower than the estimates from a previous study with Norwegian Red cows (Heringstad et al., 2007). The heritability of the GED was higher than that of the GESS for CE, which may have been caused by a larger maternal effect (Jamrozik et al., 2005)—for instance, cow body conformation is genetically associated with calving difficulty (Dadati et al., 1985; Ga et al., 2021), indicating the crucial role of GED on CE.

Genetic Correlations

Former genomic analyses reported that some genes (e.g., *SPP1*) regulate both tissue and embryonic growth (Weintraub et al., 2004; Rangaswami et al., 2006), which is important for both the male and female aspects of calving traits. Furthermore, some genes related to spermatogenesis have been shown to affect cow reproduction performance (Peddinti et al., 2010; Li et al., 2012b; Buzanskas et al., 2017). Thus, the genetic covariances of male and female were considered. The genetic correlations between the GESS and GED were significantly different from zero for SB and GL ($P < 0.05$; based on a *t*-test). The correlation coefficient was considerably high in GL, whereas that in CE was low and non-significant ($P > 0.05$; CE: 0.099 ± 0.081; SB: 0.139 ± 0.071; GL: 0.787 ± 0.040). The results indicate the homogeneity between the GESS and the GED for SB and GL. The genetic covariances of the insemination traits between the effects of service bull and dams were usually ignored in previous studies (Berry et al., 2011). We evaluated these covariances because there are evidence of low but statistically significant correlations between GESS and GED [e.g., NRR56: 0.010 ± 0.002; (Tiezzi et al., 2013)]. However, there was no significant correlation between the two terms in the current study (CR: 0.044 ± 0.066 and NRR56: -0.094 ± 0.063). The genetic correlations between the direct and maternal effects for SB and CE can be quite variable, ranging from −0.24 to 0.12 (Luo et al., 1999; Steinbock et al., 2003; Wiggans et al., 2003; Cole et al., 2007; Heringstad et al., 2007; Vanderick et al., 2014). For GL, the correlations are usually negative and stronger ($-0.13 \sim -0.85$) (Cubas et al., 1991; Bennett and Gregory, 2001; Hansen et al., 2004; Crews, 2006; Cervantes et al., 2010). These discrepancies are reasonable, because we considered both the direct effect of dam and maternal effect as dam effect.

The correlations of breeding values of the GESS of reproductive traits as well as production, health, and type traits are shown in **Table 3**. For all the traits, the genetic information of approximately 400–500 individuals with reliability greater than 40% for reproductive traits and 20% for the other traits was used for the calculation of breeding value correlations. Interestingly, the GESS was negatively correlated with most production and type traits (e.g., milk yield and overall conformation), while positive correlations were observed between GESS and health traits such as reproductive disorders. We found that the GESS on CR was unfavorably and significantly related to milk yield (-0.218 ± 0.039), indicating that selection exclusively on milk production might indirectly result in a decline of insemination performance of the service sire. Murray et al. (1977) reported a negative correlation between male fertility and milk yield (−0.26), which is in contrast to the positive correlation (0.13–0.29) reported by Raheja et al. (1989). Further verification about the biological correlation between male fertility and milk

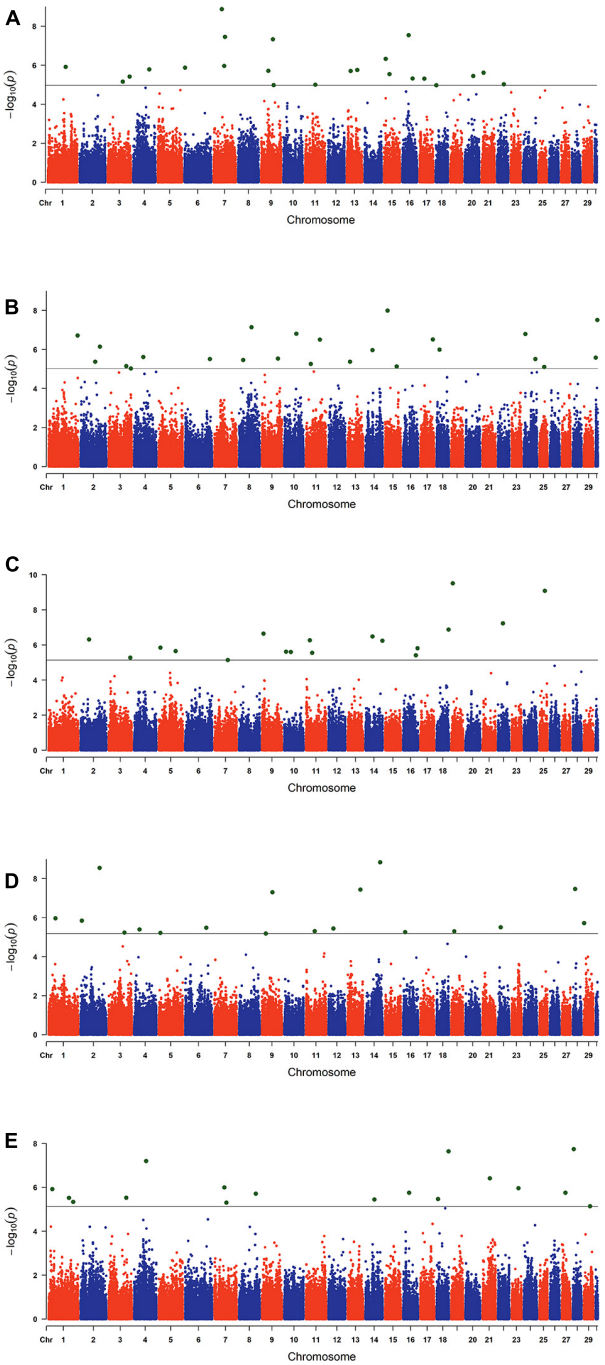

FIGURE 2 | Manhattan plots of the genome-wide association studies for genetic effect of service sire on reproductive traits. The *x*-axis and the *y*-axis represent the chromosome number and the observed $-\log_{10}(P-\text{value})$, respectively. The single-nucleotide polymorphisms were plotted against their genomic positions. The lines in the plots indicate the thresholds of false discovery rate (0.05) in the corresponding traits: **(A)** conception rate, **(B)** 56-day non-return rate, **(C)** calving ease, **(D)** stillbirth, and **(E)** gestation length.

production is needed due to the inconsistent relationships observed. The genetic correlations between GESS on success traits and reproductive disease were positive (0.174 ± 0.043 for CR and 0.203 ± 0.042 for NRR56). Progesterone is regarded as a responsible factor for cattle ovarian follicular cysts (Silvia et al., 2002). Ramal-Sanchez et al. (2020) also suggested that progesterone is related to sperm release, which affects the fertilization ability of spermatozoa. Therefore, reproduction-related hormones might account for the potential biological relation between GESS on success traits and reproductive disease. Cows with a higher incidence of ovarian cysts tend to have lower fertility. The genetic correlation between GESS on GL and overall conformation was negative, indicating that undesirable body conformation might lead to longer GL, with worse development after birth (Bourdon and Brinks, 1982). These correlations indicate that direct selection for production, health, and type traits may have a favorable effect on service sire calving performance but may lead to an unfavorable decline in service sire mating performance.

Genome-Wide Association Studies

The estimated breeding values obtained from former genetic estimation with accuracy above 10% were used for DRP calculation. Hence, the EBV of 2,996 and 1,147 individuals were used in GWAS for GESS and GED, respectively. The detailed information of the genomic regions and candidate genes for GESS and GED for the five reproductive traits are summarized in **Tables 4**, **5**. In addition, the Q–Q plots and Manhattan plots are provided in **Figures 1**–**4**. The Q–Q plots and λ of GESS on 56-day non-return rate indicated a slight inflation of the results. A total of 162 significant SNPs were detected, including two significant SNPs located in the X chromosome (pseudo-autosomal region) (Johnson et al., 2019). The *P*-values ranged from 1.34×10^{-10} to 1.05×10^{-5}, and FDR ranged from <0.001 to 0.050. Furthermore, we mapped the significant SNPs to the Cattle QTL database (see Text Footnote 3), and the overlapped QTL regions are listed in **Tables 6**, **7**.

For GESS, 100 SNPs were found to be significant, with 23, 24, 18, 18, and 17 associated with CR, NRR56, CE, SB, and GL, respectively (**Table 4**). For the success traits, 53 nearby annotated protein-coding genes located 200 kb upstream and downstream of the significant SNPs were mapped, including genes related to sperm development (e.g., *BMP4*) and early embryogenesis (e.g., *LRRC34*). Han and Peñagaricano (2016) also identified a genomic region (1.5 Mb) located on BTA13 which explained more than 0.50% of the total additive genetic variance for SCR (male fertility). This region contains the *CTCFL* gene detected in the present study. Although previous studies have attempted to identify candidate genes related to GESS on success traits (Taylor et al., 2018), novel candidate genes were identified in this study—for instance, the gene *BMP4* (bone morphogenetic protein 4) was previously indicated as a candidate gene for spermatogenesis (Hu et al., 2004; Li et al., 2014) and follicle development (Nilsson and Skinner, 2003; Shimizu et al., 2004; Fatehi et al., 2005; Li and Ge, 2011). These associations indicate the paternal and maternal effects on the pre-implantation stage

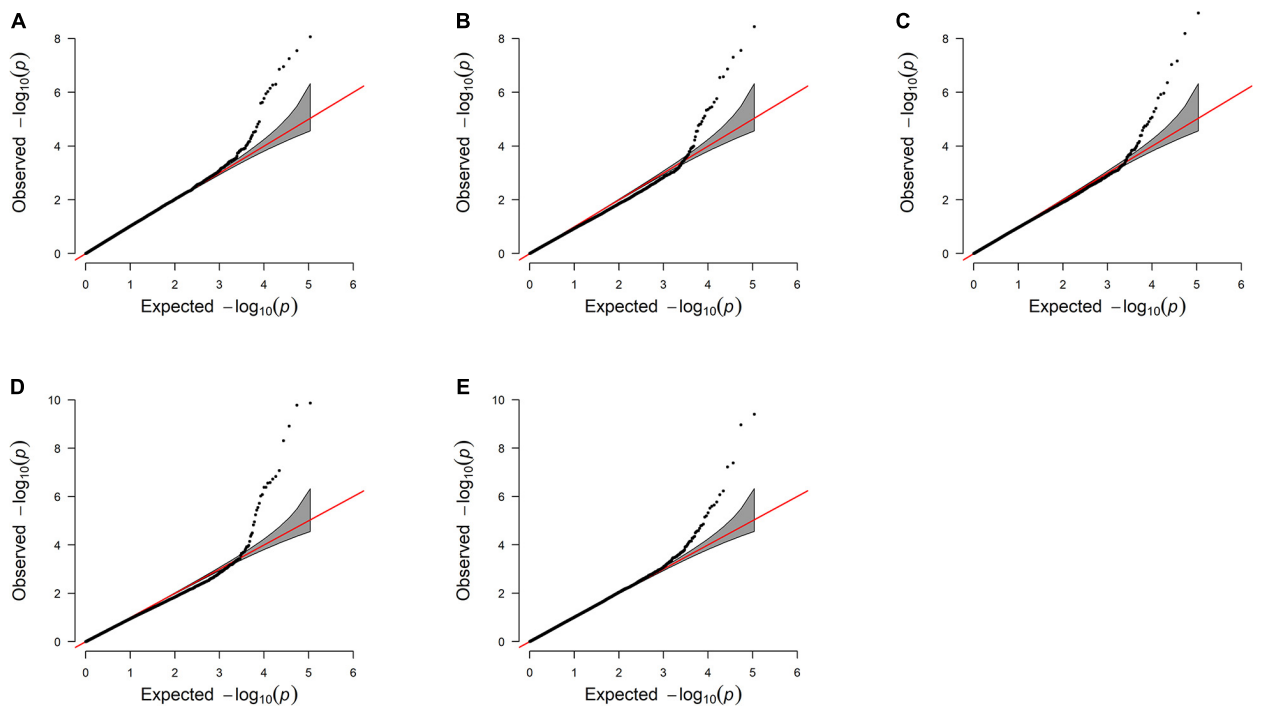

FIGURE 3 | Quantile–quantile plots of genome-wide association studies for genetic effects of the dam on reproductive traits **(A–E)**. The *x*-axis and the *y*-axis represent the expected and observed $-\log_{10}(P-\text{value})$. **(A)** Conception rate—the λ value is 1.03. **(B)** 56-day non-return rate—the λ value is 0.92. **(C)** Calving ease—the λ value is 0.96. **(D)** Stillbirth—the λ value is 0.94. **(E)** Gestation length—the λ value 1.00.

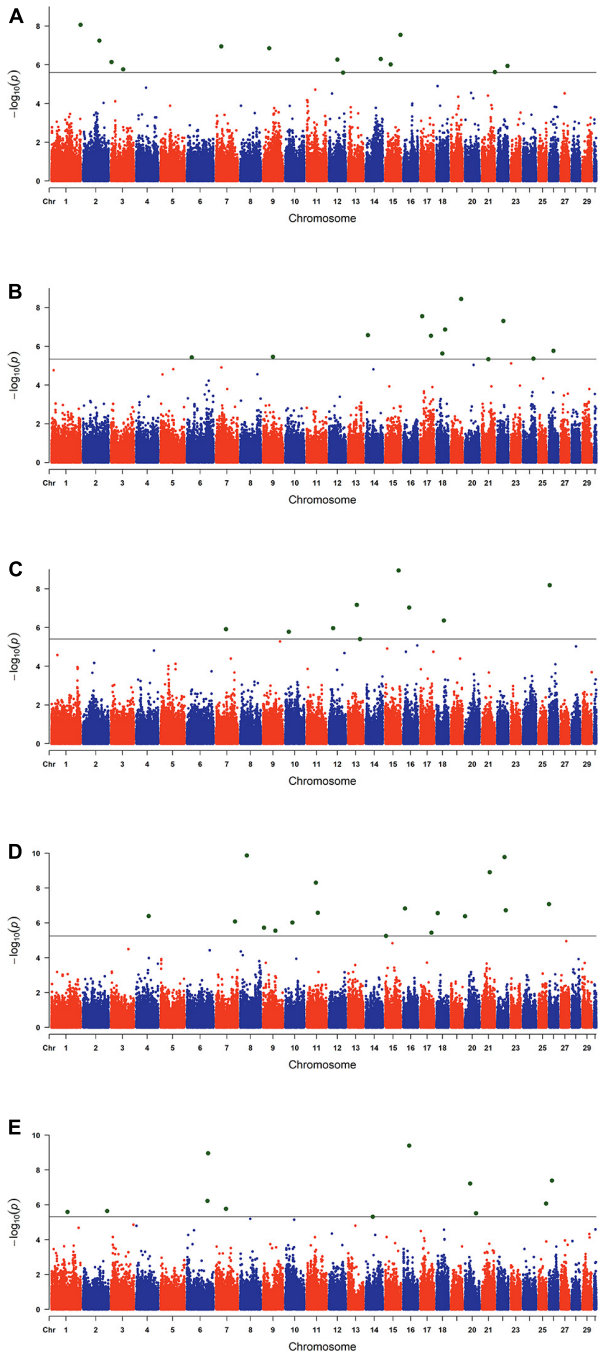

FIGURE 4 | Manhattan plots of the genome-wide association studies for the genetic effects of the dam on reproductive traits. The *x*-axis and the *y*-axis represent the chromosome number and the observed $-\log_{10}(P-\text{value})$, respectively. The single-nucleotide polymorphisms were plotted against their genomic positions. The lines in the plots indicate the thresholds of false discovery rate (0.05) in the corresponding traits: **(A)** conception rate, **(B)** 56-day non-return rate, **(C)** calving ease, **(D)** stillbirth, and **(E)** gestation length.

of embryo development, respectively (Koide et al., 2009; Li et al., 2012a). In this context, the present study also revealed the GESS on female conception. Ou et al. (2020) carried out a transcriptome analysis on spermatogonial stem cells (SSCs) and reported that *LRRC34* was highly expressed in mouse SSCs and was essential for *in vitro* SSC proliferation. *RXFP3* was identified to be differentially expressed in human sperm and was likely diminished in spermiogenesis (Heidari et al., 2018). Furthermore, other studies suggested *PPP2R2B* and *PCK1* as candidate genes affecting the semen quality traits in livestock (Huang et al., 2016; Gao et al., 2019).

One hundred fourteen nearby proteinase genes located 200 kb upstream and downstream of the significant SNPs related to the GESS of calving traits were mapped. The fetal growth and metabolism show a specific pattern during pregnancy (Bell et al., 1993) and supposedly reflect on calving traits. Some potential genes related to calving traits in our study were previously associated with carcass and meat quality traits, including *MTUS1* (Albrecht et al., 2016), *PLCH1* (Lemos et al., 2016), *F2RL1* (Srikanth et al., 2020; Zhang et al., 2017), *MYO7B* (Doyle et al., 2020; Jia et al., 2020), *WWOX* (Grigoletto et al., 2020), *TFB2M* (Jiang et al., 2006; Song et al., 2019), and *SMYD3* (De Vos, 2018). Furthermore, *F2RL1* was reported to affect the body size in Chinese Holstein cattle (Zhang et al., 2017). Some other genes were identified as essential genes for embryogenesis [*SEMA4D* (Masuda et al., 2004), and *CCNG2* (Ma et al., 2015) and *CKS2* (Martinsson-Ahlzén et al., 2008), which contribute to subsequent fetal development]. Perkins et al. (1995) collected human fetal plasma samples both at mid-gestation and parturient to explore the trend of corticotrophin-releasing hormone-binding protein (CRHBP) during the different gestation stages and reported CRHBP as being functional in both maternal and fetal circulation. *F2RL2*, *LIMS2*, and *LIMS1* were indicated by Forde et al. (2012) as pregnancy-associated genes that are differently expressed in the endometrium of cattle during early pregnancy, indicating the potential role of these genes on successful pregnancy establishment. The SNP *rs43354413* located in BTA 3 was identified as significant for GESS on NRR56 and GL, with 10 protein-coding genes in close proximity—for instance, the *SSBP3* gene (located approximately 131 kb upstream of this marker) was reported to regulate mouse embryonic stem cell differentiation to trophoblast-like cells (Liu J. et al., 2016). These findings suggest the possible function of *SSPB3* of GESS on reproduction performance. We used Chinese Holstein population and re-defined the genetic components of calving traits in current study, but some SNPs (*rs42813960* and *rs110402487*) in our study are consistent with previous genomic studies of calving traits focused on BTA18 (Müller et al., 2017), indicating the potential importance of this chromosome in calving performance. Fang et al. (2019) proposed *ZNF613* as a candidate gene for paternal contributions to GL, and in our population, we detected one GL-related marker located downstream of *ZNF613* (*rs110402487*). Furthermore, we also mapped these SNPs to the Animal QTL Database (see Text Footnote 3) and subsequently found that some of them were located in QTLs associated with related traits (**Table 6**). Particularly, *rs108993952* was related to GESS on GL and meanwhile located in the QTL related to some maternal calving traits, which is a promising candidate marker for calving performance.

Sixty-two SNPs were found to be significant for GED, with 13, 12, 9, 17, and 11 being associated with CR, NRR56, CE, SB, and GL, respectively (**Table 5**). One hundred sixty-seven protein-coding genes were found within 200 kb of those SNPs. *rs136577145* was significant for both GESS and GED on GL and was about 63 kb upstream of *rs41578821*, which was detected as significant for GED on CE. The nearby genes mapped with *rs136577145* were *TFB2M*, *CNST*, *SCCPDH*, *H3-5*, and *SMYD3*. *TFB2M*, a mitochondrial transcription specificity factor, as well as *SMYD3*, a histone lysine methyltransferase, were previously linked to bone and skeletal muscle tissue development (Norrbom et al., 2010; Fujii et al., 2011; Sun et al., 2015)—that is, both paternal- and maternal-derived effects on GL were coincident with growth ability. In addition, *WWOX* overlapped between GESS and GED on other calving traits, though no gene overlapped between GESS and GED on success traits (**Figure 5**). This is consistent with the genetic correlation estimates observed (**Table 2**). The low overlap between GESS and GED on success traits demonstrates that different genes are involved in the insemination outcomes from the dam and service sire contributions.

TABLE 6 | Significant single-nucleotide polymorphisms (SNPs) of the genetic effects of service sire and overlapping quantitative trait loci (QTLs).

Trait	SNP	BTA	QTL[a]
CR	rs109461455	9	Muscle anserine content QTL (151506)
	rs109632400	13	Interval to first estrus after calving QTL (14769)
	rs43713533	15	Body weight (yearling) QTL (68806)
	rs110228250	20	Milk protein percentage QTL (105842)
NRR56	rs41616446	15	Body weight (yearling) QTL (68771)
	rs42427669	17	Conception rate QTL (177207)
CE	rs43715311	3	Milking speed QTL (157395)
	rs43427376	5	Milk unglycosylated kappa-casein percentage QTL (119033)
	rs108984322	10	Hip height QTL (131441); udder depth QTL (135447)
SB	rs41594258	1	Milk protein yield QTL (26122)
	rs137802315	6	Milk unglycosylated kappa-casein percentage QTL (118721); milk kappa-casein percentage QTL (111024)
GL	rs108993952	7	Milking speed QTL (157567); body depth QTL (43024); calving ease (maternal) QTL (43025); daughter pregnancy rate QTL (43026); foot angle QTL (43027); milk fat percentage QTL (43028); PTA type QTL (43029); udder attachment QTL (43030); milk fat yield QTL (43031); net merit QTL (43032); length of productive life QTL (43033); rump width QTL (43034); calving ease QTL (43035); somatic cell score QTL (43036); stillbirth QTL (43037); stature QTL (43038); strength QTL (43039); udder depth QTL (43040)

CE, calving ease; SB, stillbirth; GL, gestation length.
[a]The QTL mapping was based on the Cattle QTL Database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index), and those QTLs that include significant SNPs are shown in the table.

TABLE 7 | Significant single-nucleotide polymorphisms (SNPs) of the genetic effects of dam and overlapping quantitative trait loci (QTLs).

Trait	SNP	BTA	QTL[a]
CR	rs41657989	7	Milking speed QTL (157466); clinical mastitis QTL (19014)
NRR56	rs109533406	17	Muscle potassium content QTL (152009)
CE	rs110115548	13	Milk yield QTL (16180); milk protein yield QTL (16181); milk fat percentage QTL (16182); milk protein percentage QTL (16183)
	rs41578821	16	Body weight (slaughter) QTL (102172)
SB	rs41595401	10	Body weight (yearling) QTL (68167); body weight gain QTL (68168); body depth QTL (44657); dairy form QTL (44658); feet and leg conformation QTL (44659); PTA type QTL (44660); teat placement—front QTL (44661); udder attachment QTL (44662); net merit QTL (44663); teat placement—rear QTL (44664); udder height QTL (44665); rump width QTL (44666); somatic cell score QTL (44667); stature QTL (44668); strength QTL (44669); udder cleft QTL (44670); udder depth QTL (44671)
	rs110003547	11	Age at puberty QTL (21140)
	rs42337856	11	Interval to first estrus after calving QTL (28582)
	rs109564594	26	Calving ease (maternal) QTL (52571); dairy form QTL (52572); daughter pregnancy rate QTL (52573); milk fat percentage QTL (52574); milk fat yield QTL (52575); net merit QTL (52576); length of productive life QTL (52577); milk protein percentage QTL (52578); milk protein yield QTL (52579); rump angle QTL (52580); rear leg placement—side view QTL (52581); teat placement—rear QTL (52582); calving ease QTL (52583); teat length QTL (52584); udder cleft QTL (52585)

CR, conception rate; NRR56, 56-day non-return rate; CE, calving ease; SB, stillbirth; GL, gestation length.
[a]The QTL mapping was based on the Cattle QTL Database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index), and those QTLs that include significant SNPs are shown in the table.

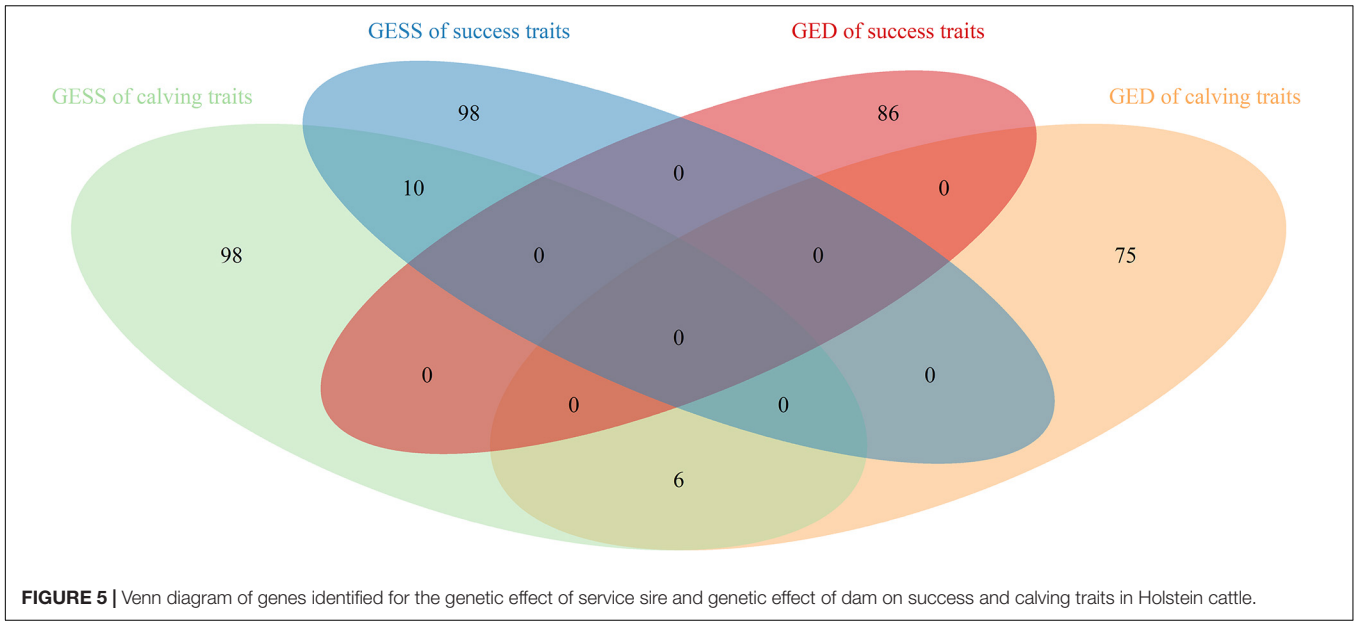

FIGURE 5 | Venn diagram of genes identified for the genetic effect of service sire and genetic effect of dam on success and calving traits in Holstein cattle.

Except for the genes mentioned before, the genomic results of GED are consistent with previous studies (e.g., *SMAD4*). For GED on success traits, 86 nearby protein-coding genes located 200 kb upstream and downstream of significant SNPs were mapped. Particularly, *SMAD4* has been reported to be linked to embryogenesis and folliculogenesis in mice (Chu et al., 2004; Pangas et al., 2006). Lee et al. (2014) reported *SMAD4* to be involved in early embryonic development in cattle and to regulate the effects of follistatin, which influences the environment-independent part of the interval from the first to the last insemination in cattle (Zhang et al., 2019). Eighty-one protein-coding genes were mapped for GED on calving traits and previously considered as essential genes for the maternal mechanism during pregnancy, including *CXCL10* (Walker et al., 2012), *HBEGF* (Jessmon et al., 2009), *PCDHA13* (Lotfan et al., 2018), *TGM2* (Purfield et al., 2019), *C1QB* (Cochran et al., 2013), *CYSTM1* (Purfield et al., 2019), and *EPHB2* (Purfield et al., 2015), which were also identified as candidate genes for maternal effect on calving traits in dairy and beef cattle. We also observed that some significant SNPs located in QTLs are associated with related traits (**Table 7**)—for instance, *rs42427669*, which was related to GED on NRR56, is located in a QTL region associated with CR (Kiser et al., 2019), while the GL-related marker *rs108993952* is located in QTLs associated with CE and SB (Cole et al., 2011).

Functional Enrichment Analyses

The enriched GO terms and KEGG pathways that passed the criteria of the Benjamini–Hochberg-corrected p-value < 0.05 are summarized in the **Supplementary Files**, and the genes shared between the main terms or pathways are presented in **Supplementary Figures 2–5**. Noticeably, both GESS and GED on success traits were enriched in neural development-related terms, such as neural crest cell migration (GO:0001755) for GESS on success traits and positive regulation of neuron projection development (GO:0010976) for GED on success traits. For GESS on calving traits, genes were enriched mainly in the thrombin-activated receptor signaling pathway (GO:0070493), microvillus (GO:0005902), neural precursor cell proliferation (GO:0061351), liver development (GO:0001889), *N*-methyltransferase activity (GO:0008170), sulfotransferase activity (GO:0008146), and cyclin-dependent protein kinase holoenzyme complex (GO:0000307). These categories include functionable genes named *CCNG2* and *CKS2* (GO:0000307), *LIMS2* (GO:0001889 and GO:0061351), *MYO7B*, *WWOX* (GO:0005902), *SMYD3*, *TFB2M* (GO:0008170), *F2RL1*, and *F2RL2* (GO:0070493) as previously discussed. In addition, the enrichment analysis for GED showed main terms such as *CXCR3* chemokine receptor binding (GO:0048248), homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156), and synapse pruning (GO:0098883), including *CXCL10* (GO:0048248), *PCDHA13* (GO:0007156), and *C1QB* (GO:0098883). Therefore, our findings provide further evidence of the possible genetic mechanism of GESS and GED on reproductive performance in Holstein cattle.

Future Prospects

Generally, only the GED for reproduction performance is analyzed in genetic evaluations, though research including the current study have shown the need to dissect GESS (Tiezzi et al., 2011). Compared to previous studies, we fitted an improved model considering the covariance between GESS and GED and also identified candidate genes associated with GESS and GED. The GESS on reproductive traits is small but significant. The low repeatability estimates indicate the poor consistency among repeated records. Therefore, more accurate records and novel traits are required. With recent improvements in data collection, GESS might become an important factor on the genetic evaluation of reproductive performance. Genomic selection is also expected to contribute to improve the accuracy of breeding value for these lowly heritable traits (Rice and Lipka, 2019).

Additional analyses with larger datasets and in independent populations (e.g., different breeds) are recommended.

CONCLUSION

The GESS on reproductive traits is heritable, with a similar genetic variance to the GED. Moreover, the approximate genetic correlation among the GESS and production, health, and type traits is unfavorable for the success traits (CR and NRR56) but favorable for the calving traits (CE, SB, and GL). A total of 100 and 62 significant SNPs were detected to be associated with GESS and GED on those five reproductive traits, respectively. Among them, five genes (*BMP4* and *CTCFL* for success traits and *WWOX*, *TFB2M*, and *SMYD3* for calving traits) are suggested as important candidate genes for GESS according to positional and functional analyses. As GESS and GED are lowly heritable, genomic prediction might be a promising alternative for breeding schemes aiming to improve fertility performance in dairy cattle.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: this manuscript utilizes proprietary data. Requests to access these datasets should be directed to YW, wangyachun@cau.edu.cn.

ETHICS STATEMENT

The studies involving animals were reviewed and approved by the Animal Welfare Committee of the China Agricultural University (Protocol Number: DK996).

AUTHOR CONTRIBUTIONS

ZC and YW conceived the study. ZC and LB designed the analyses, discussed the results, and drafted the manuscript. ZC performed all the data analyses and the data curation of genotype. ZC, HL, and RS prepared the phenotypic files. YC assisted with the modeling analyses. LL and GG provided support for the collection of raw data. All authors have read and approved the final version of the manuscript.

FUNDING

This study was supported by the China Agriculture Research System of MOF and MARA, the program for Changjiang Scholar and Innovation Research Team in University (IRT_15R62), and the National Agricultural Genetic Improvement Program (2130135). This study also received funding from Beijing Sanyuan Breeding Technology Ltd., Co. funded project (SYZYZ20190005). The funder had the following involvement with the study: study design. All authors declare no other competing interests.

ACKNOWLEDGMENTS

The authors acknowledge Beijing Sanyuan Breeding Technology Company Limited for providing access to the datasets used. The authors also grateful to Guosheng Su from Aarhus University for providing suggestions about the statistical model of the current study and to Qing Xu from Beijing Jiaotong University for giving advice on the genomic analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.713575/full#supplementary-material

Supplementary Figure 1 | The linkage disequilibrium decay of the population.

Supplementary Figure 2 | The Gene Ontology- and Kyoto Encyclopedia of Genes and Genomes-based network analysis of candidate genes of genetic effect of service sire on success traits using ClueGO application in Cytoscape.

Supplementary Figure 3 | The Gene Ontology- and Kyoto Encyclopedia of Genes and Genomes-based network analysis of candidate genes of genetic effect of service sire on calving traits using ClueGO application in Cytoscape.

Supplementary Figure 4 | The Gene Ontology- and Kyoto Encyclopedia of Genes and Genomes-based network analysis of candidate genes of genetic effects of the dam on success traits using ClueGO application in Cytoscape.

Supplementary Figure 5 | The Gene Ontology- and Kyoto Encyclopedia of Genes and Genomes-based network analysis of candidate genes of genetic effects of the dam on calving traits using ClueGO application in Cytoscape.

REFERENCES

Albrecht, E., Komolka, K., Ponsuksili, S., Gotoh, T., Wimmers, K., and Maak, S. (2016). Transcriptome profiling of Musculus longissimus dorsi in two cattle breeds with different intramuscular fat deposition. *Genomics Data* 7, 109–111. doi: 10.1016/j.gdata.2015.12.014

Averill, T. A., Rekaya, R., and Weigel, K. (2004). Genetic analysis of male and female fertility using longitudinal binary data. *J. Dairy Sci.* 87, 3947–3952. doi: 10.3168/jds.s0022-0302(04)73534-1

Azzam, S., Keele, J., and Nielsen, M. (1988). Expectations of heritability estimates for non-return rate of bulls and conception rate of cows. *J. Anim. Sci.* 66, 2767–2783.

Bagnato, A., and Oltenacu, P. (1993). Genetic study of fertility traits and production in different parities in Italian Friesian cattle. *J. Anim. Breed. Genet.* 110, 126–134.

Barton, S. C., Surani, M., and Norris, M. (1984). Role of paternal and maternal genomes in mouse development. *Nature* 311, 374–376. doi: 10.1038/311374a0

Bell, A., Ferrell, C., and Freetly, H. (1993). "Pregnancy and fetal metabolism," in *Quantitative Aspects of Ruminant Digestion and Metabolism*, eds J. Dijkstra, J. M. Forbes, and J. France (Ithaca, NY: Department of Animal Science, Cornell University), 405–431.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279–284. doi: 10.1016/s0166-4328(01)00297-2

Bennett, G., and Gregory, K. (2001). Genetic (co) variances for calving difficulty score in composite and parental populations of beef cattle: II. Reproductive, skeletal, and carcass traits. *J. Anim. Sci.* 79, 52–59. doi: 10.2527/2001.79152x

Berry, D., Evans, R., and Mc Parland, S. (2011). Evaluation of bull fertility in dairy and beef cattle using cow field data. *Theriogenology* 75, 172–181. doi: 10.1016/j.theriogenology.2010.08.002

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Boichard, D., and Manfredi, E. (1994). Genetic analysis of conception rate in French Holstein cattle. *Acta Agric. Scand. A Anim. Sci.* 44, 138–145. doi: 10.1080/09064709409410890

Bourdon, R., and Brinks, J. (1982). Genetic, environmental and phenotypic relationships among gestation length, birth weight, growth traits and age at first calving in beef cattle. *J. Anim. Sci.* 55, 543–553. doi: 10.2527/jas1982.553543x

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

Buzanskas, M. E., do Amaral Grossi , D., Ventura, R. V., Schenkel, F. S., Chud, T. C. S., Stafuzza, N. B., et al. (2017). Candidate genes for male and female reproductive traits in Canchim beef cattle. *J. Anim.l Sci. Biotechnol.* 8:67.

Calo, L., McDowell, R., VanVleck, L. D., and Miller, P. (1973). Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. *J. Anim. Sci.* 37, 676–682. doi: 10.2527/jas1973.373676x

Cervantes, I., Gutiérrez, J. P., Fernández, I., and Goyache, F. (2010). Genetic relationships among calving ease, gestation length, and calf survival to weaning in the Asturiana de los Valles beef cattle breed. *J. Anim. Sci.* 88, 96–101. doi: 10.2527/jas.2009-2066

Chu, G. C., Dunn, N. R., Anderson, D. C., Oxburgh, L., and Robertson, E. J. (2004). Differential requirements for Smad4 in TGFβ-dependent patterning of the early mouse embryo. *Development* 131, 3501–3512. doi: 10.1242/dev.01248

Cochran, S. D., Cole, J. B., Null, D. J., and Hansen, P. J. (2013). Single nucleotide polymorphisms in candidate genes associated with fertilizing ability of sperm and subsequent embryonic development in cattle. *Biol. Reprod.* 89:69.

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., et al. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC Genomics* 12:408. doi: 10.1111/j.1365-2052.2005.01337.x

Cole, J., and VanRaden, P. (2018). Symposium review: possibilities in an age of genomics: the future of selection indices. *J. Dairy Sci.* 101, 3686–3701. doi: 10.3168/jds.2017-13335

Cole, J., Wiggans, G., VanRaden, P., and Miller, R. (2007). Stillbirth (co) variance components for a sire-maternal grandsire threshold model and development of a calving ability index for sire selection. *J. Dairy Sci.* 90, 2489–2496. doi: 10.3168/jds.2006-436

Crews, D. Jr. (2006). Age of dam and sex of calf adjustments and genetic parameters for gestation length in Charolais cattle. *J. Anim. Sci.* 84, 25–31. doi: 10.2527/2006.84125x

Cubas, A., Berger, P., and Healey, M. (1991). Genetic parameters for calving ease and survival at birth in Angus field data. *J. Anim. Sci.* 69, 3952–3958. doi: 10.2527/1991.69103952x

Dadati, E., Kennedy, B., and Burnside, E. (1985). Relationships between conformation and reproduction in Holstein cows: type and calving performance. *J. Dairy Sci.* 68, 2639–2645. doi: 10.3168/jds.s0022-0302(85)81148-6

De Vos, J. A. (2018). *Genome Wide Association Study of Carcass Traits Based on Real Time Ultrasound in South African Nguni Cattle*. Hatfield: University of Pretoria.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 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.* 52:2.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Eaglen, S., Coffey, M., Woolliams, J., and Wall, E. (2013). Direct and maternal genetic relationships between calving ease, gestation length, milk production, fertility, type, and lifespan of Holstein-Friesian primiparous cows. *J. Dairy Sci.* 96, 4015–4025. doi: 10.3168/jds.2012-6229

Egger-Danner, C., Cole, J., Pryce, J., Gengler, N., Heringstad, B., Bradley, A., et al. (2015). Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. *Animal* 9, 191–207. doi: 10.1017/s1751731114002614

Eghbalsaied, S. (2011). Estimation of genetic parameters for 13 female fertility indices in Holstein dairy cows. *Trop. Anim. Health Prod.* 43, 811–816. doi: 10.1007/s11250-010-9767-z

Fang, L., Jiang, J., Li, B., Zhou, Y., Freebern, E., Vanraden, P. M., et al. (2019). Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. *Commun. Biol.* 2:100.

Fatehi, A., Van Den Hurk, R., Colenbrander, B., Daemen, A., Van Tol, H., Monteiro, R., et al. (2005). Expression of bone morphogenetic protein2 (BMP2), BMP4 and BMP receptors in the bovine ovary but absence of effects of BMP2 and BMP4 during IVM on bovine oocyte nuclear maturation and subsequent embryo development. *Theriogenology* 63, 872–889. doi: 10.1016/j.theriogenology.2004.05.013

Fleming, A., Baes, C. F., Martin, A., Chud, T., Malchiodi, F., Brito, L. F., et al. (2019). Symposium review: the choice and collection of new relevant phenotypes for fertility selection. *J. Dairy Sci.* 102, 3722–3734. doi: 10.3168/jds.2018-15470

Forde, N., Duffy, G. B., McGettigan, P. A., Browne, J. A., Mehta, J. P., Kelly, A. K., et al. (2012). Evidence for an early endometrial response to pregnancy in cattle: both dependent upon and independent of interferon tau. *Physiol. Genomics* 44, 799–810. doi: 10.1152/physiolgenomics.00067.2012

Fujii, T., Tsunesumi, S.-I., Yamaguchi, K., Watanabe, S., and Furukawa, Y. (2011). Smyd3 is required for the development of cardiac and skeletal muscle in zebrafish. *PLoS One* 6:e23491. doi: 10.1371/journal.pone.0023491

Ga, O. J., Schenkel, F., Alcantara, L., Houlahan, K., Lynch, C., and Baes, C. (2021). Estimated genetic parameters for all genetically evaluated traits in Canadian Holsteins. *J. Dairy Sci.* 104, 9002–9015. doi: 10.3168/jds.2021-20227

Gao, N., Chen, Y., Liu, X., Zhao, Y., Zhu, L., Liu, A., et al. (2019). Weighted single-step GWAS identified candidate genes associated with semen traits in a Duroc boar population. *BMC Genomics* 20:797.

Grigoletto, L., Ferraz, J., 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.

Guo, G., Guo, X., Wang, Y., Zhang, X., Zhang, S., Li, X., et al. (2014). Estimation of genetic parameters of fertility traits in Chinese Holstein cattle. *Can. J. Anim. Sci.* 94, 281–285. doi: 10.4141/cjas2013-113

Han, Y., and Peñagaricano, F. (2016). Unravelling the genomic architecture of bull fertility in Holstein cattle. *BMC Genet.* 17:143.

Hansen, M., Lund, M. S., Pedersen, J., and Christensen, L. (2004). Gestation length in Danish Holsteins has weak genetic associations with stillbirth, calving difficulty, and calf size. *Livest. Prod. Sci.* 91, 23–33. doi: 10.1016/j.livprodsci.2004.06.007

Heidari, S., Taromchi, A., Nejatbakhsh, R., and Shokri, S. (2018). Expression and localisation of RXFP 3 in human spermatozoa and impact of INSL 7 on sperm functions. *Andrologia* 50, e12928. doi: 10.1111/and.12928

Heringstad, B., Chang, Y., Svendsen, M., and Gianola, D. (2007). Genetic analysis of calving difficulty and stillbirth in Norwegian Red cows. *J. Dairy Sci.* 90, 3500–3507. doi: 10.3168/jds.2006-792

Hoekstra, J., Van der Lugt, A., Van der Werf, J., and Ouweltjes, W. (1994). Genetic and phenotypic parameters for milk production and fertility traits in upgraded dairy cattle. *Livest. Prod. Sci.* 40, 225–232. doi: 10.1016/0301-6226(94)90090-6

Hu, J., Chen, Y.-X., Wang, D., Qi, X., Li, T.-G., Hao, J., et al. (2004). Developmental expression and function of Bmp4 in spermatogenesis and in maintaining epididymal integrity. *Dev. Biol.* 276, 158–171. doi: 10.1016/j.ydbio.2004.08.034

Huang, J., Guo, F., Zhang, Z., Zhang, Y., Wang, X., Ju, Z., et al. (2016). PCK1 is negatively regulated by bta-miR-26a, and a single-nucleotide polymorphism in the 3' untranslated region is involved in semen quality and longevity of Holstein bulls. *Mol. Reprod. Dev.* 83, 217–225. doi: 10.1002/mrd.22613

Jamrozik, J., Fatehi, J., Kistemaker, G., and Schaeffer, L. (2005). Estimates of genetic parameters for Canadian Holstein female reproduction traits. *J. Dairy Sci.* 88, 2199–2208. doi: 10.3168/jds.s0022-0302(05)72895-2

Jansen, J. (1985). Genetic aspects of fertility in dairy cattle based on analysis of AI data—a review with emphasis on areas for further research. *Livest. Prod. Sci.* 12, 1–12. doi: 10.1016/0301-6226(85)90036-3

Jaton, C., Schenkel, F., Malchiodi, F., Sargolzaei, M., Price, C., Baes, C., et al. (2017). Genetic analysis for quality of frozen embryos produced by Holstein cattle donors in Canada. *J. Dairy Sci.* 100, 7320–7329. doi: 10.3168/jds.2017-12851

Jessmon, P., Leach, R. E., and Armant, D. R. (2009). Diverse functions of HBEGF during pregnancy. *Mol. Reprod. Dev.* 76, 1116–1127. doi: 10.1002/mrd.21066

Jia, C., Li, C., Fu, D., Chu, M., Zan, L., Wang, H., et al. (2020). Identification of genetic loci associated with growth traits at weaning in yak through a genome-wide association study. *Anim. Genet.* 51, 300–305. doi: 10.1111/age.12897

Jiang, J., Cole, J. B., Da, Y., VanRaden, P. M., and Ma, L. (2018). Fast Bayesian fine-mapping of 35 production, reproduction and body conformation traits with imputed sequences of 27K Holstein bulls. *BioRxiv* [Preprint]. BioRxiv: 428227,

Jiang, Z., Kunej, T., Wibowo, T. A., Michal, J. J., Zhang, Z., Gaskins, C. T., et al. (2006). "The basal nucleus-encoded mitochondrial transcription genes and meat quality in beef cattle," in *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production* (Belo Horizonte: Instituto Prociencia).

Johnson, T., Keehan, M., Harland, C., Lopdell, T., Spelman, R., Davis, S., et al. (2019). Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1. 2. *J. Dairy Sci.* 102, 3254–3258. doi: 10.3168/jds.2018-15638

King, K., Seidel, G. Jr., and Elsden, R. (1985). Bovine embryo transfer pregnancies. II. Lengths of gestation. *J. Anim. Sci.* 61, 758–762. doi: 10.2527/jas1985.614758x

Kiser, J. N., Keuter, E. M., Seabury, C. M., Neupane, M., Moraes, J. G., Dalton, J., et al. (2019). Validation of 46 loci associated with female fertility traits in cattle. *BMC Genomics* 20:576.

Koide, Y., Kiyota, T., Tonganunt, M., Pinkaew, D., Liu, Z., Kato, Y., et al. (2009). Embryonic lethality of fortilin-null mutant mice by BMP-pathway overactivation. *Biochim. Biophys. Acta (BBA) General Subjects* 1790, 326–338. doi: 10.1016/j.bbagen.2009.01.012

Kuhn, M., and Hutchison, J. (2008). Prediction of dairy bull fertility from field data: use of multiple services and identification and utilization of factors affecting bull fertility. *J. Dairy Sci.* 91, 2481–2492. doi: 10.3168/jds.2007-0743

Lee, K.-B., Zhang, K., Folger, J. K., Knott, J. G., and Smith, G. W. (2014). Evidence supporting a functional requirement of SMAD4 for bovine preimplantation embryonic development: a potential link to embryotrophic actions of follistatin. *Biol .Reprod.* 91:62.

Lemos, M. V., Chiaia, H. L. J., Berton, M. P., Feitosa, F. L., Aboujaoud, C., Camargo, G. M., et al. (2016). Genome-wide association between single nucleotide polymorphisms with beef fatty acid profile in Nellore cattle using the single step procedure. *BMC Genomics* 17:213.

Li, C. W., and Ge, W. (2011). Spatiotemporal expression of bone morphogenetic protein family ligands and receptors in the zebrafish ovary: a potential paracrine-signaling mechanism for oocyte-follicle cell communication. *Biol. Reprod.* 85, 977–986. doi: 10.1095/biolreprod.111.092239

Li, G., Khateeb, K., Schaeffer, E., Zhang, B., and Khatib, H. (2012a). Genes of the transforming growth factor-beta signalling pathway are associated with pre-implantation embryonic development in cattle. *J. Dairy Res.* 79:310. doi: 10.1017/s0022029912000210

Li, G., Peñagaricano, F., Weigel, K., Zhang, Y., Rosa, G., and Khatib, H. (2012b). Comparative genomics between fly, mouse, and cattle identifies genes associated with sire conception rate. *J. Dairy Sci.* 95, 6122–6129. doi: 10.3168/jds.2012-5591

Li, N., Pan, S., Zhu, H., Mu, H., Liu, W., and Hua, J. (2014). BMP4 promotes SSEA-1+ hUC-MSC differentiation into male germ-like cells in vitro. *Cell Prolif.* 47, 299–309. doi: 10.1111/cpr.12115

Lillehammer, M., Meuwissen, T., and Sonesson, A. (2011). A comparison of dairy cattle breeding designs that use genomic selection. *J. Dairy Sci.* 94, 493–500. doi: 10.3168/jds.2010-3518

Liu, J., Luo, X., Xu, Y., Gu, J., Tang, F., Jin, Y., et al. (2016). Single-stranded DNA binding protein Ssbp3 induces differentiation of mouse embryonic stem cells into trophoblast-like cells. *Stem Cell Res. Ther.* 7:79.

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

Lotfan, M., Ali, S. A., Yadav, M. L., Choudhary, S., Jena, M. K., Kumar, S., et al. (2018). Genome-wide gene expression analysis of 45 days pregnant fetal cotyledons vis-a-vis non-pregnant caruncles in buffalo (*Bubalus bubalis*). *Gene* 654, 127–137. doi: 10.1016/j.gene.2018.02.038

Luo, M., Boettcher, P., Dekkers, J., and Schaeffer, L. (1999). Bayesian analysis for estimation of genetic parameters of calving ease and stillbirth for Canadian Holsteins. *J. Dairy Sci.* 82, 1848. e1-e11.

Luo, M., Boettcher, P., Schaeffer, L., and Dekkers, J. (2002). Estimation of genetic parameters of calving ease in first and second parities of Canadian Holsteins using Bayesian methods. *Livest. Prod. Sci.* 74, 175–184. doi: 10.1016/s0301-6226(01)00294-9

Ma, L., Liu, X., Wang, F., He, X., Chen, S., and Li, W. (2015). Different donor cell culture methods can influence the developmental ability of cloned sheep embryos. *PLoS One* 10:e0135344. doi: 10.1371/journal.pone.0135344

Madsen, P., Sørensen, P., Su, G., Damgaard, L. H., Thomsen, H., and Labouriau, R, eds (2006). "DMU-a package for analyzing multivariate mixed models," in *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, Belo Horizonte.

Martinsson-Ahlzén, H.-S., Liberal, V., Grünenfelder, B., Chaves, S. R., Spruck, C. H., and Reed, S. I. (2008). Cyclin-dependent kinase-associated proteins Cks1 and Cks2 are essential during early embryogenesis and for cell cycle progression in somatic cells. *Mol. Cell Biol.* 28, 5698–5709. doi: 10.1128/mcb.01833-07

Masuda, K., Furuyama, T., Takahara, M., Fujioka, S., Kurinami, H., and Inagaki, S. (2004). Sema4D stimulates axonal outgrowth of embryonic DRG sensory neurones. *Genes Cells* 9, 821–829. doi: 10.1111/j.1365-2443.2004.00766.x

McClintock, S., Beard, K., Gilmour, A., and Goddard, M. (2003). Relationships between calving traits in heifers and mature cows in Australia. *Interbull. Bull.* 31:102.

Meijering, A. (1985). Sire evaluation for calving traits by Best Linear Unbiased Prediction and nonlinear methodology. *J. Anim. Breed. Genet.* 102, 95–105. doi: 10.1111/j.1439-0388.1985.tb00677.x

Miglior, F., Gong, W., Wang, Y., Kistemaker, G., Sewalem, A., and Jamrozik, J. (2009). Genetic parameters of production traits in Chinese Holsteins using a random regression test-day model. *J. Dairy Sci.* 92, 4697–4706. doi: 10.3168/jds.2009-2212

Mujibi, F., and Crews, D. Jr. (2009). Genetic parameters for calving ease, gestation length, and birth weight in Charolais cattle. *J Anim Sci* 87, 2759–2766. doi: 10.2527/jas.2008-1141

Müller, M.-P., Rothammer, S., Seichter, D., Russ, I., Hinrichs, D., Tetens, J., et al. (2017). Genome-wide mapping of 10 calving and fertility traits in Holstein dairy cattle with special regard to chromosome 18. *J. Dairy Sci.* 100, 1987–2006. doi: 10.3168/jds.2016-11506

Murray, B., Schaeffer, L., and Burnside, E. (1977). Estimation of breeding values for non return rates of Canadian Holstein Friesian sires. *Dairy Ind. Res. Rep.* 1977, 30–33.

Muuttoranta, K., Tyrisevä, A.-M., Mäntysaari, E. A., Pösö, J., Aamand, G. P., and Lidauer, M. H. (2019). Genetic parameters for female fertility in Nordic Holstein and Red Cattle dairy breeds. *J. Dairy Sci.* 102, 8184–8196. doi: 10.3168/jds.2018-15858

Nilsson, E. E., and Skinner, M. K. (2003). Bone morphogenetic protein-4 acts as an ovarian follicle survival factor and promotes primordial follicle development. *Biol. Reprod.* 69, 1265–1272. doi: 10.1095/biolreprod.103.018671

Norman, H., Hutchison, J., and Wright, J. (2008). *Sire Conception Rate: New National Ai Bull Fertility Evaluation.* AIPL Res Report SCR1 (7-08). Washington, DC: USDA.

Norrbom, J., Wallman, S., Gustafsson, T., Rundqvist, H., Jansson, E., and Sundberg, C. (2010). Training response of mitochondrial transcription factors in human skeletal muscle. *Acta Physiol.* 198, 71–79. doi: 10.1111/j.1748-1716.2009.02030.x

Ou, J., Li, Y., Wang, Z., Jin, C., Li, K., Lu, Y., et al. (2020). Lrrc34 is highly expressed in sscs and is necessary for SSC expansion in vitro. *Chin. Med. Sci. J.* 35, 20–30.

Pangas, S. A., Li, X., Robertson, E. J., and Matzuk, M. M. (2006). Premature luteinization and cumulus cell defects in ovarian-specific Smad4 knockout mice. *Mol. Endocrinol.* 20, 1406–1422. doi: 10.1210/me.2005-0462

Peddinti, D., Memili, E., and Burgess, S. C. (2010). Proteomics-based systems biology modeling of bovine germinal vesicle stage oocyte and cumulus cell interaction. *PLos One* 5:e11240. doi: 10.1371/journal.pone.0011240

Perkins, A., Wolfe, C., Eben, F., Soothill, P., and Linton, E. (1995). Corticotrophin-releasing hormone-binding protein in human fetal plasma. *J. Endocrinol.* 146, 395–401. doi: 10.1677/joe.0.1460395

Purfield, D., Bradley, D., Kearney, J., Evans, R., and Berry, D. (2015). Genome-wide association using high density genotypes for calving difficulty in dairy and beef cattle. *Genet. Sel. Evol.* 47:47.

Purfield, P. C., Evans, R. E., Carthy, T. R., and Berry, D. P. (2019). Genomic regions associated with gestation length detected using whole-genome sequence data differ between dairy and beef cattle. *Front. Genet.* 10:1068.

Raheja, K., Nadarajah, K., and Burnside, E. (1989). Relationship of bull fertility with daughter fertility and production traits in Holstein dairy cattle. *J. Dairy Sci.* 72, 2679–2682. doi: 10.3168/jds.s0022-0302(89)79409-1

Ramal-Sanchez, M., Bernabo, N., Tsikis, G., Blache, M.-C., Labas, V., Druart, X., et al. (2020). Progesterone induces sperm release from oviductal epithelial cells by modifying sperm proteomics, lipidomics and membrane fluidity. *Mol. Cell Endocrinol.* 504:110723. doi: 10.1016/j.mce.2020.110723

Rangaswami, H., Bulbule, A., and Kundu, G. C. (2006). Osteopontin: role in cell signaling and cancer progression. *Trends Cell Biol.* 16, 79–87. doi: 10.1016/j.tcb.2005.12.005

Rice, B., and Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *Plant Genome* 12, 1–14.

Shimizu, T., Yokoo, M., Miyake, Y., Sasada, H., and Sato, E. (2004). Differential expression of bone morphogenetic protein 4–6 (BMP-4,-5, and-6) and growth differentiation factor-9 (GDF-9) during ovarian development in neonatal pigs. *Domest. Anim. Endocrinol.* 27, 397–405. doi: 10.1016/j.domaniend.2004.04.001

Silvia, W., Hatler, T., Nugent, A., and Da Fonseca, L. L. (2002). Ovarian follicular cysts in dairy cows: an abnormality in folliculogenesis. *Domest. Anim. Endocrinol.* 23, 167–177. doi: 10.1016/s0739-7240(02)00154-6

Sokal, R. R., and Rohlf, F. J. (1981). *Biometry: The Principles And Practice of Statistics in Biological Research*. San Francisco, CA: W.H. Freeman.

Song, Q., Zhang, W., Wu, F., Zhang, J., Xu, M., Li, H., et al. (2019). Cloning and expression levels of TFAM and TFB2M genes and their correlation with meat and CarCass quality traits in Jiaxing BlaCk pig. *Ann. Anim. Sci.* 19, 327–341. doi: 10.2478/aoas-2018-0056

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* 11:316. doi: 10.3390/genes11030316

Steinbock, L., Näsholm, A., Berglund, B., Johansson, K., and Philipsson, J. (2003). Genetic effects on stillbirth and calving difficulty in Swedish Holsteins at first and second calving. *J. Dairy Sci.* 86, 2228–2235. doi: 10.3168/jds.s0022-0302(03)73813-2

Su, G., Lund, M. S., and Sorensen, D. (2007). Selection for litter size at day five to improve litter size at weaning and piglet survival rate. *J. Dairy Sci.* 85, 1385–1392. doi: 10.2527/jas.2006-631

Sun, C., and Su, G. (2010). Comparison on models for genetic evaluation of non-return rate and success in first insemination of the Danish Holstein cows. *Livest. Sci.* 127, 205–210. doi: 10.1016/j.livsci.2009.09.015

Sun, Y., Luo, W., Xie, H., Zhang, Y., and Cai, H. (2015). Analysis of association between polymorphism of TFB2M gene and meat quality, growth and slaughter traits in Guizhou White Goat, a well-known Chinese indigenous goat breed. *Pak. J. Zool.* 47, 1605–1610.

Taylor, J. F., Schnabel, R. D., and Sutovsky, P. (2018). Genomics of bull fertility. *Animal* 12, s172–s183.

Tiezzi, F., Maltecca, C., Penasa, M., Cecchinato, A., and Bittante, G. (2013). Genetic analysis of dairy bull fertility from field data of Brown Swiss cattle. *J. Dairy Sci.* 96, 7325–7328. doi: 10.3168/jds.2013-6885

Tiezzi, F., Penasa, M., Maltecca, C., Cecchinato, A., and Bittante, G. (2011). Exploring different model structures for the genetic evaluation of dairy bull fertility. *Agric. Conspec. Sci.* 76, 239–243.

Van Tassell, C., Wiggans, G., and Misztal, I. (2003). Implementation of a sire-maternal grandsire model for evaluation of calving ease in the United States. *J. Dairy Sci.* 86, 3366–3373. doi: 10.3168/jds.s0022-0302(03)73940-x

Vanderick, S., Troch, T., Gillon, A., Glorieux, G., and Gengler, N. (2014). Genetic parameters for direct and maternal calving ease in W alloon dairy cattle based on linear and threshold models. *J. Anim. Breed. Genet.* 131, 513–521. doi: 10.1111/jbg.12105

VanRaden, P., Sanders, A., Tooker, M., Miller, R., Norman, H., Kuhn, M., et al. (2004). Development of a national genetic evaluation for cow fertility. *J. Dairy Sci.* 87, 2285–2292. doi: 10.3168/jds.s0022-0302(04)70049-1

Vieira-Neto, A., Galvão, K., Thatcher, W., and Santos, J. (2017). Association among gestation length and health, production, and reproduction in Holstein cows and implications for their offspring. *J. Dairy Sci.* 100, 3166–3181. doi: 10.3168/jds.2016-11867

Walker, C. G., Littlejohn, M. D., Mitchell, M. D., Roche, J. R., and Meier, S. (2012). Endometrial gene expression during early pregnancy differs between fertile and subfertile dairy cow strains. *Physiol. Genomics* 44, 47–58. doi: 10.1152/physiolgenomics.00254.2010

Weigel, K., and Rekaya, R. (2000). Genetic parameters for reproductive traits of Holstein cattle in California and Minnesota. *J. Dairy Sci.* 83, 1072–1080. doi: 10.3168/jds.s0022-0302(00)74971-x

Weintraub, A. S., Lin, X., Itskovich, V. V., Aguinaldo, J. G. S., Chaplin, W. F., Denhardt, D. T., et al. (2004). Prenatal detection of embryo resorption in osteopontin-deficient mice using serial noninvasive magnetic resonance microscopy. *Pediatric Res.* 55, 419–424. doi: 10.1203/01.pdr.0000112034.98387.b2

Wiggans, G., Cooper, T., VanRaden, P., and Cole, J. (2011). Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. *J. Dairy Sci.* 94, 6188–6193. doi: 10.3168/jds.2011-4481

Wiggans, G., Misztal, I., and Van Tassell, C. (2003). Calving ease (co) variance components for a sire-maternal grandsire threshold model. *J. Dairy Sci.* 86, 1845–1848. doi: 10.3168/jds.s0022-0302(03)73771-0

Wu, X., Fang, M., Liu, L., Wang, S., Liu, J., Ding, X., et al. (2013). Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. *BMC Genomics* 14:897.

Zhang, X., Chu, Q., Guo, G., Dong, G., Li, X., Zhang, Q., et al. (2017). Genome-wide association studies identified multiple genetic loci for body size at four growth stages in Chinese Holstein cattle. *PLoS One* 12:e0175971.

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. *Heredity* 123, 202–214.

Conflict of Interest: LL is employed by Beijing Dairy Cattle Center. GG is employed by Beijing Sunlon Livestock Development Company Limited.

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.

frontiers in Genetics

METHODS
published: 11 October 2021
doi: 10.3389/fgene.2021.717457

Single-Trait and Multiple-Trait Genomic Prediction From Multi-Class Bayesian Alphabet Models Using Biological Information

*Zigui Wang and Hao Cheng**

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

Genomic prediction has been widely used in multiple areas and various genomic prediction methods have been developed. The majority of these methods, however, focus on statistical properties and ignore the abundant useful biological information like genome annotation or previously discovered causal variants. Therefore, to improve prediction performance, several methods have been developed to incorporate biological information into genomic prediction, mostly in single-trait analysis. A commonly used method to incorporate biological information is allocating molecular markers into different classes based on the biological information and assigning separate priors to molecular markers in different classes. It has been shown that such methods can achieve higher prediction accuracy than conventional methods in some circumstances. However, these methods mainly focus on single-trait analysis, and available priors of these methods are limited. Thus, in both single-trait and multiple-trait analysis, we propose the multi-class Bayesian Alphabet methods, in which multiple Bayesian Alphabet priors, including RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO, can be used for markers allocated to different classes. The superior performance of the multi-class Bayesian Alphabet in genomic prediction is demonstrated using both real and simulated data. The software tool JWAS offers open-source routines to perform these analyses.

Keywords: multiple-trait, multi-class, genomic prediction, Bayesian Alphabet, biological information

OPEN ACCESS

Edited by:
Peng Xu,
Xiamen University, China

Reviewed by:
Mario Calus,
Wageningen University and Research, Netherlands
Ming Fang,
Jimei University, China

***Correspondence:**
Hao Cheng
qtlcheng@ucdavis.edu

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

Received: *31 May 2021*
Accepted: *23 August 2021*
Published: *11 October 2021*

Citation:
Wang Z and Cheng H (2021) Single-Trait and Multiple-Trait Genomic Prediction From Multi-Class Bayesian Alphabet Models Using Biological Information. Front. Genet. 12:717457. doi: 10.3389/fgene.2021.717457

1. INTRODUCTION

Genomic prediction, proposed by Meuwissen et al. (2001), utilizes genomic information, such as single-nucleotide polymorphisms (SNPs), to estimate genotypic values or breeding values of complex traits. In the last decades, with the fast development of genotyping and sequencing technology, high-density genotype data has become much easier to access (Harris et al., 2011; Kranis et al., 2013). Accompanied by the high-density data, genomic prediction has been widely used in many areas, including animal breeding (e.g., Hayes et al., 2009a; Erbe et al., 2012), plant breeding (e.g., Wang et al., 2018; Moeinizade et al., 2020), and human disease risk prediction (e.g., Abraham et al., 2014, 2016).

A large number of genomic prediction methods with different statistical assumptions have been developed. Among these methods, genomic best linear unbiased prediction (GBLUP) (Habier et al., 2007; VanRaden, 2008; Hayes et al., 2009b), where a genomic relationship matrix is used to accommodate the covariances among breeding values, is widely used. GBLUP, however, assumes

a priori that all marker effects share the same normal distribution, which may not be biologically meaningful, especially for traits controlled by a few causal variants. Furthermore, a collection of Bayesian Alphabet methods (Meuwissen et al., 2001; Fernando and Garrick, 2013; Cheng et al., 2018b; Gianola and Fernando, 2020) have been developed to incorporate different priors on marker effects, e.g., BayesA and BayesB (Meuwissen et al., 2001). Notice that GBLUP is equivalent to a Bayesian Alphabet model with a normal prior for the marker effects (Fernando, 1998; Habier et al., 2007; Strandén and Garrick, 2009). These methods, however, are still developed mainly based on statistical consideration and ignore the abundant biological information. To bridge the gap between the statistical model for genomic prediction and underlying biological architectures, researchers have proposed several methods to incorporate biological information into genomic prediction and have shown that incorporating biological information has the potential to improve the prediction accuracy in some cases (Zhang et al., 2014; Gao et al., 2015; Edwards et al., 2016).

One purpose of incorporating biological information is to relax the assumption that each locus is equally likely to affect the trait, i.e., all loci share the same prior distribution. This assumption is less biologically meaningful, e.g., some loci may be known to lead non-synonymous coding changes or have functional effects on candidate genes (MacLeod et al., 2016). One strategy to achieve this purpose is weighting markers based on the biological information and then integrating the weighting information into the model construction (Zhang et al., 2014; Gao et al., 2015). Zhang et al. (2014) incorporated the QTL list obtained in previous genome-wide association studies (GWAS) into GBLUP, i.e., when constructing genomic relationship matrix, markers were weighted based on the frequency of corresponding genomic regions being reported in the QTL list (Zhang et al., 2014). Gao et al. (2015) incorporated previous GWAS results by using locus-specific inclusion probability based on the *p*-values from GWAS.

In addition to weighting markers, another strategy to incorporate biological information is marker allocation. It has been observed that molecular markers from different genomic regions have different prediction abilities (Erbe et al., 2012; Morota et al., 2014; Do et al., 2015; Abdollahi-Arpanahi et al., 2016) and the marker allocation is beneficial if a particular class is enriched for QTL. To better fit these genomic regions with different genetic architectures, recent studies have tried to allocate genome-wide molecular markers into multiple classes based on the prior biological information and conduct genomic prediction based on these marker classes jointly. Speed and Balding (2014) proposed such a method under the GBLUP framework called MultiBLUP, which divides breeding values into multiple classes to allow different effect-size variances. A Bayesian regression method called BayesRC (MacLeod et al., 2016) was also proposed to allocate SNPs into multiple classes, where a BayesR prior was assigned to each class. It has been shown that allocating markers into different classes can improve predictive accuracy in some circumstances (Speed and Balding, 2014; MacLeod et al., 2016). The idea to allocate markers into multiple classes has also been used in a haplotype-based genomic prediction model (Xu et al., 2020), in which effects of haplotype blocks are estimated using both numerical dosage and categorical coding strategies (Martini et al., 2017) for each genomic class.

To our knowledge, most methods that allocate SNPs into different classes, focus on single-trait analysis and available priors of these methods are limited. Thus, the primary goal of this research is to present a more general Bayesian Alphabet method that can handle both single-trait and multiple-trait analysis, while is able to assign multiple Bayesian Alphabet priors, including RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO, to markers in different SNP classes. The new genomic prediction method we implemented is called multi-class Bayesian Alphabet, where the term "Bayesian Alphabet" denotes a collection of Bayesian Alphabet priors adopted for marker effects. Our multi-class Bayesian Alphabet works for both single-trait and multiple-trait analysis. The performance of the multi-class Bayesian Alphabet is studied using real and simulated data.

2. MATERIALS AND METHODS

2.1. Multi-Class Bayesian Alphabet Models

For simplicity, the general mean is assumed as the only fixed effect, thus the general form of the multi-class Bayesian Alphabet model for *i*th genotyped observation can be written as:

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \sum_{l=1}^{g} \sum_{f_l \in C_l} m_{if_l} \boldsymbol{\alpha}_{f_l} + \boldsymbol{e}_i \tag{1}$$

where $\boldsymbol{y}_i$ is a vector of phenotypic values of t traits for observation i; $\boldsymbol{\mu}$ is a vector of overall means for t traits; m_{if_l} is the genotype covariate at locus f_l (coded as 0,1,2) in SNP class C_l for observation i; g is the number of SNP classes; $\boldsymbol{\alpha}_{f_l}$ is a vector of the corresponding allele substitution effects (marker effects) of t traits for locus f_l; and $\boldsymbol{e}_i$ is a vector of residuals for observation i. Note that when the number of traits $t = 1$, the general form above simplifies to the single-trait model, and all vectors of effects in Equation 1 become scalars. The fixed effect $\boldsymbol{\mu}$ is assigned a flat prior. The residuals, $\boldsymbol{e}_i$, are a priori assumed to be independently and identically distributed multivariate normal vectors with null mean and covariance matrix $\boldsymbol{R}$, which is assigned an inverse Wishart prior distribution, $\boldsymbol{W}^{-1}(\boldsymbol{S}_e, \nu_e)$, with degrees of freedom $\nu_e = 4$ and scale matrix $\boldsymbol{S}_e$ such that the prior mean of $\boldsymbol{R}$ equals half of the phenotypic variance. Note that when number of traits $t = 1$, the prior for $\boldsymbol{R}$ follows a scaled inverted chi-square distribution.

To incorporate known biological information, marker effects of SNPs in the same class are assumed to have identical Bayesian Alphabet prior. Different from conventional Bayesian Alphabet methods, our multi-class Bayesian Alphabet methods allow assigning different Bayesian Alphabet priors to marker effect $\boldsymbol{\alpha}_{f_l}$ in different SNP classes. These priors are discussed in the following section 2.2.

2.2. Bayesian Prior for Marker Effects

Multiple priors are implemented in our multi-class Bayesian Alphabet models, including BayesA, BayesB, BayesCΠ, RR-BLUP, and Bayesian LASSO. In multiple-trait analysis, with BayesB and BayesCΠ priors, each locus is allowed to affect any combination of traits (Cheng et al., 2018b). In multiple-trait BayesB and BayesCΠ, the vector of marker effects at locus f_l can be written as $\boldsymbol{\alpha}_{f_l} = \boldsymbol{D}_{f_l}\boldsymbol{\beta}_{f_l}$, where $\boldsymbol{D}_{f_l}$ is a diagonal matrix whose diagonal elements are $\boldsymbol{\delta}_{f_l} = (\delta_{f_l1}, \delta_{f_l2}..., \delta_{f_lt})$, where δ_{f_lt} is an indicator variable indicating whether the marker effect of locus f_l for trait t is zero or not. We use numeric labels "1," "2,"$\cdots$, "z" to represent all possible combinations for $\boldsymbol{\delta}_{f_l}$, in which case the prior distribution for $\boldsymbol{\delta}_{f_l}$ is: $p(\boldsymbol{\delta}_{f_l} = \text{"i"}) = \Pi_1 I(\boldsymbol{\delta}_{f_l} = \text{"1"}) + \Pi_2 I(\boldsymbol{\delta}_{f_l} = \text{"2"}) + ... + \Pi_l I(\boldsymbol{\delta}_{f_l} = \text{"}z\text{"})$ where Π_i is the prior probability that the vector $\boldsymbol{\delta}_{f_l}$ corresponds to the vector labeled "i" and $\sum \Pi_i = 1$. A uniform prior distribution is assigned to $\Pi = (\Pi_1, \Pi_2, ...\Pi_l)$ (Cheng et al., 2018b). In multiple-trait BayesB, the prior for $\boldsymbol{\beta}_{f_l}$ is a multivariate normal distribution with null mean and locus-specific covariance matrix $\boldsymbol{G}_{f_l}$, which is assigned an inverse Wishart prior, $W_t^{-1}(\boldsymbol{S}_\beta, \nu_\beta)$. In multiple-trait BayesCΠ, instead of locus-specific covariance matrix $\boldsymbol{G}_{f_l}$, $\boldsymbol{\beta}_{f_l}$ is assumed to follow a multivariate normal prior with null mean and common covariance matrix $\boldsymbol{G}$, which is assumed to have an inverse Wishart prior distribution, $W^{-1}(\boldsymbol{S}_\beta, \nu_\beta)$, with degrees of freedom $\nu_\beta = 4$ and scale matrix $\boldsymbol{S}_\beta$ such that the prior mean of genetic variance equals half of the phenotypic variance. In single-trait analysis, $\boldsymbol{D}_{f_l}$, $\boldsymbol{G}_{f_l}$, and marker effect $\boldsymbol{\beta}_{f_l}$ become scalars. The prior of β_{f_l} becomes a univariate normal distribution; the prior of G_{f_l} becomes an inverted chi-square distribution, and D_{f_l} is an indicator variable indicating whether the maker effect is zero or not. In both single-trait and multiple-trait analysis, BayesA and RR-BLUP are just special cases of BayesB and BayesCΠ respectively, where all markers are assumed to have effects on all traits (Fernando and Garrick, 2013). The Bayesian LASSO prior is also included in the multi-class Bayesian Alphabet. In Bayesian LASSO, the multivariate Laplace prior distribution with a null mean is assigned to marker effect vector $\boldsymbol{\alpha}_{f_l}$ (Gianola and Fernando, 2020) in multiple-trait analysis. In single-trait Bayesian LASSO, the prior for α_{f_l} is a double exponential distribution (Tibshirani, 1996; Gianola, 2013).

2.3. Data Analysis

2.3.1. Real Data

Two public datasets are used to evaluate the performance of multi-class Bayesian Alphabet models. The first dataset, which is used to evaluate the single-trait analysis, is composed of genotypic and phenotypic data from Michigan State University Pig Resource Population (MSUPRP) raised at the Michigan State University Swine Teaching and Research Farm, East Lansing,MI (Edwards et al., 2008). After quality control (Duarte et al., 2014), 928 individuals and 42,246 SNPs remain. The trait *13-week tenth rib backfat (mm)* is considered in this analysis. The original data is available at https://msu.edu/~steibelj/JP_files/GBLUP.html. The genome annotation information for the pig dataset used in this paper is obtained from the Ensembl (Rainer et al., 2019) database using the GALLO package (Fonseca et al., 2020). Five annotation regions are identified in the pig dataset, and will be used in our analysis. The number of SNPs in the protein coding, RNA, processed pseudogene, intergenic, and pseudogene regions are 15084, 1840, 107, 24838, and 377, respectively.

The second dataset, which is used to evaluate the multiple-trait analysis, is from the Rice Diversity Panel with 370 *Oryza sativa* individual accessions (Zhao et al., 2011). Three traits *plant height (PH)*, *flowering time in Arkansas (FTA)*, and *panicle number per plant (PN)* are considered. After removing the genotypes missing for these three traits or with minor allele frequency < 0.05, 33,519 SNPs are included in our analysis. The phenotypic and genotypic data are publicly available at http://www.ricediversity.org/. The genome annotation information for the rice dataset is obtained from Ensembl (Rainer et al., 2019) database using the biomart package (Durinck et al., 2009). Four annotation regions are identified in the rice dataset, and will be used in our analysis. The number of SNPs in protein coding, RNA, non-translating CDS, and intergenic regions are 14129, 3, 176, and 19211, respectively.

We identified total 6 genomic annotations: protein coding, processed pseudogene, pseudogene, non-coding RNA, non-translating CDS, and intergenic. According to Howe et al. (2020), the "protein coding" class is comprised of the SNPs within the gene that contains an open reading frame (ORF). In other words, these SNPs may be processed into messenger RNAs (mRNAs) which, after their export to the cytosol, are translated into proteins (Harrow et al., 2009). The "pseudogene" class contains SNPs within the genes that have coding-sequence deficiencies like frameshifts and premature stop codons but resemble protein-coding genes (Howe et al., 2020; Tutar, 2012). The "processed pseudogene" class includes the SNPs in the pseudogene that lack introns and is thought to arise from reverse transcription of messenger RNA followed by reinsertion of DNA into the genome (Howe et al., 2020). The "non-coding RNA" class contains SNPs within RNA that are not translated into a protein (Howe et al., 2020). The "non-translating CDS" class represents SNPs in coding sequence regions that are not translated to a protein (Howe et al., 2020). All other SNPs were allocated to the class "intergenic".

2.3.2. Simulated Data

To comprehensively compare multi-class Bayesian Alphabet with conventional Bayesian Alphabet for genomic prediction, we conducted simulations based on the real genotypes from Michigan State University Pig Resource Population (MSUPRP) described above (Edwards et al., 2008). The simulation strategies in MacLeod et al. (2016) were applied. 500 QTLs were randomly selected from SNP class "protein coding", i.e., SNPs with the annotation "protein coding". In addition, 20 QTLs were randomly selected across the genome. The same QTL positions were used in our simulation. Two correlated traits of heritabilities equal to 0.5 and 0.9 were simulated, where pleiotropic QTL effects were sampled from a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{G} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. The trait of heritability 0.5 was used in our single-trait analysis, and both traits were used in our multiple-trait analysis. There were total

TABLE 1 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet on single-trait simulated data.

Method	RR-BLUP	BayesA	BayesB	BayesCπ	Bayesian LASSO	Ensemble
Conventional	0.542*	0.542	0.547*	0.547*	0.541*	0.545*
Multi-class	0.563*	0.542	0.565*	0.565*	0.563*	0.563*

*The comparison of mean prediction accuracy across 150 single-trait validation datasets (30 simulated data × five-fold cross validation) between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. *Denotes that significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).*

TABLE 2 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet on multiple-trait simulated data.

Method	RR-BLUP	BayesA	BayesB	BayesCΠ	Bayesian LASSO	Ensemble
Conventional	0.552*	0.554	0.565*	0.564*	0.552*	0.561*
Multi-class	0.572*	0.553	0.578*	0.577*	0.572*	0.575*

*The comparison of mean prediction accuracy across 150 multiple-trait validation datasets (30 simulated data × five-fold cross validation) between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. *Denotes that significant differences were found between multi-class Baysian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ, Baysian LASSO, and ensemble approach, respectively ($p < 0.1$).*

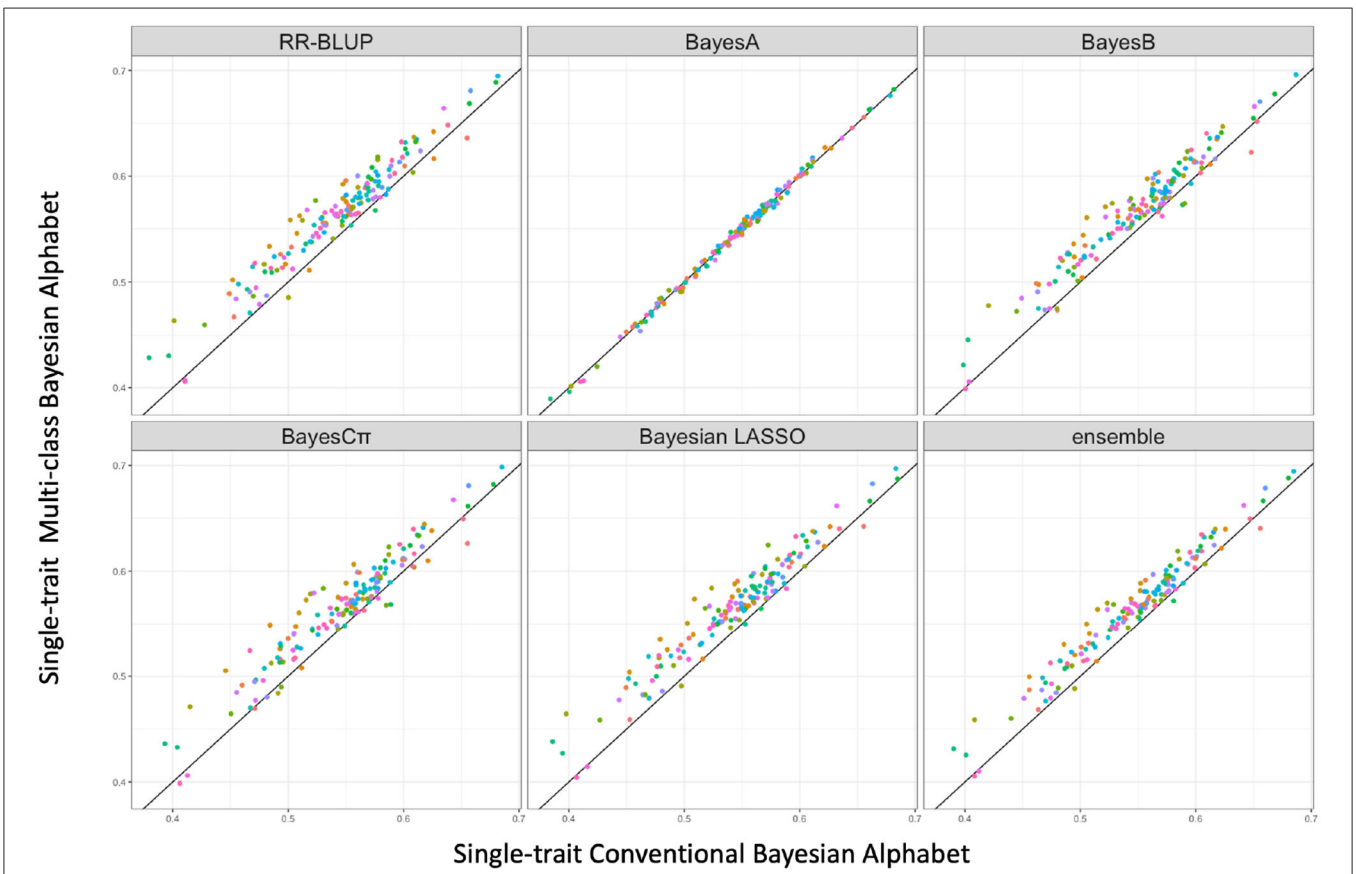

FIGURE 1 | The pairwise predication accuracy comparison between conventional and multi-class Bayesian Alphabet on single-trait simulated data. The 30 simulated datasets were distinguished by color; the x-axis represents the genomic prediction accuracy obtained from conventional Bayesian Alphabet methods; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet methods; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. Significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).

30 different datasets being simulated based on the simulation processes described above.

2.3.3. Cross Validation

The dataset was randomly split into training and validation datasets following an 8:2 ratio for each replicate. 50 replicates and 5 replicates were applied to the real and simulated datasets, respectively. The prediction accuracy was calculated as the mean Pearson correlation between the estimated breeding values and phenotypic records of observations in validation datasets. Conventional and multi-class Bayesian Alphabet methods were compared using RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO priors. In addition to the above five Bayesian methods, an ensemble approach that uses average estimated breeding values across five Bayesian methods, was used to integrate multiple predictions into one summary prediction (Azodi et al., 2019).

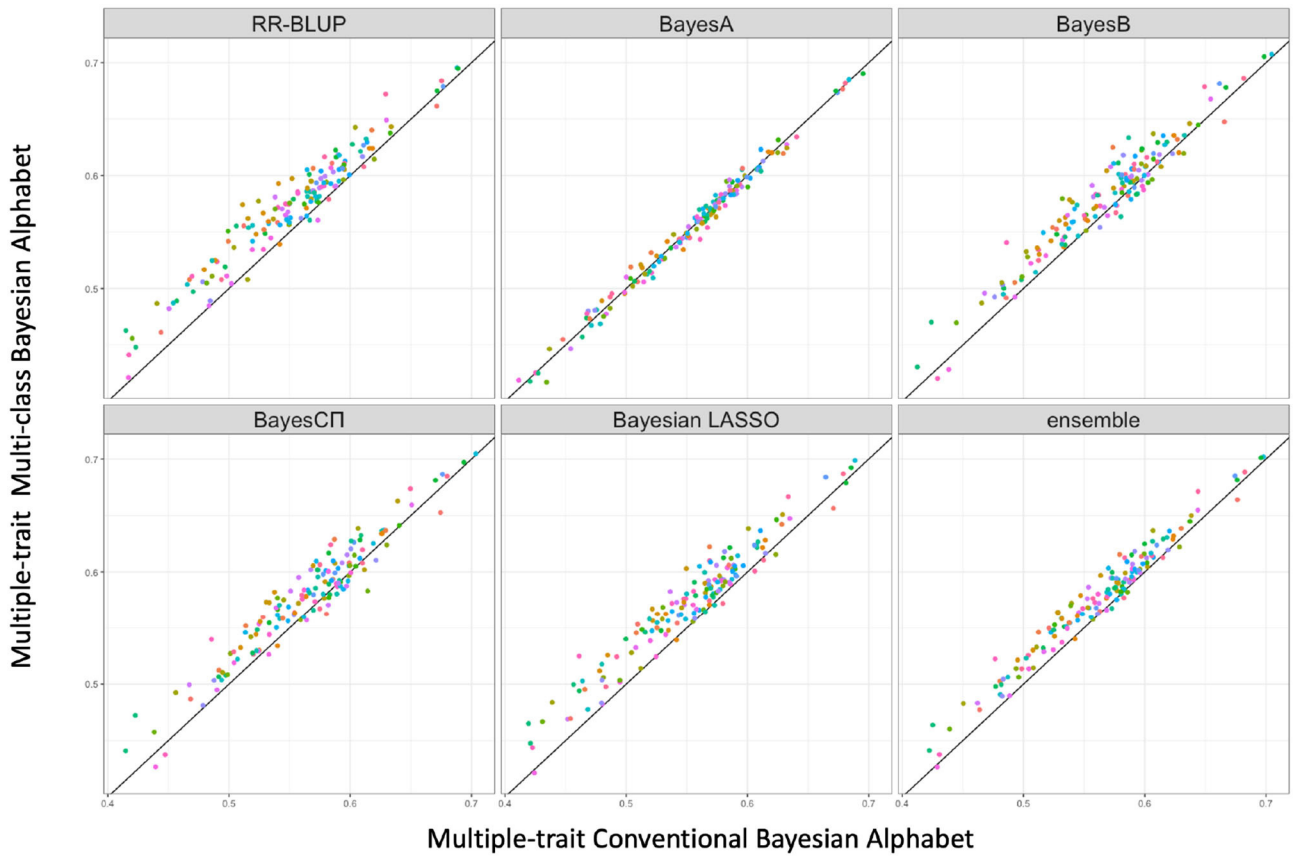

FIGURE 2 | The pairwise predication accuracy comparison between conventional Bayesian Alphabet and multi-class Bayesian Alphabet on multiple-trait simulated data. The 30 simulated datasets were distinguished by color; the x-axis represents the genomic prediction accuracy obtained from conventional Bayesian Alphabet method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. Significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).

TABLE 3 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet for real pig data (single-trait) and real rice data (multiple-trait).

Data	Method	RR-BLUP	BayesA	BayesB	BayesCΠ	Bayesian LASSO	Ensemble
Pig	Conventional	0.516	0.565	0.568	0.532	0.517	0.550
	Multi-class	0.516	0.565	0.569	0.532	0.516	0.550
Rice	Conventional	0.378	0.353	0.372	0.384	0.378	0.377
	Multi-class	0.374	0.357	0.363	0.375	0.373	0.373

The comparison of mean prediction accuracy on the trait 13-week tenth rib backfat (mm) from pig data and trait FTA of rice real data across 50 validation datasets between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. No significant differences were found between multi-class and conventional Bayesian Alphabet methods for both real pig and rice data ($p < 0.1$).

The molecular markers were allocated into multiple classes using the genome annotation information. SNP classes were defined using the genome annotation information, i.e., SNPs with the same genome annotation were allocated in one class.

We have implemented these methods in JWAS (Cheng et al., 2018a), an open-source package for single-trait and multiple-trait genome-enabled prediction and analyses. The software tool JWAS offers open-source routines to perform these analyses. The documentation and examples of JWAS can be found at https://github.com/reworkhow/JWAS.jl. MCMC chains of length 100,000 with a burn-in of the first 50,000 iterations were used. The Gelman-Rubin test (Gelman and Rubin, 1992) has been used to verify the convergence of the MCMC chain.

3. RESULT

3.1. Simulated Data

Multi-class Bayesian Alphabet methods using genome annotation information were performed for both single-trait and multiple-trait prediction on the simulated data. In both single-trait and multiple-trait analysis, 5-fold cross validation was applied on 30 simulated datasets. The comparisons between multi-class and conventional Bayesian Alphabet methods are shown in **Table 1** for single-trait analysis and **Table 2** for multiple-trait analysis. The pairwise comparisons across all 30 simulated datasets are also shown in **Figure 1** for single-trait analysis and **Figure 2** for multiple-trait analysis. The 30 simulated datasets are distinguished by color. The paired t-test with a significance level 0.1 is used to declare the significant difference between prediction accuracies from multi-class and conventional Bayesian Alphabet methods.

In the single-trait analysis, significant differences in prediction accuracies were detected between multi-class and conventional Bayesian Alphabet methods with RR-BLUP, BayesB, BayesCπ, Bayesian LASSO priors, and the ensemble approach ($p < 0.1$). In detail, the mean prediction accuracies of multi-class Bayesian Alphabet were higher than conventional Bayesian Alphabet in 30 out of all 30 datasets with RR-BLUP, BayesB, BayesCπ, Bayesian LASSO priors, and ensemble approach. Multi-class Bayesian Alphabet significantly outperforms conventional Bayesian Alphabet in the ensemble approach due to the better performance of multi-class Bayesian Alphabet using these 4 priors.

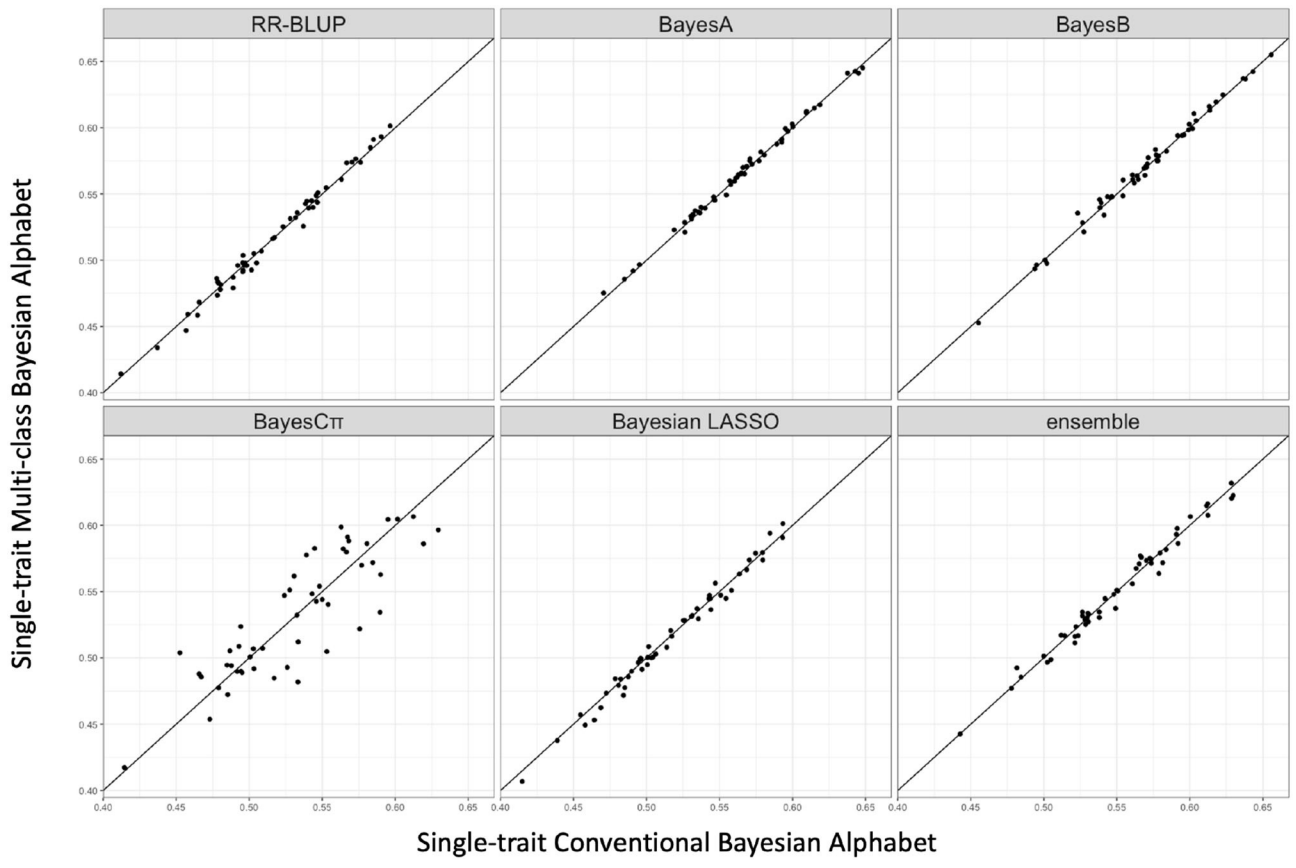

FIGURE 3 | The pairwise prediction accuracy comparisons between conventional Bayesian method and multi-class Bayesian Alphabet using genome annotation classes for the trait *13-week tenth rib backfat (mm)* from pig real data. The x-axis represents the genomic prediction accuracy obtained from conventional Bayesian method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. No significant differences were found ($p < 0.1$).

In the multiple-trait analysis, no significant differences were observed for the higher heritability trait, and results for the lower heritability trait were presented. Overall, higher prediction accuracies were usually observed for the same prior in multiple-trait analysis compared to single-trait analysis. A significant difference in prediction accuracies was detected between multi-class and conventional Bayesian Alphabet methods with RR-BLUP, BayesB, BayesCΠ, Bayesian LASSO prior ($p < 0.1$) as well as the ensemble approach. Similar to single-trait simulation result, the mean prediction accuracies of multi-class Bayesian Alphabet were higher than conventional Bayesian Alphabet in 30 out of all 30 simulated datasets with RR-BLUP, BayesB, BayesCπ Bayesian LASSO priors and the ensemble approach. The simulated data result shows that the multi-class Bayesian Alphabet has the potential to improve the prediction accuracy for both single-trait and multiple-trait analysis.

3.2. Real Data

Multi-class Bayesian Alphabet methods were performed on the pig data (Edwards et al., 2008) for single-trait analysis and the rice data (Zhao et al., 2011) for multiple-trait analysis. In the multiple-trait analysis, three traits *PH*, *FTA* and *PN* showed similar patterns on the comparison between conventional and multi-class Bayesian Alphabet methods, so only results of trait *FTA* were presented for simplicity. In both single-trait and multiple-trait analysis, 50-fold cross validation was applied. The comparison between multi-class and conventional Bayesian Alphabet methods are shown in **Table 3** for single-trait analysis and multiple-trait analysis. The pairwise comparisons across all 50 validation datasets are also shown in **Figure 3** for single-trait analysis, and **Figure 4** for multiple-trait analysis. The paired t-test with a significance level 0.1 was used to declare the significant difference between prediction accuracies from multi-class and conventional Bayesian Alphabet methods.

As shown in **Table 3**, in both real pig (single-trait) and rice (multiple-trait) data analysis, the prediction accuracies of multi-class Bayesian Alphabet using genome annotation information were not significantly different from conventional Bayesian Alphabet methods for all priors and ensemble approach.

We further studied the effect of SNP allocation on prediction accuracy by using other types of known biological information. For example, we allocated SNPs on the same chromosome to the same class such that number of chromosomes classes are fitted in multi-class Bayesian Alphabet methods. As shown

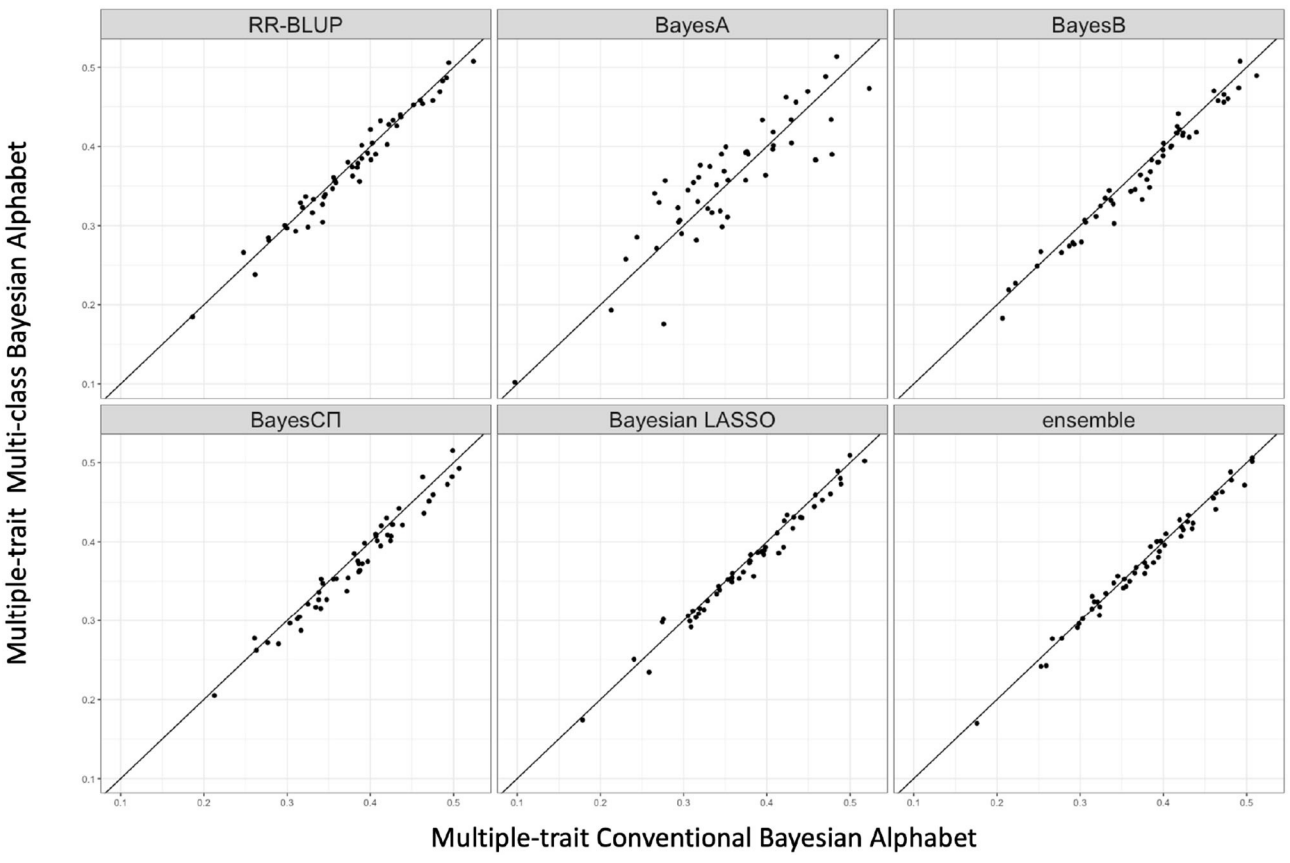

FIGURE 4 | The pairwise accuracy comparisons between conventional Bayesian method and multi-class Bayesian Alphabet using genome annotation classes for the trait *FTA* of rice real data. The x-axis represents the genomic prediction accuracy obtained from conventional Bayesian method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. No significant differences were found ($p < 0.1$).

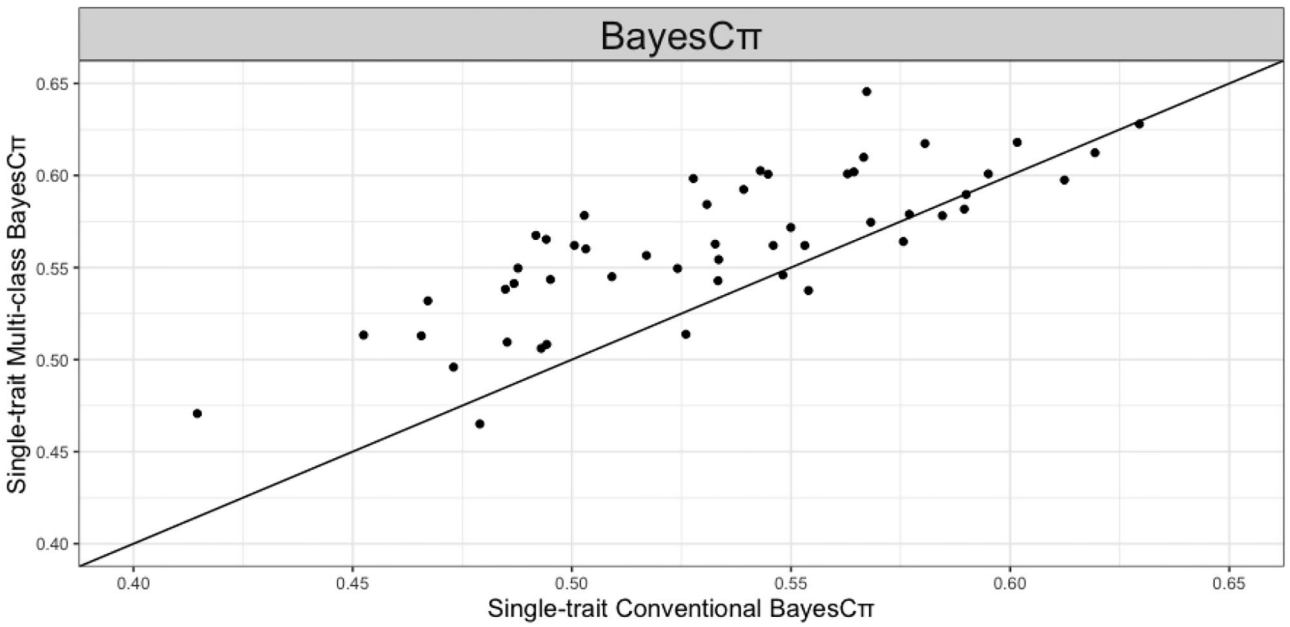

FIGURE 5 | The pairwise prediction accuracy comparisons of conventional BayesCπ and multi-class BayesCπ using chromosome classes for the trait *13-week tenth rib backfat (mm)* of real pig data. The x-axis represents the genomic prediction accuracy obtained from conventional BayesCπ; the y-axis represents the genomic prediction accuracy obtained from multi-class BayesCπ; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class BayesCπ. Multi-class BayesCπ significantly outperformed the conventional BayesCπ ($p < 0.1$).

in **Figure 5**, in the real pig (single-trait) data analysis, when BayesCπ prior is used, multi-class Bayesian Alphabet using chromosome classes has significantly higher prediction accuracy than the conventional Bayesian Alphabet ($p < 0.1$). To further understand why higher prediction accuracy is achieved in multi-class BayesCπ using chromosome classes, a genome-wide association study (GWAS) was performed on the same dataset, and one significant signal was detected on chromosome 6 (Chen et al., 2017). Thus, we ran another multi-class Bayesian alphabet analysis by allocating SNPs on chromosome 6 to one class and the remaining to another for a 2-class Bayesian Alphabet analysis. Higher prediction accuracy was observed in this 2-class Bayesian Alphabet analysis. It indicates that assigning SNPs into classes based on GWAS results may be one useful strategy to incorporate biological information.

4. DISCUSSION

Most genomic prediction methods usually assume all marker effects share the same prior distribution. This assumption, however, is not biologically meaningful and may potentially reduce the prediction performance when genetic architectures vary across different genomic regions (Speed and Balding, 2014). To address this issue, some methods such as MultiBLUP (Speed and Balding, 2014) and BayesRC (MacLeod et al., 2016) were proposed to allocate markers into different classes, and the superior performances of these methods were observed. Most of these methods, however, focus on single-trait analysis and have limitations in the priors used for marker effects. Thus, in this study, we presented the multi-class Bayesian Alphabet methods, which can perform both single-trait and multiple-trait analysis and provide multiple Bayesian Alphabet priors for markers allocated to different classes.

The effect of allocating markers into different classes on genomic prediction has been studied in some previous studies (Morota et al., 2014; Speed and Balding, 2014; MacLeod et al., 2016; Xu et al., 2020). Different effect-size prior distributions are assigned to molecular markers being split into multiple classes based on genetic architectures. In this paper, we use genome annotation to allocate markers into multiple classes. Note that given the different biological information, the number of classes and markers inside each class might be different. For example, we can use the GWAS results, like Zhang et al. (2014) and Gao et al. (2015), to allocate markers into two classes: one with identified causal variants and another class with the remaining markers.

The comparisons between prediction accuracies from multi-class and conventional Bayesian Alphabet are shown in **Tables 1**–**3**. Multi-class Bayesian Alphabet performs consistently equivalent to or better than conventional Bayesian Alphabet in both real and simulated datasets. The different performances of the multi-class Bayesian Alphabet may be caused by the genetic architectures across different genomic regions in the datasets. The methods that allocate markers into different classes outperform the conventional methods because these methods allow different priors on marker effects according to genetic architectures (Speed and Balding, 2014; MacLeod et al., 2016). If genetic architectures

are similar across the SNP classes, assigning different priors will not bring significant improvement. For example, in comparisons without much difference between multi-class and conventional methods, e.g., multi-class BayesCπ using genome annotation information in the real pig data analysis, relatively small range (0.0001 to 0.03) for the estimated marker effect variances was observed across SNP classes. However, in comparison with significant differences, e.g., multi-class BayesCπ using chromosome information in the real pig data analysis, relatively large range (0.0001 to 0.15) for the estimated marker effect variances was observed across SNP classes.

Our multi-class Bayesian Alphabet method allows the coexist of the different types of priors in one model. For example, a BayesA prior can be assigned to one SNP class and a BayesCΠ prior to another. In addition, the same marker can be allocated to multiple SNP classes. Compared to other methods that allocate markers into multiple classes, our multi-class Bayesian Alphabet provides more flexibility for model construction given the genetic architectures of the traits of interest and increasing biological knowledge on the genome for both single-trait and multiple-trait analysis. However, a naive comparison among multiple multi-class Bayesian Alphabet methods is computationally intensive. For example, with 6 SNP classes and 5 types of prior, there are 5^6 possible combinations, and the computational intensity increases dramatically as the number of SNP classes grows. An efficient algorithm to choose biologically meaningful priors for each SNP class, is needed. In addition, biological knowledge generated from other projects may help to narrow down the prior candidates for each SNP class. In our multi-class Bayesian Alphabet methods tested in this paper, where computational intensities are similar to conventional methods, equivalent or better performances are consistently observed. Given that our single-trait and multiple-trait multi-class Bayesian Alphabet methods are biologically meaningful and their implementation is available in an open-source package, we expect it would be widely adopted for genomic prediction.

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

HC conceived the study. HC and ZW implemented the method. ZW undertook the analysis and wrote the draft. Both authors contributed to the final version of the manuscript, read, and approved the final manuscript.

FUNDING

This work was supported by the United States Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive Grant Nos. 2018-67015-27957 and 2021-67015-33412.

ACKNOWLEDGMENTS

We want to thank Tianjing Zhao for her suggestions on the SNP allocation and Jiayi Qu for her advice on the manuscript writing.

REFERENCES

Abdollahi-Arpanahi, R., Morota, G., Valente, B. D., Kranis, A., Rosa, G. J. M., and Gianola, D. (2016). Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. *Genet. Select. Evol.* 48:10. doi: 10.1186/s12711-016-0187-z

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., Livera, A. M. D., Yetukuri, L., et al. (2016). Genomic prediction of coronary heart disease. *Eur. Heart J.* 37, 3267–3278. doi: 10.1093/eurheartj/ehw450

Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet.* 10:e1004137. doi: 10.1371/journal.pgen.1004137

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 (Bethesda).* 9, 3691–3702. doi: 10.1534/g3.119.400498

Chen, C., Steibel, J. P., and Tempelman, R. J. (2017). Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. *Genetics* 206, 1791–1806. doi: 10.1534/genetics.117.202259

Cheng, H., Fernando, R., and Garrick, D. (2018a). "Jwas: Julia implementation of whole-genome analyses software," in *Proceedings of the World Congress on Genetics Applied to Livestock Production* (Auckland).

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., and Fernando, R. (2018b). Genomic prediction from multiple-trait bayesian regression methods using mixture priors. *Genetics* 209, 89–103. doi: 10.1534/genetics.118.300650

Do, D. N., Janss, L. L. G., Jensen, J., and Kadarmideen, H. N. (2015). SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. *J. Anim. Sci.* 93, 2056–2063. doi: 10.2527/jas.2014-8640

Duarte, J. L. G., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., and Steibel, J. P. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* 15:246. doi: 10.1186/1471-2105-15-246

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomart. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

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

Edwards, S. M., Sarensen, I. F., Sarup, P., Mackay, T. F. C., and Sarensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203, 1871–1883. doi: 10.1534/genetics.116.187161

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129. doi: 10.3168/jds.2011-5019

Fernando, R. (1998). "Genetic evaluation and selection using genotypic, phenotypic and pedigree information," in *6th Wld. Cong. Genet. App. Liv. Prod.* (Armidale, NSW), 329–336.

Fernando, R. L., and Garrick, D. (2013). Genome-wide association studies and genomic prediction. *Methods Mol. Biol.* 1019, 237–274. doi: 10.1007/978-1-62703-447-0_10

Fonseca, P., Suarez-Vega, A., Marras, G., and Canovas, A. (2020). GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *GigaScience* 9:giaa149.

Gao, N., Li, J., He, J., Xiao, G., Luo, Y., Zhang, H., et al. (2015). Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. *BMC Genetics* 16:120. doi: 10.1186/s12863-015-0278-9

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Gianola, D., and Fernando, R. L. (2020). A multiple-trait bayesian lasso for genome-enabled analysis and prediction of complex traits. *Genetics* 214, 305–331. doi: 10.1534/genetics.119.302934

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Harris, B., Creagh, F., Winkelman, A., and Johnson, D. (2011) Experiences with the illumina high density bovine beadchip. *Interbull Bulletin*.

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., et al. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biol.* 10:201. doi: 10.1186/gb-2009-10-1-201

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

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/S0016672308009981

Howe, K. L., Achuthan, P., Allen, J., Allen, J., and Alvarez-Jarreta, J. (2020). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891.

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., et al. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14:59. doi: 10.1186/1471-2164-14-59

MacLeod, I. M., Bowman, P. J., Jagt, C. J. V., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6

Martini, J. W. R., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J. C., et al. (2017). Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics* 18:3. doi: 10.1186/s12859-016-1439-1

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

Moeinizade, S., Kusmec, A., Hu, G., Wang, L., and Schnable, P. S. (2020). Multi-trait genomic selection methods for crop improvement. *Genetics* 215, 931–945. doi: 10.1534/genetics.120.303305

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics* 15:109. doi: 10.1186/1471-2164-15-109

Rainer, J., Gatto, L., and Weichenberger, C. X. (2019). ensembldb: an R package to create and use ensembl-based annotation resources. *Bioinformatics* 35, 3151–3153. doi: 10.1093/bioinformatics/btz031

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 24, 1550–1557. doi: 10.1101/gr.169375.113

Strandén, I., and Garrick, D. J. (2009). Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92, 2971–2975. doi: 10.3168/jds.2008-1929

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tutar, Y. (2012). Pseudogenes. *Comp. Funct. Genomics* 2012:424526. doi: 10.1155/2012/424526

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

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic selection methods for crop improvement: current status and prospects. *Crop J.* 6, 330–340. doi: 10.1016/j.cj.2018.03.001

Xu, L., Gao, N., Wang, Z., Xu, L., Liu, Y., Chen, Y., et al. (2020). Incorporating genome annotation into genomic prediction for carcass traits in Chinese simmental beef cattle. *Front. Genet.* 11:481. doi: 10.3389/fgene.2020.00481

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., et al. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* 9:e93017. doi: 10.1371/journal.pone.0093017

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467. doi: 10.1038/ncomms1467

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 in Genetics

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fgene.2021.728764

Towards a Cost-Effective Implementation of Genomic Prediction Based on Low Coverage Whole Genome Sequencing in Dezhou Donkey

Changheng Zhao[1], Jun Teng[1], Xinhao Zhang[1,2], Dan Wang[1], Xinyi Zhang[1], Shiyin Li[1], Xin Jiang[1], Haijing Li[2], Chao Ning[1] and Qin Zhang[1]**

[1]Shandong Provincial Key Laboratory of Animal Biotechnology and Disease Control and Prevention, College of Animal Science and Veterinary Medicine, Shandong Agricultural University, Tai'an, China, [2]National Engineering Research Center for Gelatin-based TCM, Dong-E E-Jiao Co., Ltd., Dong'e County, China

OPEN ACCESS

Edited by:
Ruidong Xiang,
The University of Melbourne, Australia

Reviewed by:
Irene Van Den Berg,
Agriculture Victoria, Australia
Johannes W. R. Martini,
International Maize and Wheat Improvement Center, Mexico

***Correspondence:**
Chao Ning
ningchao@sdau.edu.cn
Qin Zhang
qzhang@sdau.edu.cn

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

Received: *22 June 2021*
Accepted: *20 September 2021*
Published: *03 November 2021*

Citation:
Zhao C, Teng J, Zhang X, Wang D, Zhang X, Li S, Jiang X, Li H, Ning C and Zhang Q (2021) Towards a Cost-Effective Implementation of Genomic Prediction Based on Low Coverage Whole Genome Sequencing in Dezhou Donkey. Front. Genet. 12:728764. doi: 10.3389/fgene.2021.728764

Low-coverage whole genome sequencing is a low-cost genotyping technology. Combined with genotype imputation approaches, it is likely to become a critical component of cost-effective genomic selection programs in agricultural livestock. Here, we used the low-coverage sequence data of 617 Dezhou donkeys to investigate the performance of genotype imputation for low-coverage whole genome sequence data and genomic prediction based on the imputed genotype data. The specific aims were as follows: 1) to measure the accuracy of genotype imputation under different sequencing depths, sample sizes, minor allele frequency (MAF), and imputation pipelines and 2) to assess the accuracy of genomic prediction under different marker densities derived from the imputed sequence data, different strategies for constructing the genomic relationship matrixes, and single-vs. multi-trait models. We found that a high imputation accuracy (>0.95) can be achieved for sequence data with a sequencing depth as low as 1x and the number of sequenced individuals ≥400. For genomic prediction, the best performance was obtained by using a marker density of 410K and a G matrix constructed using expected marker dosages. Multi-trait genomic best linear unbiased prediction (GBLUP) performed better than single-trait GBLUP. Our study demonstrates that low-coverage whole genome sequencing would be a cost-effective approach for genomic prediction in Dezhou donkey.

Keywords: dezhou donkey, low coverage whole genome sequencing, genotype imputation, genomic prediction, GBLUP

INTRODUCTION

Dezhou donkey, originating from Dezhou area, Shandong Province, China, is one of the major donkey breeds in China. It is famous for its large body size (and thus good meat production ability) and excellent skin quality (for producing donkey-hide gelatin). It has been introduced as a breeding stock into many areas of China and has also brought considerable economic benefits to farmers (Wang et al., 2020a). Therefore, Dezhou donkey plays an important role in the donkey industry in

China. However, selective breeding based on animal breeding theory had never been practiced in Dezhou donkey in the past. In recent years, along with the increased importance of the donkey industry in livestock agriculture in China, selective breeding is gradually becoming an important issue in donkey production, and some breeding work is being carried out in the Dezhou donkey population.

Starting with the pioneered work of Meuwissen et al. (2001), genomic selection (GS) has been widely used in selective breeding in almost all major farm animal species and has brought great increases of genetic progress and economic benefit for many animal breeding industries (Schaeffer, 2006; Stock and Reents, 2013; Wiggans et al., 2017). Typically, GS is carried out using a high-density (or medium-density) single-nucleotide polymorphism (SNP) array. Many commercial SNP arrays have been developed for major farm animal species (Stock and Reents, 2013). However, there are still some species, such as donkey, for which no such array is available, which inhibits the application of GS in these species.

Recently, along with the rapid development of next-generation sequencing technology and reduction of sequencing cost, GS using genotypes revealed by whole genome sequencing (WGS), instead of SNP array, has drawn interests of animal GS community (Hickey 2013; Daetwyler et al., 2014; Georges 2014). The motivations of using whole genome sequence data are to increase the selection accuracy, to facilitate GS across breeds/populations, and to improve persistence of accuracy across generations (Meuwissen and Goddard, 2010; Hayes et al., 2013). To capture the whole genome variants, a sequencing depth of 10x to 20x is generally required (Rashkin et al., 2017; Jiang et al., 2019). However, at present, sequencing with such depth is still too expensive for a large-scale GS application. An alternative is to perform low-coverage whole genome sequencing (lcWGS) at about 1x or less, and then recovering the missing genotypes by imputation to ensure that all individuals have genotypes for a shared set of variants. This approach has been used in human and some animal species for genome-wide association studies and genomic selection/prediction and proved to be a feasible alternative to normal sequencing (Pasaniuc et al., 2012; Nicod et al., 2016; Liu et al., 2018; Zhang et al., 2021). Since the cost of lcWGS can even be lower than that of a SNP array (e.g., in China, the current price for sequencing a cattle genome at 1x is about ¥ 250 RMB per sample, while the price for genotyping with the Neogen GGP Bovine 100k SNP array is ¥ 280 RMB per sample), it is considered as a cost-effective genotyping approach for GS [referred to as GS 2.0 by Hickey (2013)].

A critical issue of lcWGS-based GS is the accuracy of imputation of missing genotypes, which is affected by several factors, such as sequencing depth, sample size, minor allele frequency (MAF), and imputation method. A number of imputation methods for lcWGS data have been proposed (Davies et al., 2016; Ros-Freixedes et al., 2017; Hui et al., 2020). However, most of these methods require a high-density reference haplotype panel, which is not available for most animal species, including donkey. Davies et al. (2016) proposed a method called STITCH for imputation without requiring a reference haplotype panel. It makes use of the fact that SNPs in sequences are not independent of each other, and it constructs founder haplotypes directly from the sequencing read data and then perform imputation based on a hidden Markov model. This method provides an opportunity of using lcWGS technology for species for which a reference haplotype panel is not available.

In this study, we evaluated the imputation accuracy of lcWGS data with respect to different sequencing depths, sample sizes, MAFs, and imputation pipelines using 617 Dezhou donkey animals that were sequenced with an average depth of 3.5x. We then used the imputed genotypes to investigate the performance of genomic selection for birth weight (BW) and weaning weight (WW) in the Dezhou donkey population under different marker densities, strategies for constructing genomic relationship matrices, and single-vs. two-trait models.

MATERIALS AND METHODS

Animals

The animals used in this study were from the Dong-E E-Jiao Donkey Farm in Shandong Province, China. Animals that had records on both BW and WW were selected. These animals along with their known parents formed the study population for this research, which consisted of 617 animals, of which 594 had records on both traits. These 594 animals (303 males and 291 females) were born between 2015 and 2019. The animals were weaned at 6 months after birth, and their weaning weight was measured at the age of 6 ± 1 month. Weaning weight recorded outside this age range was regarded as invalid record. The means and standard deviations of the two traits were 30.507 ± 4.235 kg (ranging from 15.0 to 42.2 kg) and 116.752 ± 18.227 kg (ranging from 63.5 to 165.5 kg), respectively.

Blood samples were collected from all these animals. Total DNA was isolated using the QIAamp DNA Investigator Kit (QIAGEN, Hilden, Germany) and following the manufacturer's instruction. DNA quality was evaluated by spectrophotometry and agarose gel electrophoresis.

All of the above experiments were carried out according to the guideline of the experimental animal management of Shandong Agricultural University (SDAUA-2018-018).

Low-Coverage Whole Genome Sequencing

DNA templates were ultrasonically sheared using a Covaris E220 (Covaris, Woburn, MA, United States) to yield to 150-bp fragments and then prepared for sequencing libraries following the workflow of the NEBNext Ultra DNA Library Preparation Protocol. Multiple Ampure Bead XP cleanups (Beckman Coulter, Brea, CA, United States) were conducted to remove any adapter dimer that might have developed. The quality and concentration of libraries were determined on an Agilent Bioanalyzer 2,100 (Agilent Technologies, Santa Clara, CA). The genomic library for each sample was PE150 sequenced using the Illumina NovaSeq 6,000 sequencing system.

Read quality was assessed using the FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) with focus on base quality scores ($q > 30$), GC content (skewness <5%), N content

(<5%), and sequence duplication levels (<100). The resulting data reached a nucleotide length of 150 bp and a base quality score of higher than 30 and were aligned to the donkey reference genome (Wang et al., 2020b) by BWA (Li and Durbin, 2009). SAMtools (Li et al., 2009) was used to transfer the formats and sort and index files. The 617 animals had an average sequencing depth of 3.5x (ranging from 1.9x to 6.4x) (**Supplementary Figure S1**).

Pipelines for Genotype Imputation

We compared two imputation pipelines, i.e., Bcftools + Beagle and BaseVar + STITCH. In the first pipeline, we called SNPs using Bcftools (Li, 2011) and then conducted genotype imputation using Beagle v4.1 (Browning and Browning, 2016). In the second pipeline, we called SNPs using BaseVar (Liu et al., 2018) and imputed the missing genotypes (with probabilities) using STITCH v1.6.3. The resulted SNP data from both pipelines were filtered with MAF ≥0.01 and a Hardy–Weinberg equilibrium (HWE) p-value > 1e-6 using PLINK (Chang et al., 2015).

Evaluation of Imputation Accuracy

We evaluated the imputation accuracy under different sequencing depths, sample sizes, and MAFs using the sequence data of additional 18 Dezhou donkey animals provided by the Donkey Research Institute, Liaocheng University, Shandong Province, China. The average sequencing depth of the 18 animals was 13.5x (ranging from 11.2x to 16.3x). Chromosomes 1, 19, and 30, which represented the long, short, and medium chromosomes among the donkey chromosomes, respectively, were chosen to evaluate the imputation accuracy. The imputation accuracy was measured with two criteria, i.e., genotypic concordance and genotypic accuracy. Genotypic concordance is defined as the proportion of correctly imputed genotypes (Fridley et al., 2010), and genotypic accuracy is defined as squared Pearson correlation coefficient (r^2) between expected dosages (posterior expectation of the imputed allele dosages) and typed genotypes (Browning and Browning, 2009). To evaluate the imputation accuracy for different sequencing depths, in addition to the original sequence data with an average depth of 3.5x, we randomly sampled reads from the sequencing read data to generate sequence data with different lower sequencing depths (0.5x, 1x, 1.5x, and 2x) using Picard (https://broadinstitute.github.io/picard/). For the depths of 0.5x, 1x, and 1.5x, three repeated samplings were performed. To test the effect of sample size (number of low coverage sequenced individuals) on imputation accuracy, three different sample sizes (200, 400, and 617) were considered. The samples with sizes of 200 and 400 were randomly sampled from the 617 animals, and three repeated samplings were performed. To test the effect of MAF on imputation accuracy, we restored the SNPs that were previously filtered out with MAF >0.01 and divided the SNPs into 15 MAF bins: (0–0.001), (0.001–0.002), (0.002–0.005), (0.005–0.01), (0.01–0.02), (0.02–0.05), (0.05–0.1), (0.1–0.15), (0.15–0.2), (0.2–0.25), (0.25–0.3), (0.3–0.35), (0.35–0.4), (0.4–0.45), and (0.45–0.5). The average imputation accuracy in each bin was calculated.

Genomic Prediction

The imputation-based sequence data was used to investigate the performance of genomic prediction using the 594 animals having records on both BW and WW. The genomic estimated breeding values (GEBVs) were obtained by using the genomic best linear unbiased prediction (GBLUP) method (VanRaden, 2008) under single-trait model as well as two-trait model.

The single-trait GBLUP model is as follows:

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

where $\mathbf{y}$ is the vector of observed phenotypes of BW or WW; $\mathbf{b}$ is the vector of fixed effects, which include the effects of sex, year-seasons when the trait was measured (years for BW: 2015–2019, years for WW: 2016–2019, and four seasons each year), and age (in days, as covariate, for WW) when the trait was measured; $\mathbf{u}$ is the vector of genomic breeding values with distribution of $N(0, \mathbf{G}\sigma_a^2)$, where σ_a^2 is the additive genetic variance and $\mathbf{G}$ is the genomic relationship matrix; $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices for $\mathbf{b}$ and $\mathbf{u}$, respectively; and $\mathbf{e}$ is the vector of random residuals with distribution of $N(0, \mathbf{I}\sigma_e^2)$.

The two-trait GBLUP model is as follows:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & 0 \\ 0 & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

where the meanings of the vectors and matrices are the same as those in the single-trait model with the subscripts 1 and 2 referring BW and WW, respectively. It was assumed that $\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \sim N(0, \mathbf{G}\otimes\mathbf{M})$, where $\mathbf{M} = \begin{bmatrix} \sigma^2_{\mathbf{a}_1} & \sigma_{\mathbf{a}_{12}} \\ \sigma_{\mathbf{a}_{12}} & \sigma^2_{\mathbf{a}_2} \end{bmatrix}$ is the variance–covariance matrix of the genomic breeding values of the two traits, and $\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \sim N(0, \mathbf{I}\otimes\mathbf{R})$, where $\mathbf{R} = \begin{bmatrix} \sigma^2_{\mathbf{e}_1} & \sigma_{\mathbf{e}_{12}} \\ \sigma_{\mathbf{e}_{12}} & \sigma^2_{\mathbf{e}_2} \end{bmatrix}$ is the residual variance–covariance matrix of the two traits.

Since STITCH provides for each SNP and each individual the imputed genotype (the most likely genotype) as well as the expected genotype dosages (posterior expectation of the genotype dosages), the $\mathbf{G}$ matrix can be constructed using either the imputed genotypes or the expected genotype dosages. The genotype-based $\mathbf{G}$ matrix [denoted as $\mathbf{G}(\mathrm{g})$] was constructed using the method of VanRaden (2008) as follows:

$$\mathbf{G}(\mathrm{g}) = \mathbf{WW}' \Big/ \sum 2p_j\left(1 - p_j\right)$$

where, $\mathbf{W}$ is the centralized maker genotype matrix with its ij th element equal to

$$w_{ij} = m_{ij} - 2p_j$$

where m_{ij} (= 2, 1, or 0) is the original genotype of individual i for SNP j, and p_j is the minor allele frequency of SNP j.

For constructing $\mathbf{G}$ using expected dosages [denoted as $\mathbf{G}(\mathrm{d})$], following the idea of the formula for $\mathbf{G}(\mathrm{g})$, we proposed the following formula:

$$\mathbf{G}(\mathrm{d}) = \mathbf{DD}^{'}/s_d$$

where, $\mathbf{D}$ is the centralized marker dosage matrix whose elements are zero-centered expected dosages. s_d is the sum of variances for every column of $\mathbf{D}$.

To evaluate the effect of marker density on the performance of genomic prediction, we used four levels of marker densities

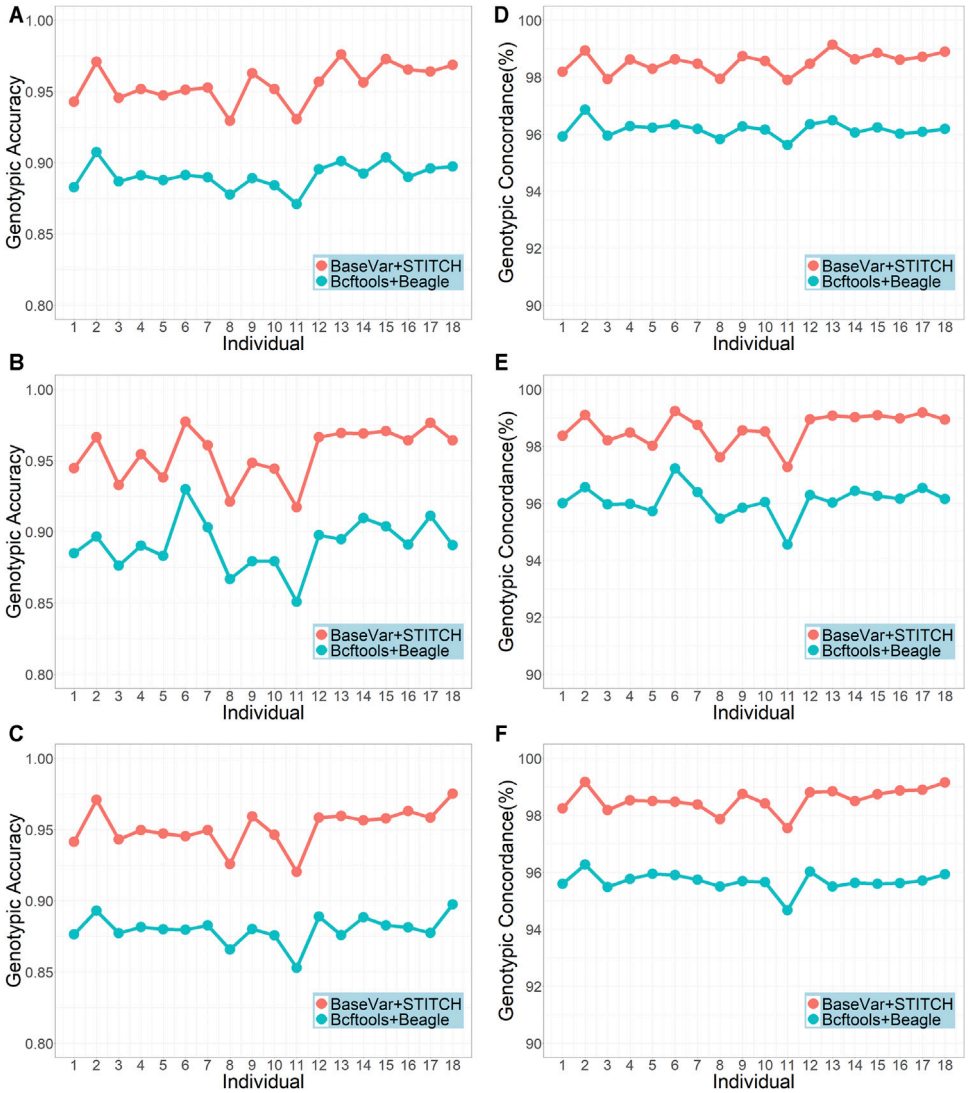

FIGURE 1 | Genotypic accuracy and genotypic concordance using the two imputation pipelines (sample size = 617 and average sequencing depth = 3.5x). **(A–C)** represent genotype accuracy for chromosomes 1, 19, and 30, respectively; **(D–F)** represent genotype concordance for chromosomes 1, 19 and 30, respectively.

to construct the **G** matrices. From the original sequence data with an average depth of 3.5x, we obtained 2.3M SNPs after imputation and quality control. We then reduced the marker density by down-sampling SNPs from the 2.3M SNPs. We applied linkage disequilibrium (LD) pruning with three LD levels: r^2 = 0.2, 0.4, and 0.8, by PLINK (Chang et al., 2015), which produced 130, 220, and 410K SNPs, respectively.

In addition, we also evaluated the performance of genomic prediction using the 1x sequencing data, which was sampled from the original sequence data and contained 1.4M SNPs after imputation and quality control.

We used GMAT (Wang et al., 2020a) to construct the **G** matrix. The variance and covariance components involved in the models and GEBVs were estimated by AI-REML using the DMU package (Madsen et al., 2014; http://dmu.agrsci.dk).

Cross-Validation

In this study, a 12-fold cross-validation (CV) was applied to assess the accuracy of the genomic prediction. The 594 animals were divided into 12 subsets. One of them was taken in turn to be used as a validation population, and the remaining 11 subsets used as a training population. For the two-trait model analysis, we left out the observations on both BW and WW for the animals in the validation set and calculated their GEBVs for both traits simultaneously. The accuracy of genomic prediction for the validation animals was assessed by $r_{y_c,\mathrm{GEBV}}$, the correlation between corrected phenotypic values (y_c) and GEBVs. The corrected phenotype for each animal was calculated as the original phenotypic value corrected for fixed effects [sex, year-season, and age (for WW)], which were estimated by conventional BLUP using the DMU package (Madsen et al., 2014; http://dmu.agrsci.dk). The model for conventional BLUP

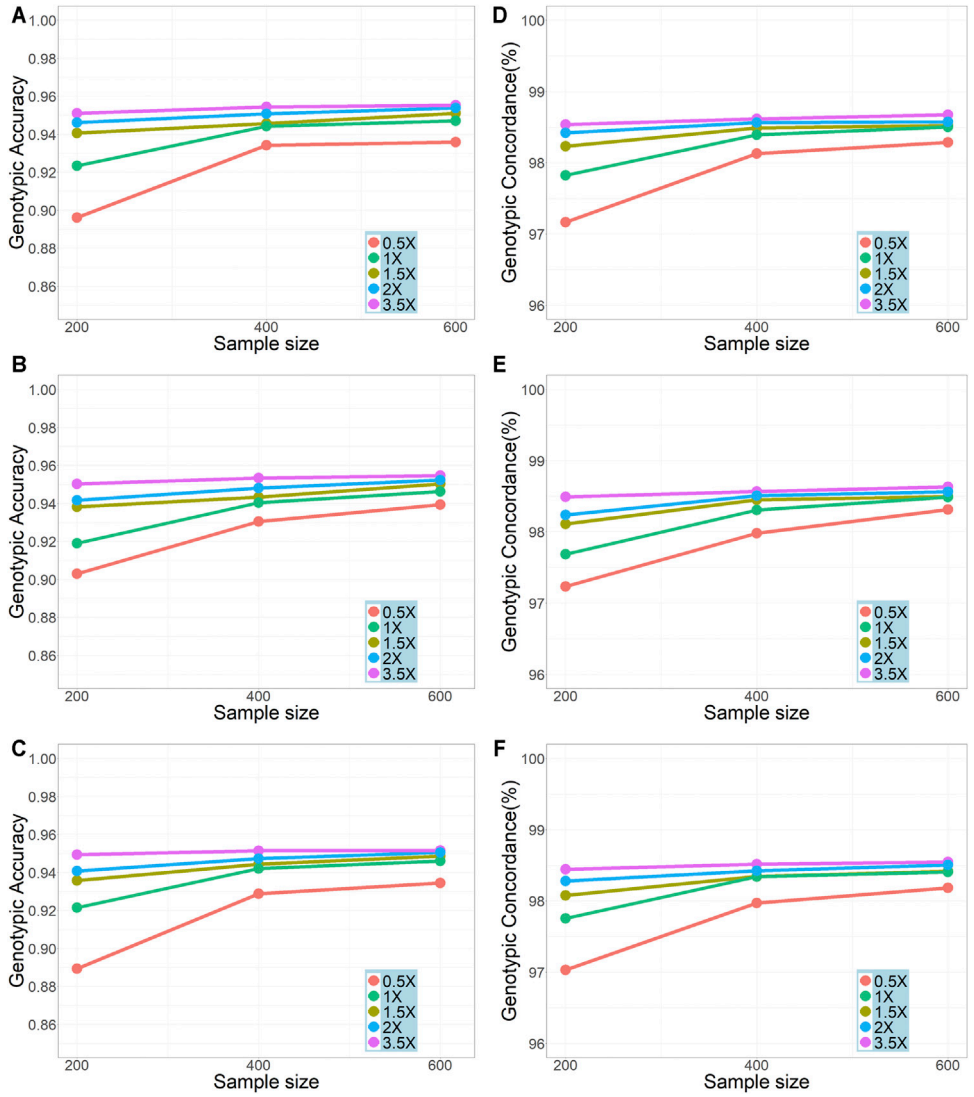

FIGURE 2 | Effects of sample size and sequencing depth on imputation genotypic accuracy and genotypic concordance using the pipeline of BaseVar + STITCH. **(A–C)** represent genotype accuracy for chromosomes 1, 19, and 30, respectively; **(D–F)** represent genotype concordance for chromosomes 1, 19, and 30, respectively.

was the same as that for GBLUP except that the **G** matrix was replaced by the pedigree-based **A** matrix. The bias of predictions was assessed by the regression of y_c on GEBV ($b_{y_c,GEBV}$), with $b_{y_c,GEBV} = 1$ indicating unbiased prediction (Su et al., 2012).

RESULTS

Accuracies of Genotype Imputation

Comparison of Different Pipelines

The two genotype imputation pipelines, BaseVar + STITCH and Bcftools + Beagle, were compared using the original sequencing data of the 617 animals with an average sequencing depth of 3.5x. **Figure 1** shows that the BaseVar + STITCH pipeline was remarkably better than the Bcftools + Beagle pipeline. The average genotypic accuracy from BaseVar + STITCH was about 0.06 higher than that from Bcftools + Beagle, and the average genotypic concordance was about 0.02 higher. Therefore, the BaseVar + STITCH pipeline was used for the subsequent analyses.

The Effects of Sample Size and Sequencing Depth

We compared the genotypic accuracy and genotypic concordance for imputation with different sample sizes (200, 400, and 600) and sequencing depths (0.5x, 1x, 1.5x, 2x, and 3.5x) (**Figure 2**). In all scenarios, the genotype accuracies were over 0.90 (with only one exception on chromosome 30 in the scenario of sequencing depth = 0.5x and sample size = 200) and the genotypic concordances were over 0.97. In general, as expected, the genotypic accuracy and genotypic concordance increased with the increase of sample size and sequencing depth. The improvement of imputation accuracy was most obvious when the sample size was

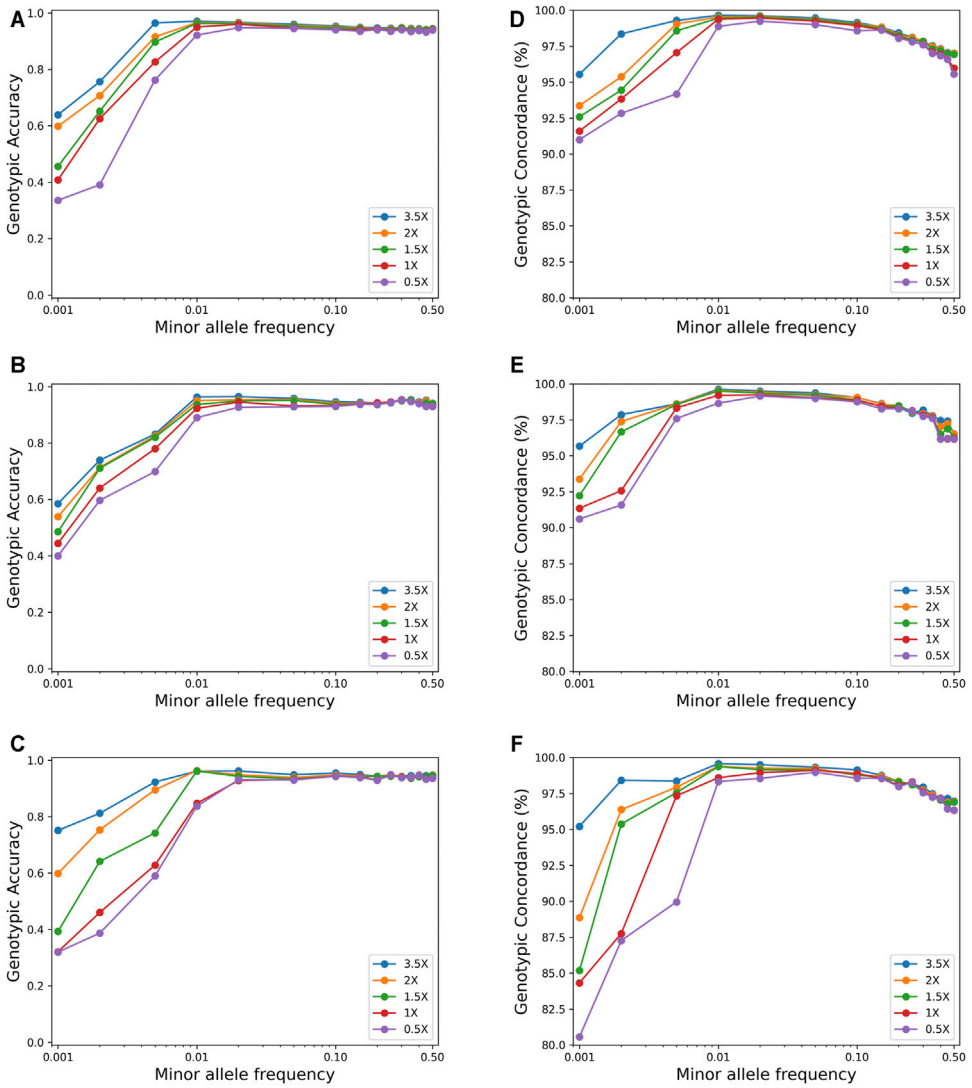

FIGURE 3 | Effects of minor allele frequency on imputation genotype accuracy and genotype concordance using the pipeline of BaseVar + STITCH (sample size = 617). **(A–C)** represent genotype accuracy for chromosomes 1, 19, and 30, respectively; **(D–F)** represent genotype concordance for chromosomes 1, 19, and 30, respectively.

increased from 200 to 400 and the sequencing depth increased from 0.5x to 1x. For sequencing depths of 0.5x, 1x, and 1.5x, the results from the three repeatedly sampled data were almost the same (see **Supplementary Table S1** for chromosome 19 and sample size of 200), so did the results from the repeated samples of sizes 200 and 400 (see **Supplementary Table S2**). It should be noted that, with a sample size of ≥400, a genotypic accuracy greater than 0.94 and a genotypic concordance greater than 0.98 could be achieved even when the sequencing depth was as low as 1x. However, with a sequencing depth of 0.5x, even for a sample size of 600, the genotype accuracy was less than 0.94.

The Effect of MAF

Figure 3 shows the effect of MAF on imputation accuracy for a sample size of 600. For SNPs with MAF <0.01, both the genotypic accuracy and the genotypic concordance were greatly affected by MAF, and the accuracy increased rapidly with the increase of MAF. However, for SNPs with MAF >0.01, the imputation accuracy was not affected by MAF, while the genotypic concordance decreased slightly with the increase of MAF.

Variance Component Estimation

Table 1 presents the estimates of variance components and heritabilities based on the single-trait model with the two types of **G** matrix [**G**(g) and **G**(d)] constructed using five different marker sets (130, 220, 410K, and 2.3M from the 3.5x sequence data and 1.4M from the 1x sequence data). For the 3.5x sequence data, the estimates under the four marker sets were very similar, with the additive variance and heritability estimates from the 2.3M marker set being consistently slightly smaller than those

TABLE 1 | Estimates of variance components and heritabilities and their standard errors (in brackets) under single-trait model using different marker sets and different **G** matrices for birth weight (BW) and weaning weight (WW).

Marker set[a]	Genotype-based G matrix			Expected dosage-based G matrix		
	σ_a^2	σ_e^2	h^2	σ_a^2	σ_e^2	h^2
			BW			
130K	10.266 (2.448)	8.024 (1.730)	0.561 (0.108)	10.100 (2.409)	8.030 (1.730)	0.557 (0.108)
220K	10.424 (2.456)	7.904 (1.728)	0.569 (0.108)	10.253 (2.417)	7.912 (1.727)	0.564 (0.108)
410K	10.750 (2.457)	7.656 (1.709)	0.584 (0.106)	10.581 (2.420)	7.664 (1.709)	0.580 (0.106)
2.3M	10.166 (2.424)	8.154 (1.703)	0.555 (0.107)	10.011 (2.388)	8.159 (1.702)	0.551 (0.107)
1.4M	9.456 (2.374)	8.732 (1.692)	0.520 (0.107)	9.316 (2.340)	8.739 (1.692)	0.516 (0.107)
			WW			
130K	68.419 (26.549)	139.977 (21.542)	0.328 (0.115)	68.223 (26.116)	140.874 (21.553)	0.326 (0.115)
220K	69.136 (26.566)	140.659 (20.358)	0.330 (0.115)	68.919 (26.549)	141.777 (21.542)	0.327 (0.115)
410K	69.866 (26.423)	141.046 (21.431)	0.331 (0.115)	69.670 (25.831)	141.328 (21.421)	0.330 (0.114)
2.3M	62.042 (25.269)	146.029 (20.907)	0.298 (0.112)	61.083 (24.888)	146.074 (20.900)	0.295 (0.111)
1.4M	54.305 (23.542)	152.396 (20.072)	0.263 (0.106)	53.514 (23.204)	152.422 (20.067)	0.260 (0.105)

[a]*Marker sets 130K–2.3M were derived from the original sequence data with an average depth of 3.5x; marker set 1.4M was from sequence data with a depth of 1x.*
σ_a^2, additive genetic variance; σ_e^2, residual variance; h2, heritability.

TABLE 2 | Estimates of variance (covariance) components, heritabilities, and genetic correlation and their standard errors (in brackets) under two-trait models using the 410K marker set and expected dosage-based G matrix.

Trait	σ_a^2	σ_e^2	h^2	Cov_a	Cov_e	r_g	r_p
Birth weight	11.769 (2.467)	7.016 (1.687)	0.627 (0.102)	27.399 (6.796)	9.839 (4.806)	0.839 (0.076)	0.588
Weaning weight	90.728 (25.747)	122.553 (19.855)	0.425 (0.105)				

σa2, additive genetic variance; σe2, residual variance; h2, heritability; Cova, additive genetic covariance between BW and WW; Cove, residual covariance between BW and WW; rg, genetic correlation (= $\frac{Cov_a}{\sigma_a(BW)\times\sigma_a(WW)}$); rp, phenotypic correlation (= $\frac{Cov_a+Cov_e}{\sqrt{\sigma_a^2(BW)+\sigma_e^2(BW)}\times\sqrt{\sigma_a^2(WW)+\sigma_e^2(WW)}}$).

TABLE 3 | Accuracies and biases of genomic prediction and their standard errors (in brackets) under single-trait model with different marker sets.

Marker set[a]	Genotype-based G matrix		Expected dosage-based G matrix	
	Accuracy[b]	Bias[c]	Accuracy[b]	Bias[c]
		Birth weight		
130K	0.285 (0.041)	0.063 (0.189)	0.285 (0.041)	0.063 (0.189)
220K	0.290 (0.040)	0.069 (0.186)	0.290 (0.040)	0.069 (0.186)
410K	0.297 (0.039)	0.061 (0.175)	0.297 (0.039)	0.062 (0.176)
2.3M	0.283 (0.039)	0.043 (0.174)	0.283 (0.039)	0.043 (0.174)
1.4M	0.277 (0.040)	0.037 (0.178)	0.277 (0.040)	0.038 (0.178)
		Weaning weight		
130K	0.225 (0.031)	0.163 (0.165)	0.225 (0.031)	0.164 (0.166)
220K	0.226 (0.032)	0.163 (0.166)	0.226 (0.031)	0.163 (0.165)
410K	0.229 (0.031)	0.168 (0.163)	0.229 (0.031)	0.169 (0.163)
2.3M	0.223 (0.032)	0.149 (0.179)	0.223 (0.032)	0.149 (0.179)
1.4M	0.221 (0.031)	0.183 (0.178)	0.221 (0.031)	0.183 (0.178)

[a]*Marker sets 130K–2.3M were derived from the original sequence data with an average depth of 3.5x; marker set 1.4M was from sequence data with a depth of 1x.*
[b]*Accuracy is defined as the correlation between GEBVs and corrected phenotypes (y_c).*
[c]*Bias is defined as 1-regression coefficient of GEBVs on y_c.*

from the other three marker sets. However, the additive variance and heritability estimates from the 1.4M marker set were all smaller than those from the other marker sets. For all marker sets, the estimates of additive genetic variances and heritabilities based on **G**(d) were consistently smaller than those based on **G**(g), although the differences were very small and not significant.

For the two-trait model, the variance and co-variance components of the two traits were estimated based on the dosage-based **G** matrix and the 410K marker set (**Table 2**).

TABLE 4 | Accuracies and biases of genomic prediction and their standard errors (in brackets) under single-trait and two-trait models.

Model[a]	Birth weight		Weaning weight	
	Accuracy[b]	Bias[c]	Accuracy[b]	Bias[c]
Two-trait	0.337 (0.037)	0.020 (0.141)	0.301 (0.038)	0.117 (0.164)
Single-trait	0.297 (0.039)	0.062 (0.176)	0.229 (0.031)	0.169 (0.163)

[a]*For the two-trait model, only the expected dosage-based* ***G*** *matrix constructed using the 410K marker set derived from the 3.5x sequence data was used. For comparison, the results of the single-trait model using the same* ***G*** *matrix is represented here.*
[b]*Accuracy is defined as the correlation between GEBVs and corrected phenotypes (y_c).*
[c]*Bias is defined as 1-regression coefficient of GEBV on y_c.*

The estimates of heritability from the two-trait model (0.627 for BW and 0.425 for WW) were higher than those from the single-trait model (0.580 for BW and 0.330 for WW). The estimate of genetic correlation between BW and WW was 0.839.

Accuracy and Bias of Genomic Prediction

The GEBVs for BW and WW were calculated under the single-trait model and two-trait model, respectively. For the single-trait model, we again considered both types of **G** matrix [**G**(g) and **G**(d)] constructed using the five different marker sets. The average accuracies and biases derived from 12-fold cross-validation are given in **Table 3**. In general, the differences in accuracy and bias between different marker sets were small and not significant, while the 410K marker set resulted in the highest accuracies, and the 1.4M marker set resulted in the lowest accuracies. No differences in prediction accuracy and bias were observed between the two types of **G** matrices. For the two-trait model, only the **G**(d) matrix constructed using the 410K marker set was used (**Table 4**). Compared with the results under the single-trait model with the same **G** matrix, the two-trait model remarkably improved the accuracies (0.337 vs. 0.297 for BW and 0.301 vs. 0.229 for WW) and reduced the biases (0.020 vs. 0.062 for BW and 0.117 vs. 0.169 for WW). The difference tendencies mentioned above were quite consistent across the 12 folds (see **Supplementary Table S3**).

DISCUSSION

Low-coverage whole genome sequencing followed by imputation provides a cost-effective way for genome-wide high-density genotyping, especially for species (such as donkey) for which a SNP array is not available. In this study, we investigated the strategies for genotype imputation and evaluated the performance of genomic prediction using imputation-based sequence data in a donkey population.

Strategies of Imputation for Low-Coverage Sequence Data

Imputation is necessary for lcWGS data due to the high missing rates, which involves two steps, i.e., SNP calling and imputation. A proper pipeline is essential to ensure high imputation performance. In this study, we compared two pipelines, Bcftools + Beagle and BaseVar + STITCH. In the first pipeline, both Bcftools and Beagle have been widely used for SNP calling and imputation for sequence data, respectively. However, it is not clear whether they are suitable for lcWGS data. On the other hand, BaseVar and STITCH were designed specifically for lcWGS data. We demonstrated that BaseVar + STITCH outperformed Bcftools + Beagle (**Figure 1**). Furthermore, we showed that in our Dezhou donkey population, using this pipeline, high imputation accuracy (genotypic accuracy >0.94 and genotypic concordance >98%) can be achieved with a sample size of 400 and a sequencing depth of 1x (**Figure 2**). Similar results were also reported by Zhang et al. (2021). In other words, with a sample size of over 400, a sequencing depth of 1x could be sufficient to ensure high imputation accuracy using BaseVar + STITCH.

Genomic Prediction Using Imputation-Based Sequence Data

Using the imputation-based sequence data, we evaluated the performance of genomic prediction using GBLUP with respect to two types of **G** matrices [**G**(g) and **G**(d)], five different marker sets (130, 220, 410K, and 2.3M derived from the 3.5x sequence data and 1.4M derived from the 1x sequence data), and single-vs. two-trait GBLUP model.

Comparison of the Two Types of G Matrices

We found that the accuracies and biases of genomic prediction derived from the two types of **G** matrices were almost the same in all scenarios. Note that the variance component estimates from the two types of **G** matrices were also very similar. This implicates that for our given data, the two types of **G** matrices did not lead to different results. It remains to be seen whether this results also holds for other data sets.

Comparison of the Five Marker Sets

For the four marker sets from the 3.5x sequence data, the prediction accuracy increased slightly (although not significant) when the marker density increased from 130 to 410K, but did not further increase when the density increased to 2.3M. The densities of 130, 220, and 410K correspond to medium to high density of SNP array, while the 2.3M corresponds to the density of sequence data. Some studies showed that, in the frame of GBLUP, the genomic prediction accuracy could be improved using high-density SNP array compared to using medium-density array (VanRaden et al., 2011; Su et al., 2012; Perez-Enciso et al., 2015), but there were also studies that showed no or very small such improvement (VanRaden et al., 2013; Boison et al., 2017). It has been shown

that, in the frame of GBLUP, using sequence data could hardly improve the accuracy compared with using SNP array (Ober et al., 2012; Perez-Enciso, 2014; van Binsbergen et al., 2015; Frischknecht et al., 2018). However, this does not mean that sequence data is of no value for genomic prediction. Several studies have shown that sequence data would be beneficial when variants are preselected based on, e.g., GWAS or a Bayesian selection model (MacLeod et al., 2016; Hayes and Daetwyler, 2019). In addition, sequence data can be meaningful for cross-breed/population genomic selection (Druet et al., 2014; MacLeod et al., 2016). On the other hand, the prediction accuracies using the 1.4M marker set from the 1x sequence data were slightly lower than those from the 3.5x sequence data. This should be due to the lower imputation accuracy for the 1x sequence data than 3.5x (see **Figure 2**). However, since the reduction in accuracy was rather small, in consideration of the sequencing cost, sequencing at depth of 1x would be a preferred choice for a lcWGS-based genomic selection.

Single-vs. Two-Trait GBLUP Model

Noticeable increases in genomic prediction accuracy were observed when using a two-trait model compared with using a single-trait model. The comparison was made only for the scenario of using an expected dosage-based **G** matrix and the 410k marker set derived from the 3.5x sequence data. However, such advantage should hold for other scenarios. It has been shown in several incidences that a multi-trait model can increase the accuracy of breeding value estimation, either by conventional BLUP or by GBLUP (Calus and Veerkamp, 2011; Jia and Jannink, 2012; Guo et al., 2014), in particular for traits with high genetic correlation, such as the two traits investigated in this study. This increase in accuracy with multi-trait model will be particularly beneficial for the situation where the reference population size is limited.

It should be pointed out that, although the differences in the performance of genomic prediction between different scenarios seemed reasonable, some of the differences were actually not significant, possibly due to the small dataset available for this study. It is the practical situation for some species/breeds/populations for which only a small dataset is available for investigating genomic prediction. Therefore, despite the limitations of having a small dataset, our findings would provide meaningful inspirations for such situations.

CONCLUSION

In this study, we demonstrated that the pipeline BaseVar + STITCH is a good choice for SNP calling and imputation for low-coverage sequence data. A sufficient high imputation accuracy could be achieved for sequence data with a sequencing depth as low as 1x, when the size of the sequencing population is over 400. Thus, lcWGS combined with imputation provides a cost-effective way for whole genome high-density genotyping and can be applied for large-scale genomic selection in farm animals. This is particularly beneficial for those animal species for which a SNP array is not available. In the frame of GBLUP, increasing marker density from a density comparable with a high-density SNP array (e.g., 400K) to sequence density with millions of SNPs did not increase the accuracy of genomic prediction. The multi-trait model GBLUP improves the accuracy of genomic prediction over the single-trait model, which would be particularly meaningful for the situation where the reference population size is limited.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The animal study was reviewed and approved by the experimental animal management of Shandong Agricultural University.

AUTHOR CONTRIBUTIONS

QZ and CN designed the study. CZ, XYZ, XJ, HL and SL collected the sample and performed the experiments. CZ and JT analyzed and interpreted the data. CZ, JT, XHZ, CN, DW and QZ drafted the manuscript.

FUNDING

The study was funded by the Project for Improved Agricultural Breeding of Shandong Province (2019LZGC011), China Postdoctoral Science Foundation (2020M682217), Shandong Provincial Postdoctoral Program for Innovative Talent, Shandong Provincial Natural Science Foundation (ZR2020QC176 and ZR2020QC175), and National Natural Science Foundation of China (32002172).

ACKNOWLEDGMENTS

The authors thank the Supercomputing Center at Shandong Agricultural University for technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.728764/full#supplementary-material

Supplementary Figure S1 | The number of individuals of different sequencing depths. The sequencing was performed in two batches. The first batch consisted of 317 animals and the average sequencing depth was 4.5x; the second batch consisted of 300 animals and the average depth was 2.4x. Taken together, the overall average depth was 3.5x.

REFERENCES

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

Browning, B. L., and Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* 84, 210–223. doi:10.1016/j.ajhg.2009.01.005

Browning, B. L., and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98, 116–126. doi:10.1016/j.ajhg.2015.11.020

Calus, M. P., and Veerkamp, R. F. (2011). Accuracy of Multi-Trait Genomic Selection Using Different Methods. *Genet. Sel. Evol.* 43, 26. doi:10.1186/1297-9686-43-26

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

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., et al. (2014). Whole-genome Sequencing of 234 Bulls Facilitates Mapping of Monogenic and Complex Traits in Cattle. *Nat. Genet.* 46, 858–865. doi:10.1038/ng.3034

Davies, R. W., Flint, J., Myers, S., and Mott, R. (2016). Rapid Genotype Imputation from Sequence without Reference Panels. *Nat. Genet.* 48, 965–969. doi:10.1038/ng.3594

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward Genomic Prediction from Whole-Genome Sequence Data: Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions. *Heredity* 112, 39–47. doi:10.1038/hdy.2013.13

Fridley, B. L., Jenkins, G., Deyo-Svendsen, M. E., Hebbring, S., and Freimuth, R. (2010). Utilizing Genotype Imputation for the Augmentation of Sequence Data. *PLoS One* 5, e11018. doi:10.1371/journal.pone.0011018

Frischknecht, M., Meuwissen, T. H. E., Bapst, B., Seefried, F. R., Flury, C., Garrick, D., et al. (2018). Short Communication: Genomic Prediction Using Imputed Whole-Genome Sequence Variants in Brown Swiss Cattle. *J. Dairy Sci.* 101 (2), 1292–1296. doi:10.3168/jds.2017-12890

Georges, M. (2014). Towards Sequence-Based Genomic Selection of Cattle. *Nat. Genet.* 46, 807–809. doi:10.1038/ng.3048

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of Single-Trait and Multiple-Trait Genomic Prediction Models. *BMC Genet.* 15, 30. doi:10.1186/1471-2156-15-30

Hayes, B., Daetwyler, H., Fries, R., Guldbrandtsen, B., Sando Lund, M., Boichard, D., et al. (2013). *The 1000 Bull Genomes Project toward Genomic Selection from Whole Genome Sequence Data in Dairy and Beef Cattle.* San Diego, CA, USA: Plant and Animal Genome XXI Conference.

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

Hickey, J. M. (2013). Sequencing Millions of Animals for Genomic Selection 2.0. *J. Anim. Breed. Genet.* 130, 331–332. doi:10.1111/jbg.12054

Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L., and Kivisild, T. (2020). Evaluating Genotype Imputation Pipeline for Ultra-low Coverage Ancient Genomes. *Sci. Rep.* 10, 18542. doi:10.1038/s41598-020-75387-w

Jia, Y., and Jannink, J.-L. (2012). Multiple-trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics* 192, 1513–1522. doi:10.1534/genetics.112.144246

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., and Ding, X. (2019). Optimal Sequencing Depth Design for Whole Genome Re-sequencing in Pigs. *BMC Bioinformatics* 20, 556. doi:10.1186/s12859-019-3164-z

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, 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

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., et al. (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* 175, 347–359.e14. doi:10.1016/j.cell.2018.08.016

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting Biological Priors and Sequence Variants Enhances QTL Discovery and Genomic Prediction of Complex Traits. *BMC Genomics* 17, 144. doi:10.1186/s12864-016-2443-6

Madsen, P., Jensen, J., Labouriau, R., Christensen, O. F., and Sahana, G. (2014). "DMU - A Package for Analyzing Multivariate Mixed Models in Quantitative Genetics and Genomics," in Proceedings, 10th World Congress of Genetics Applied to Livestock Production, 2014 (Vancouver, Canada.

Meuwissen, T., and Goddard, M. (2010). Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* 185, 623–631. doi:10.1534/genetics.110.116590

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:10.1093/genetics/157.4.1819

Nicod, J., Davies, R. W., Cai, N., Hassett, C., Goodstadt, L., Cosgrove, C., et al. (2016). Genome-wide Association of Multiple Complex Traits in Outbred Mice by Ultra-low-coverage Sequencing. *Nat. Genet.* 48, 912–918. doi:10.1038/ng.3595

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., et al. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*. *Plos Genet.* 8 (5), e1002685. doi:10.1371/journal.pgen.1002685

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., et al. (2012). Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-wide Association Studies. *Nat. Genet.* 44, 631–635. doi:10.1038/ng.2283

Pérez-Enciso, M. (2014). Genomic Relationships Computed from Either Next-Generation Sequence or Array SNP Data. *J. Anim. Breed. Genet.* 131, 85–96. doi:10.1111/jbg.12074

Pérez-Enciso, M., Rincón, J. C., and Legarra, A. (2015). Sequence- vs. Chip-Assisted Genomic Selection: Accurate Biological Information Is Advised. *Genet. Sel. Evol.* 47, 43. doi:10.1186/s12711-015-0117-5

Rashkin, S., Jun, G., Chen, S., and Abecasis, G. R. (2017). Optimal Sequencing Strategies for Identifying Disease-Associated Singletons. *Plos Genet.* 13, e1006811. doi:10.1371/journal.pgen.1006811

Ros-Freixedes, R., Gonen, S., Gorjanc, G., and Hickey, J. M. (2017). A Method for Allocating Low-Coverage Sequencing Resources by Targeting Haplotypes rather Than Individuals. *Genet. Sel. Evol.* 49, 78. doi:10.1186/s12711-017-0353-y

Schaeffer, L. R. (2006). Strategy for Applying Genome-wide Selection in Dairy Cattle. *J. Anim. Breed. Genet.* 123, 218–223. doi:10.1111/j.1439-0388.2006.00595.x

Stock, K., and Reents, R. (2013). Genomic Selection: Status in Different Species and Challenges for Breeding. *Reprod. Dom Anim.* 48 (Suppl. 1), 2–10. doi:10.1111/rda.12201

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., and Lund, M. S. (2012). Comparison of Genomic Predictions Using Medium-Density (~54,000) and High-Density (~777,000) Single Nucleotide Polymorphism Marker Panels in Nordic Holstein and Red Dairy Cattle Populations. *J. Dairy Sci.* 95, 4657–4665. doi:10.3168/jds.2012-5379

van Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., van Eeuwijk, F. A., Schrooten, C., and Veerkamp, R. F. (2015). Genomic Prediction Using Imputed Whole-Genome Sequence Data in Holstein Friesian Cattle. *Genet. Sel. Evol.* 47, 71. doi:10.1186/s12711-015-0149-x

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., et al. (2013). Genomic Imputation and Evaluation Using High-Density Holstein Genotypes. *J. Dairy Sci.* 96, 668–678. doi:10.3168/jds.2012-5702

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., and Weigel, K. A. (2011). Genomic Evaluations with many More Genotypes. *Genet. Sel. Evol.* 43, 10. doi:10.1186/1297-9686-43-10

Wang, C., Li, H., Guo, Y., Huang, J., Sun, Y., Min, J., et al. (2020a). Donkey Genomes Provide New Insights into Domestication and Selection for Coat Color. *Nat. Commun.* 11, 6014. doi:10.1038/s41467-020-19813-7

Wang, D., Tang, H., Liu, J.-F., Xu, S., Zhang, Q., and Ning, C. (2020b). Rapid Epistatic Mixed-Model Association Studies by Controlling Multiple Polygenic Effects. *Bioinformatics* 36, 4833–4837. doi:10.1093/bioinformatics/btaa610

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

Zhang, W., Li, W., Liu, G., Gu, L., Ye, K., and Zhang, Y. (2021). Evaluation for the Effect of Low-Coverage Sequencing on Genomic Selection in Large Yellow Croaker. *Aquaculture* 534, 736323. doi:10.1016/j.aquaculture.2020.736323

Conflict of Interest: XHZ and HL were employed by National Engineering Research Center for Gelatin-based TCM, Dong-E E-Jiao Co., Ltd.

The remaining author declares 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 in Genetics

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fgene.2021.752424

Genome-wide Association Study for Carcass Primal Cut Yields Using Single-step Bayesian Approach in Hanwoo Cattle

Masoumeh Naserkheil[1], *Hossein Mehrban*[2], *Deukmin Lee*[3]* *and Mi Na Park*[1]*

[1]Animal Breeding and Genetics Division, National Institute of Animal Science, Cheonan-si, South Korea, [2]Department of Animal Science, Shahrekord University, Shahrekord, Iran, [3]Department of Animal Life and Environment Sciences, Hankyong National University, Anseong-si, South Korea

The importance of meat and carcass quality is growing in beef cattle production to meet both producer and consumer demands. Primal cut yields, which reflect the body compositions of carcass, could determine the carcass grade and, consequently, command premium prices. Despite its importance, there have been few genome-wide association studies on these traits. This study aimed to identify genomic regions and putative candidate genes related to 10 primal cut traits, including tenderloin, sirloin, striploin, chuck, brisket, top round, bottom round, shank, flank, and rib in Hanwoo cattle using a single-step Bayesian regression (ssBR) approach. After genomic data quality control, 43,987 SNPs from 3,745 genotyped animals were available, of which 3,467 had phenotypic records for the analyzed traits. A total of 16 significant genomic regions (1-Mb window) were identified, of which five large-effect quantitative trait loci (QTLs) located on chromosomes 6 at 38–39 Mb, 11 at 21–22 Mb, 14 at 6–7 Mb and 26–27 Mb, and 19 at 26–27 Mb were associated with more than one trait, while the remaining 11 QTLs were trait-specific. These significant regions were harbored by 154 genes, among which *TOX*, *FAM184B*, *SPP1*, *IBSP*, *PKD2*, *SDCBP*, *PIGY*, *LCORL*, *NCAPG*, and *ABCG2* were noteworthy. Enrichment analysis revealed biological processes and functional terms involved in growth and lipid metabolism, such as growth (GO:0040007), muscle structure development (GO:0061061), skeletal system development (GO:0001501), animal organ development (GO:0048513), lipid metabolic process (GO:0006629), response to lipid (GO:0033993), metabolic pathways (bta01100), focal adhesion (bta04510), ECM–receptor interaction (bta04512), fat digestion and absorption (bta04975), and Rap1 signaling pathway (bta04015) being the most significant for the carcass primal cut traits. Thus, identification of quantitative trait loci regions and plausible candidate genes will aid in a better understanding of the genetic and biological mechanisms regulating carcass primal cut yields.

Keywords: candidate genes, QTL, carcass primal cut yield, single-step GWAS, hanwoo

OPEN ACCESS

Edited by:
Hao Cheng,
University of California, Davis,
United States

Reviewed by:
Gota Morota,
Virginia Tech, United States
Sungbong Jang,
University of Georgia, United States

***Correspondence:**
Deukmin Lee
dhlee@hknu.ac.kr
Mi Na Park
mina0412@korea.kr

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

Received: *03 August 2021*
Accepted: *02 November 2021*
Published: *26 November 2021*

Citation:
Naserkheil M, Mehrban H, Lee D and Park MN (2021) Genome-wide Association Study for Carcass Primal Cut Yields Using Single-step Bayesian Approach in Hanwoo Cattle. Front. Genet. 12:752424. doi: 10.3389/fgene.2021.752424

Abbreviations: HIC, hanwoo improvement center; GWAS, genome-wide association study; ssGBLUP, single-step genomic best linear unbiased prediction; ssBR, single-step Bayesian regression; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; BFT, backfat thickness; CW, carcass weight; EMA, eye muscle area; MS, marbling score; GO, gene ontology; KEGG, kyoto encyclopedia of genes and genomes; MCMC, markov chain monte carlo; GV, genomic variance; AGV, additive genetic variance; WPPA, window posterior probability of association.

INTRODUCTION

Hanwoo is an indigenous and popular meat-type cattle in Korea, and is particularly renowned for its rapid growth rate and quality attributes such as juiciness, tenderness, characteristic flavor, and extensive marbling of its beef (Jo et al., 2012). In recent years, both carcass and meat quality traits in Hanwoo have been extensively studied because of their economic relevance for optimizing the profitability of the beef industry. The current selection index in Hanwoo focuses on the improvement of carcass traits, such as backfat thickness (BFT), carcass weight (CW), eye muscle area (EMA), and marbling score (MS) as major selection criteria for breeding programs (Kim et al., 2017). However, other important traits such as carcass primal cuts have received inadequate attention in the Hanwoo breeding program, which affects both the quantity and quality of meat, and consequently, command premium prices. To meet consumer demand, the importance of primal cut yields is growing in the beef industry of developed countries because of its market value. Hence, cattle breeders need to address these traits, which determine selection decisions to increase carcass cut-out value and consumer acceptance of meat. Meanwhile, the existence of genetic variation and moderate to high heritability in the yield of primal cuts has been reported (Choi et al., 2015). In this sense, the improvement of primal cut yields requires knowledge of the underlying genetic background influencing these invaluable traits.

Over the last decade, with the development of high-throughput single nucleotide polymorphism (SNP) genotyping technologies, genome-wide association studies (GWAS) have become an affordable and powerful tool for detecting and localizing candidate genes and causal mutations associated with quantitative traits in different species (Matukumalli et al., 2009). Several statistical methods to conduct GWAS have been developed and applied, among which a simple regression model has been widely used, where one marker is tested at a time for significance (Meyer and Tier, 2012). However, this method was challenged by false positives and overestimation of quantitative trait loci (QTL) effects. Therefore, the marker effect models in the Bayesian approaches have been proposed for GWAS analysis (Habier et al., 2011; Wang et al., 2012; Moser et al., 2015; Wang et al., 2016) as they offer methods to overcome these challenges (Strömberg, 2009; Peters et al., 2012). One such method could have a higher power to detect SNPs with moderate effects on a trait of interest. In addition, Bayesian methods are flexible in accounting for the uncertainties of variables and parameters and allow for inferences to be made by finding their marginal posterior distributions (Yi and Shriner, 2008; Blasco and Blasco, 2017). Recently, Fernando et al. (2014; 2016) developed a class of single-step Bayesian regression methods (ssBR), which not only combines all available information as single-step genomic best linear unbiased prediction [ssGBLUP; (Misztal et al., 2009)] does, but also accommodates Bayesian models. This method can also be extended to GWAS and controls the proportion of false positives by computing the posterior probability of association of a trait with each SNP or each window of consecutive SNPs. Numerous association studies have been carried out on growth and carcass traits in beef cattle using different GWAS approaches (Peters et al., 2012; Lee et al., 2013; Magalhaes et al., 2016; Weng et al., 2016; Roberts, 2018; Bedhane et al., 2019; Naserkheil et al., 2020). However, GWAS have not yet been conducted to identify significant genomic regions for carcass primal cut traits, which are highly relevant to Hanwoo cattle breeding. Hence, the objective of this study was to perform GWAS to detect genomic regions and candidate genes associated with primal cut yields in Hanwoo cattle using the ssBR approach.

MATERIALS AND METHODS

Animal and Phenotype Data

A total of 3,467 Hanwoo steers born between 2008 and 2017 were included in this study. All steers were slaughtered at approximately 24 months of age and were progeny of 442 sires and 3,357 dams. The pedigree data consisted of 15,117 animals after tracing the pedigree file back 10 generations and pruning with SECATEURS software (Meyer, 2003). During the pruning process, 3,692 individuals were removed from the pedigree. All phenotypic data were collected by the Hanwoo Improvement Center (HIC) of the National Agricultural Cooperative Federation, South Korea. The ten primal cut yields considered in the present study (kg, composed of both unique and composite meat cuts from the forequarters and hindquarters) included tenderloin, sirloin, striploin, chuck, brisket, top round, bottom round, shank, flank, and rib; the locations of each cut on the carcass are illustrated in **Supplementary Figure S1**. The descriptive statistics and heritability estimates of the primal cut traits are shown in **Table 1**.

Genotype Data

In total, 12,764 animals were genotyped initially using three different SNP platforms, Illumina BovineSNP50K version 2 (n = 3,720), version 3 (n = 4,121), and customized Hanwoo version 1 (n = 4,923). Individuals (and SNPs) with a call rate of less than 90% and those without a valid phenotype were also excluded. The genotyped animals with Illumina BovineSNP50K version 2 were used as a reference populations to impute target animals (The genotyped animals using Illumina BovineSNP50K version 3 and customized Hanwoo version 1) using FImpute V3 software (Sargolzaei et al., 2014), and 52,791 SNPs on the 29 chromosomes were finally obtained. The analyses included genotypes for 2,957 steers with phenotypes and 788 their paternal ancestors. Quality control procedures were conducted using the BLUPF90 software (Misztal et al., 2015). SNPs with minor allele frequency less than 0.01 (8,783 SNPs), and a maximum difference between the observed and expected frequency of 0.15 as a departure of heterozygous from the Hardy-Weinberg equilibrium (21 SNPs) were discarded.

TABLE 1 | Descriptive statistics and heritability estimates for the primal cut traits in Hanwoo cattle.

Trait (unit)	No. of records	Mean (SE)	Min	Max	SD	CV (%)	h^2 (SD)
Tenderloin (Kg)	3,466	6.04 (0.01)	3	9	0.76	12.65	0.47 (0.03)
Sirloin (Kg)	3,465	34.23 (0.07)	16.8	50.7	4.11	12.02	0.49 (0.04)
Striploin (Kg)	3,465	7.85 (0.02)	4.3	12.4	1.17	14.96	0.46 (0.04)
Chuck (Kg)	3,463	14.61 (0.06)	6.7	34.8	3.76	25.72	0.32 (0.03)
Brisket (Kg)	3,466	23.76 (0.05)	12.6	38.6	3.01	12.67	0.59 (0.04)
Top round (Kg)	3,467	20.22 (0.04)	10.5	30.2	2.43	12	0.58 (0.04)
Bottom round (Kg)	3,467	32.99 (0.07)	16.6	49.6	3.92	11.89	0.58 (0.03)
Shank (Kg)	3,466	14.66 (0.03)	9	21.7	1.77	12.09	0.61 (0.04)
Flank (Kg)	3,465	28.29 (0.08)	12.5	50.3	4.83	17.08	0.35 (0.03)
Rib (Kg)	3,467	57.55 (0.13)	21.7	89.3	7.53	13.09	0.40 (0.04)

SE, standard error; SD, standard deviation; CV, coefficient of variation; h^2, heritability.

After quality control, 3,745 animals with the genotypes on 43,987 SNPs were remained for subsequent analyses.

Association Analyses

The single-step Bayesian regression (ssBR) method proposed by Fernando et al. (2014; 2016) was utilized to perform GWAS analyses, which combined all available phenotypes, genotypes, and pedigree information in a single-step. The estimation of genetic and residual variances as well as GWAS analyses were performed using univariate single-step Bayes B, with π being 0.99. The model for the single-step Bayesian GWAS (Fernando et al., 2014; Cheng et al., 2015; Fernando et al., 2016; Fernando et al., 2017) was as follows for genotyped animals:

$$y_i = \sum_{j=1}^{p_\beta} X_{ij}\beta_j + \sum_{k=1}^{p_u} Z_{ik}u_k + \sum_{l=1}^{p} M_{il}\alpha_l + e_i \quad (1)$$

and non-genotyped animals:

$$y_i = \sum_{j=1}^{p_\beta} X_{ij}\beta_j + \sum_{k=1}^{p_u} Z_{ik}u_k + \sum_{l=1}^{p} \hat{M}_{il}\alpha_l + \sum_{m=1}^{p_\varepsilon} Z_{n[i,m]}\varepsilon_m + e_i \quad (2)$$

where y_i is a phenotype for individual i; β_j is the *j*th effects including slaughter date (180 levels) and slaughter age (days from birth to slaughter) was considered as covariates; X_{ij} is the incidence covariate corresponding to the β_j for individual i; Z_{ik} is the incidence covariate corresponding to the *k*th random animal effect for individual i; $\mathbf{u} = [u1, u2, \ldots, u_{p_u}]$ is the vector of random animal effect assumed normally distributed N ($\mathbf{0}$, $\mathbf{A}$ σ_u^2), $\mathbf{A}$ is the numerator relationship matrix and σ_u^2 is additive genetic variance; M_{il} is the genotype covariate (coded as 0,1,2) at locus l for individual i; $\hat{M}_{il}$is the imputed genotype covariate at locus l for non-genotyped individual i; α_l is the allele substitution effect or marker effect for locus l assumed t-Student distributed t (0, σ_α^2) with probability 1- π and zero with π = 0.99, σ_α^2 is marker variance; $Z_{n[i,m]}$ is the incidence covariate corresponding to the *m*th imputation residual for individual i and ϵ_m is the imputation residual; *p* is the number of genotyped loci; p_β is the number of effect levels for $\boldsymbol{\beta}$; p_u is the number of random animal effect levels; p_ε is the number of non-genotyped animals; and e_i is the *i*th random residual effect for individual i assumed normally distributed N (0, σ_e^2), and σ_e^2 is residual variance.

The effects of $\boldsymbol{\beta}$ are assigned to the flat priors. In addition, the additive genetic (σ_u^2), residual (σ_e^2), and marker (σ_α^2) variances were assumed to have a scaled inverted chi-square prior with scale parameters S_α^2 and ν_α degrees of freedom. The prior means for additive genetic and residual variances were estimated using an animal model. In addition, the prior means was equal to $\sigma_u^2/[(1-\pi)\sum_{l=1}^{p} 2p_l(1-p_l)]$ for marker variance, as proposed by Habier et al. (2011), where p_i is the allelic frequency at the *l*th locus. The degrees of freedom were four for residual and marker variances, and five for additive genetic variance.

The analysis was performed using the JWAS Julia package for whole-genome analyses (Cheng et al., 2018) to obtain the posterior distributions of SNP effects using Markov chain Monte Carlo (MCMC). This method with 110,000 iterations was implemented to provide the posterior mean effects of the SNPs within each 1-Mb window and variance components after discarding the first 10,000 samples for burn-in and a thinning interval of 10. In total, 2,522 windows (1-Mb) across the 29 autosomes were included in the analyses. The window posterior probabilities of association (WPPA) for each window were also calculated.

The markers effect in each MCMC was estimated by using single-step Bayes B in addition to the polygenic additive genetic variance (σ_a^2) and residual variance (σ_e^2). The direct genomic value (DGV) that is attributed to markers is estimated as:

$$DGV_i = \sum_{l=1}^{p} M_{il}\alpha_l$$

The genomic variance (σ_m^2) was estimated using Gibbs sampling technique described by Sorensen et al. (2001). Then total genetic variance was estimated by summation of σ_m^2 and σ_u^2. In addition, the phenotypic variance was estimated by summation total genetic and environment variances. The heritability was obtained using total genetic variance divided by phenotypic variance.

The percentage of genomic variance (GV%) explained by each 1-Mb window in any particular iteration was calculated by dividing the genomic variance of the window by the genomic variance of the whole genome in the same iteration. Similarity, the proportion of additive genetic variance (AGV%) determined using each window markers were also obtained.

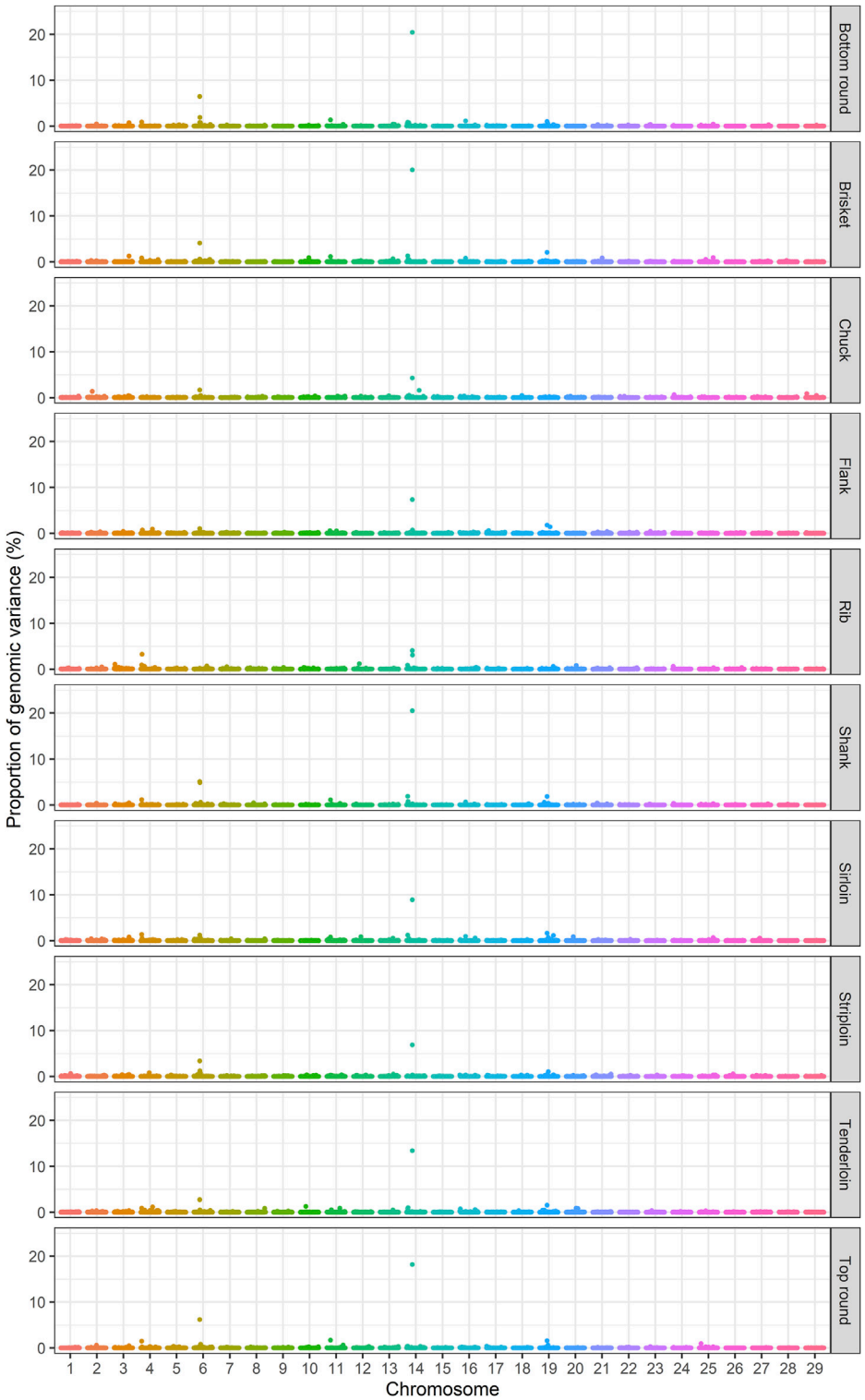

FIGURE 1 | Manhattan plots of the percentage of genomic variance explained by 1-Mb windows for primal cut traits in Hanwoo cattle.

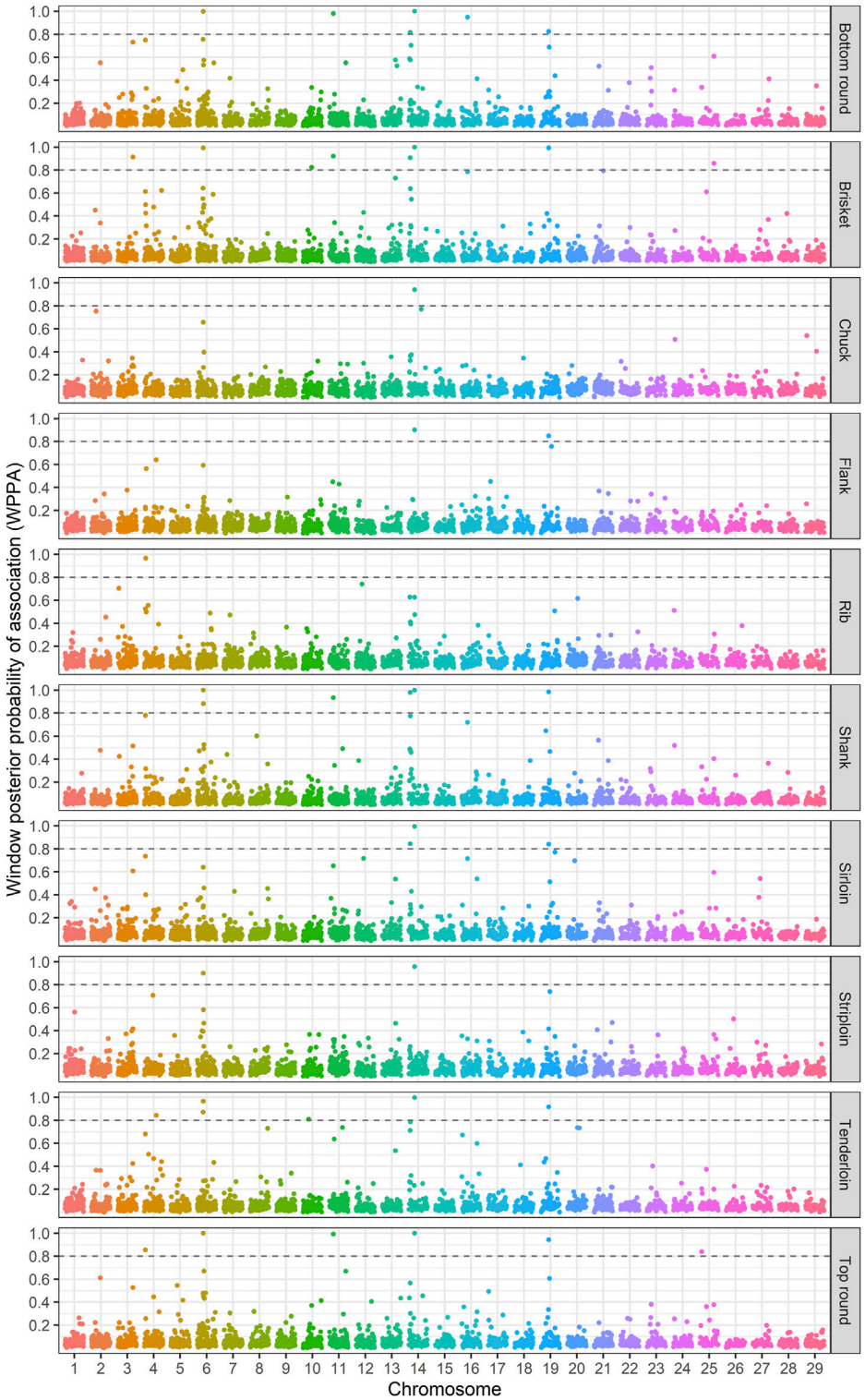

FIGURE 2 | Plots of window posterior probabilities of association (WPPA) obtained by the single-step Bayesian regression method for primal cut traits in Hanwoo cattle. The dash line is threshold 0.8 for significantly of windows.

TABLE 2 | Gene identification and proportion of variance explained by 1-Mb windows associated with the primal cut traits in Hanwoo cattle.

Trait	Chr	QTL region (Mb)	Number of SNPs	GV%	AGV %	WPPA	Candidate genes
Tenderloin	14	26–27	24	13.39	9.93	0.99	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Tenderloin	6	38–39	19	2.7	2.00	0.96	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL
Tenderloin	19	26–27	21	1.52	1.12	0.91	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Tenderloin	6	37–38	25	2.74	2.04	0.87	LOC104972722, TIGD2, FAM13A, LOC104972723, LOC104972724, LOC100847719, HERC3, NAP1L5, PYURF, PIGY, HERC5, HERC6, PPM1K, ABCG2, LOC781421
Tenderloin	4	77–78	18	1.13	0.84	0.84	RAMP3, WAP, TBRG4, NACAD, CCM2, LOC100140586, LOC101904529, MYO1G, PURB, MIR4657, H2AFV, PPIA, ZMIZ2, OGDH, TMED4, DDX56, LOC104972145, NPC1L1, NUDCD3, LOC104972146, CAMK2B, YKT6, GCK, MYL7, POLD2, AEBP1, POLM
Tenderloin	10	32–33	20	1.24	0.92	0.81	C10H15orf41, LOC107131394, LOC101904022, LOC104973117, MEIS2, LOC104973118
Sirloin	14	26–27	24	8.93	5.88	0.99	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Sirloin	14	6–7	33	1.23	0.81	0.84	
Sirloin	19	26–27	21	1.62	1.07	0.84	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Striploin	14	26–27	24	6.86	4.62	0.95	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Striploin	6	38–39	19	3.40	2.29	0.90	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL
Chuck	14	26–27	24	4.26	3.06	0.94	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Brisket	14	26–27	24	20.03	13.67	1.00	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Brisket	6	38–39	19	4.11	2.80	0.99	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL
Brisket	19	26–27	21	2.07	1.41	0.99	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Brisket	11	21–22	27	1.15	0.79	0.92	GALM, SRSF7, GEMIN6, LOC107132913, DHX57, MORN2, ARHGEF33, SOS1, MIR2284Z-2, LOC104973309, CDKL4, LOC782845, MAP4K3, LOC107132914, TMEM178A
Brisket	3	98–99	24	1.27	0.87	0.91	AGBL4, BEND5, LOC104971807, SPATA6, TRNAT-AGU, LOC107132336, SLC5A9, LOC107131387, SKINT1, TRNAR-ACG, LOC101906301, TRABD2B
Brisket	14	6–7	33	1.31	0.89	0.90	
Brisket	25	32–33	19	0.92	0.63	0.85	LOC104975903, LOC104975897, RN18S1, LOC104975899, LOC107131836, LOC107131838, LOC107131837, LOC107131839
Brisket	10	49–50	24	0.92	0.63	0.82	RORA, LOC107132858, LOC107132859, LOC104973153, ICE2, LOC107132852, LOC101902861, ANXA2
Top round	14	26–27	24	18.18	12.40	1.00	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Top round	6	38–39	19	6.19	4.22	0.99	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL
Top round	11	21–22	27	1.69	1.15	0.99	GALM, SRSF7, GEMIN6, LOC107132913, DHX57, MORN2, ARHGEF33, SOS1, MIR2284Z-2, LOC104973309, CDKL4, LOC782845, MAP4K3, LOC107132914, TMEM178A
Top round	19	26–27	21	1.57	1.07	0.94	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Top round	4	6–7	14	1.46	1.00	0.85	LOC101904266, LOC104970217, LOC781773, LOC107132367
Top round	25	4–5	25	0.93	0.64	0.83	SEC14L5, NAGPA, C25H16orf89, ALG1, LOC101905725, EEF2KMT, LOC521021, LOC107131809, LOC104975833
Bottom round	14	26–27	24	20.42	15.21	1.00	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Bottom round	6	38–39	19	6.46	4.83	0.99	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL

(Continued on following page)

TABLE 2 | (*Continued*) Gene identification and proportion of variance explained by 1-Mb windows associated with the primal cut traits in Hanwoo cattle.

Trait	Chr	QTL region (Mb)	Number of SNPs	GV%	AGV %	WPPA	Candidate genes
Bottom round	11	21–22	27	1.36	1.02	0.98	GALM, SRSF7, GEMIN6, LOC107132913, DHX57, MORN2, ARHGEF33, SOS1, MIR2284Z-2, LOC104973309, CDKL4, LOC782845, MAP4K3, LOC107132914, TMEM178A
Bottom round	16	25–26	8	1.13	0.84	0.94	HLX, TRNAQ-CUG, DUSP10
Bottom round	19	26–27	21	1.00	0.75	0.82	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Bottom round	14	6–7	33	0.87	0.65	0.81	
Shank	14	26–27	24	20.51	14.02	1.00	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Shank	6	38–39	19	5.09	3.48	0.99	PKD2, SPP1, MEPE, IBSP, LOC104972726, TRNAA-CGC, LAP3, MED28, FAM184B, NCAPG, DCAF16, LCORL
Shank	19	26–27	21	1.86	1.27	0.98	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Shank	14	6–7	33	1.90	1.30	0.98	
Shank	11	21–22	27	1.13	0.77	0.93	GALM, SRSF7, GEMIN6, LOC107132913, DHX57, MORN2, ARHGEF33, SOS1, MIR2284Z-2, LOC104973309, CDKL4, LOC782845, MAP4K3, LOC107132914, TMEM178A
Shank	6	39–40	24	4.89	3.35	0.88	LOC782905
Flank	14	26–27	24	7.37	5.34	0.90	FAM110B, LOC101902490, UBXN2B, CYP7A1, TRNAG-CCC, SDCBP, NSMAF, LOC101902713, LOC107133116, TOX, TRNAC-GCA
Flank	19	26–27	21	1.80	1.30	0.84	PITPNM3, FAM64A, AIPL1, WSCD1, LOC104975014, NLRP1, LOC788205, MIS12, DERL2, DHX33, C1QBP, RPAIN, NUP88, MIR199C, RABEP1, LOC101904050, SCIMP, LOC107131511
Rib	4	8–9	18	3.23	2.26	0.96	CDK14, LOC104971924, FZD1, LOC782091, LOC100140224

*GV%, proportion of the genomic variance explained by window; AGV%, proportion of the additive genetic variance explained by window; WPPA, window posterior probability of association. *Table was decreasingly sorted based on the WPPA, within each trait.*

Identification of Candidate Genes and Functional Enrichment Analysis

Genome windows with WPPA ≥0.8 (Wang et al., 2020) were considered as possible QTL regions associated with the studied traits. Candidate genes were searched for 1-Mb window using the Ensembl database and the Map Viewer tool of the bovine genome based on the starting and ending coordinates of significant windows. Further information on the function of these genes was obtained from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/gene/), and GeneCards (www.genecards.org). A Manhattan plot was created using the ggplot2 package (Wickham, 2009) in R software. To understand and identify the biological processes and pathways, gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment were carried out using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and web tool g:Profiler (Raudvere et al., 2019). Only GO terms with a significant p value of 0.05 and genes involved in biological processes, molecular functions, and cellular components were highlighted.

RESULTS

The number of records, means, minimum, maximum, standard deviations, phenotypic coefficient of variation, and heritability estimates for the 10 primal cut traits are provided in **Table 1**. The mean values of these traits ranged from 6.04 to 57.55 with standard deviation between 0.76 and 7.53. The coefficients of variation ranged from 11.89 to 25.72%, indicating considerable phenotypic variation of the investigated traits in the Hanwoo cattle population. Moderate to high heritability estimates were obtained for primal cut yields, which ranged from 0.32 ± 0.03 to 0.61 ± 0.04. Estimated variance components are also summarized in **Supplementary Table S1**.

The identification of genomic regions related to the traits of interest was performed using a ssBR approach. Only windows with WPPA ≥0.8, were considered significant. Manhattan plots displayed the proportion of genomic variance (**Figure 1**), additive genetic variance (**Supplementary Figure S2**) explained and the posterior probability of association by each 1-Mb window for the studied traits (**Figure 2**). Positional candidate genes for primal cut traits were also detected within the significant windows. A summary of significant windows associated with the traits under study, such as the number of SNPs in each window, explained GV % or AGV%, and WPPA, as well as candidate genes, is shown in **Table 2**. A total of 16 relevant genomic regions (1-Mb window) were found to be associated with the 10 traits recorded in this study. These regions were distributed over nine different chromosomes: 3, 4, 6, 10, 11, 14, 16, 19, and 25. Of these significant windows, five genomic windows were pleiotropic QTL, meaning that the QTL had an effect on multiple traits,

TABLE 3 | Gene Ontology (GO) terms and KEGG pathways significantly enriched using candidate genes associated with the primal cut traits.

Term ID	Term name	Count	Genes	*p*-value
Biological process				
GO: 0040007	Growth	5	CCM2, HLX, SOS1, DERL2, SDCBP	5.40E-06
GO: 0045927	positive regulation of growth	3	HLX, DERL2, SDCBP	1.50E-04
GO: 0007517	muscle organ development	2	HLX, FZD1	1.80E-02
GO: 0009888	tissue development	10	PKD2, SPP1, MEPE, IBSP, CCM2, HLX, SOS1, ANXA2, FZD1, SDCBP	6.20E-10
GO: 0061061	muscle structure development	4	HLX, FZD1, RORA, MED28	2.80E-05
GO: 0060284	regulation of cell development	4	CAMK2B, C1QBP, DUSP10, SDCBP	8.30E-05
GO: 0001501	skeletal system development	2	ANXA2, MEPE	2.70E-02
GO: 0031214	biomineral tissue development	4	ANXA2, MEPE, SPP1, IBSP	3.70E-05
GO: 0048513	animal organ development	16	CCM2, HLX, SOS1, SDCBP, POLM, HERC6, MEIS2, RORA, ANXA2, FZD1, IBSP, MEPE, OGDH, PKD2, SPP1, TMEM178A	1.20E-11
GO: 0006629	lipid metabolic process	5	NPC1L1, PYURF, PIGY, RORA, CYP7A1	1.60E-05
GO: 0033993	response to lipid	4	RORA, DUSP10, RAMP3, SPP1	3.10E-05
GO: 0042157	lipoprotein metabolic process	3	NPC1L1, PYURF, PIGY	5.40E-05
GO: 0006096	glycolytic process	2	GCK, OGDH	3.30E-03
GO: 0009247	glycolipid biosynthetic process	2	PYURF, PIGY	7.70E-03
GO: 0045598	regulation of fat cell differentiation	2	RORA, DUSP10	6.70E-03
GO: 0051179	localization	6	MIS12, PITPNM3, C1QBP, DERL2, NUP88, RABEP1	2.40E-03
GO: 0070887	cellular response to chemical stimulus	9	RORA, C1QBP, CYP7A1, DERL2, FZD1, IBSP, PKD2, RAMP3, SDCBP	4.30E-08
GO: 0007049	cell cycle	4	MIS12, NCAPG, PKD2, SDCBP	4.20E-04
Molecular Function				
GO: 0005262	calcium channel activity	2	ANXA2, PKD2	3.12E-02
GO: 0005515	protein binding	4	SPP1, MEPE, PKD2, MED28	4.76E-02
GO: 0016874	Ligase activity	3	HERC5, HERC3, HERC6	2.76E-03
GO: 0008289	lipid binding	2	ANXA2, RORA	3.52E-02
GO: 0003824	catalytic activity	5	HERC5, HERC3, PPM1K, ABCG2, HERC6	7.40E-03
GO: 0000166	nucleotide binding	4	CDKL4, DHX57, SRSF7, MAP4K3	5.10E-03
Cellular Component				
GO: 0005789	endoplasmic reticulum membrane	8	CAMK2B, PYURF, TMEM178A, DERL2, PKD2, CYP7A1, PIGY, TMED4	1.30E-03
GO: 0005622	Intracellular	8	HERC5, PYURF, HERC3, PPM1K, PIGY, NAP1L5, ABCG2, HERC6	3.00E-02
GO: 0005783	endoplasmic reticulum	8	CAMK2B, PYURF, TMEM178A, DERL2, ALG1, PKD2, CYP7A1, PIGY	8.30E-03
GO: 0015629	actin cytoskeleton	3	NCAPG, PKD2, MED28	1.10E-02
GO: 0005576	extracellular region	5	IBSP, SPP1, MEPE, LAP3, PKD2	4.52E-02
GO: 0043226	Organelle	8	HERC5, PYURF, HERC3, PPM1K, PIGY, NAP1L5, ABCG2, HERC6	4.83E-02

(Continued on following page)

TABLE 3 | (*Continued*) Gene Ontology (GO) terms and KEGG pathways significantly enriched using candidate genes associated with the primal cut traits.

Term ID	Term name	Count	Genes	*p*-value
GO: 0099568	cytoplasmic region	2	PKD2, MED28	3.50E-02
KEGG pathways				
bta01100	Metabolic pathways	8	POLD2, OGDH, GALM, LAP3, ALG1, CYP7A1, PIGY, GCK	5.03E-07
bta04512	ECM -receptor interaction	2	SPP1, IBSP	3.80E-03
bta04510	Focal adhesion	4	SPP1, IBSP, MYL7, SOS1	2.10E-05
bta04975	Fat digestion and absorption	1	NPC1L1	8.20E-05
bta04020	Calcium signaling pathway	1	CAMK2B	7.10E-03
bta00010	Glycolysis/Gluconeogenesis	2	GALM, GCK	8.30E-03
bta04015	Rap1 signaling pathway	1	LCP2	3.30E-03

which were located on chromosomes 6 at 38–39, 11 at 21–22, 14 at 6–7 Mb and 26–27, and 19 at 26–27 Mb. The genomic window located on chromosome 14 at 26–27 Mb explained a large proportion of the genomic variance across the nine analyzed traits, including tenderloin, sirloin, striploin, chuck, brisket, top round, bottom round, shank, and flank. The largest QTL window was observed for shank, which explained 20.51% of the genomic variance, was located in the region of 26–27 Mb on chromosome 14. The QTL window with the smallest proportion of genomic variance (0.87%) was identified for bottom round, located on chromosome 14 at 6–7 Mb. Candidate genes responsible for the genomic variance explained by the 1-Mb window were identified using the *Bos taurus* genome map. A total of 154 genes were identified within the significant regions to be associated with the traits of interest. Of these, 92 genes were codified proteins, 3 were miRNA (microRNA), 6 were tRNA (RNA transporter), and 53 were pseudogenes (**Table 2**). Functional enrichment analysis revealed 18 biological processes, six molecular functions, seven cellular components, and seven KEGG pathways. The following biological process terms were highlighted: growth (GO:0040007), positive regulation of growth (GO:0045927), tissue development (GO:0009888), muscle structure development (GO:0061061), skeletal system development (GO:0001501), animal organ development (GO:0048513), cellular response to chemical stimulus (GO:0070887), lipid metabolic process (GO:0006629), response to lipid (GO:0033993), glycolipid biosynthetic process (GO:0009247), and cell cycle (GO:0007049) being the most significant for the traits under study. In the case of the molecular function, calcium channel activity (GO:0005262), protein binding (GO:0005515), ligase activity (GO:0016874), lipid binding (GO:0008289), catalytic activity (GO:0003824), and nucleotide binding (GO:0000166) were significant terms. The significant enrichments for cellular component, including endoplasmic reticulum membrane (GO:0005789), intracellular (GO:0005622), endoplasmic reticulum (GO:0005783), actin cytoskeleton (GO:0015629), extracellular region (GO:0005576), organelle (GO:0043226), and cytoplasmic region (GO:0099568) were obtained. Moreover, KEGG pathway analysis revealed that the identified candidate genes involved in primal cut yields were enriched in metabolic pathways (bta01100), focal adhesion (bta04510), ECM-receptor interaction (bta04512), glycolysis/gluconeogenesis (bta00010), fat digestion and absorption (bta04975), Rap1 signaling pathway (bta04015), and calcium signaling pathway (bta04020). Functional gene set annotation and enrichment pathways are presented in **Table 3**.

DISCUSSION

The aim of this study was to identify genomic regions associated with primal cut yields using a ssBR approach in Hanwoo cattle. The marker effect model in a Bayesian framework would seem to be useful for GWAS because they account for uncertainty in parameters required to construct posterior distributions for QTL inference, thereby improving accuracy of genomic predictions and the power of QTL detection (Fernando and Garrick, 2013; Mehrban et al., 2017). In recent years, interest in exploring genomic regions that control economically important traits in beef cattle has increased due to advances in high-throughput genotyping techniques and the constant availability of molecular data, statistical methods, and ease of application of GWAS. Primal cut traits have recently been proposed as potential indicator of carcass weight and overall carcass merit (Berry et al., 2019) given that the genetic correlations between these traits and carcass weight are generally moderate to strong (Choi et al., 2015; Judge et al., 2019). Nevertheless, selection for the weight of primal cuts requires knowledge of the genetic basis for these traits, which may be useful in future genomic evaluations targeting the improvement of weight in the more valuable primal cuts, and consequently increasing the profitability of the meat production system. Our results showed that primal cut yields were moderate to highly heritable, being in accordance with those reported in Hanwoo cattle (Choi et al., 2015), Chianina cattle (Sarti et al., 2013), Simmental cattle (Zhu et al., 2019), and Irish

cattle (Berry et al., 2019; Judge et al., 2019). Among all identified window regions, the QTL on chromosome 14 at position 26–27 Mb had a larger impact than any of the other QTLs and was associated with a greater number of traits. This region, which is related to tenderloin, sirloin, striploin, chuck, brisket, top round, bottom round, shank, and flank, explained between 4.26 and 20.51% of the genomic variance across all these traits. A total of 11 genes were detected on chromosome 14 at 26–27 Mb regions. Among these, *FAM110B*, *UBXN2B*, *NSMAF*, *TOX*, *SDCBP*, and *CYP7A1* were notable. This region also seems to be most significantly associated with carcass and growth traits in beef cattle (Magalhaes et al., 2016; Roberts, 2018; Zhang et al., 2019; Grigoletto et al., 2020). Positional candidate genes of *FAM110B*, *UBXN2B*, *NSMAF*, *CYP7A1*, *SDCBP*, and *TOX* have been previously reported to be associated with carcass weight in Hanwoo cattle (Lee et al., 2013; Bhuiyan et al., 2018; Naserkheil et al., 2020). For instance, Lee et al. (2013) identified the six most significant SNPs associated with carcass weight in Hanwoo that were located in or nearby *TOX*, *FAM110B*, and *SDCBP*. Similarly, Srikanth et al. (2020) reported that the most significant SNPs on chromosome 14 were located in *UBXN2B*, *CYP7A1*, *SDCBP*, and *TOX*, which have been regarded as positional candidate genes for carcass weight in Hanwoo cattle. It was also reported that *CYP7A1* and *SDCBP* are positional candidate genes for carcass weight and eye muscle area in Hanwoo cattle (Bhuiyan et al., 2018; Srivastava et al., 2020), weaning weight in Brangus cattle (Weng et al., 2016), and feed efficiency traits in Nellore cattle (Brunes et al., 2021). The *TOX* gene located within this region is associated with reproductive traits in Nellore (de Camargo et al., 2015), residual feed intake and mid-test metabolic weight in SimAngus (Seabury et al., 2017), and development of puberty in Brahman cattle (Fortes et al., 2012). In addition, the *UBXN2B* gene was found to be associated with mid-test metabolic weight in SimAngus (Seabury et al., 2017) and carcass weight, carcass fat, and carcass conformation in Simmental cattle (Purfield et al., 2019), which is known as a protein-coding gene involved in endoplasmic reticulum biogenesis. *FAM110B* has been reported to be associated with fat thickness in composite beef cattle (Roberts, 2018). This gene functions in the cell cycle and cell growth and might play an important role in increasing carcass weight in beef cattle by increasing cell number and size. A pleiotropic QTL on chromosome 6 at 38–39 Mb was associated with tenderloin, striploin, brisket, top round, bottom round, and shank, which explained the largest (6.46%) and smallest (2.7%) proportion of genomic variance for bottom round and tenderloin, respectively. This region harbors relevant candidate genes, including *PKD2*, *SPP1*, *MEPE*, *IBSP*, *LAP3*, *NCAPG*, and *LCORL*. Most of the positional genes detected on chromosome 6 have previously been associated with many economically important traits in beef and dairy cattle. In a study on Brangus beef cattle, Weng et al. (2016) reported that most positional genes associated with direct birth weight, weaning weight, and yearling weight are located on chromosome 6. Similarly, Saatchi et al. (2014) identified a large-effect pleiotropic QTL located on chromosome 6 at 37–42 Mb was associated with direct birth weight, calving ease, carcass weight, rib eye muscle area, and weaning weight across several cattle breeds. In addition, it has previously been reported that a large number of significant SNPs associated with skeleton trait in Simmental cattle that were harbored by *LAP3*, *FAM184B*, *LCORL*, and *NCAPG* genes (Xia et al., 2017). A major QTL on chromosome 6, extending from 36 to 39 Mb related to carcass weight, was also identified in Japanese Black cattle (Nishimura et al., 2012). Interestingly, the *NCAPG* and *LCORL* genes, while being associated with the skeletal type traits (Hoshiba et al., 2013; Doyle et al., 2020) have also been regarded as positional candidate genes for direct calving ease, feed intake, gain, meat, and carcass traits (Lindholm-Perry et al., 2011; Bongiorni et al., 2012; Purfield et al., 2019), as well as growth and lipid deposition (Snelling et al., 2010; Weikard et al., 2010) across several breeds. Moreover, Liu et al. (2015) identified *LCORL* as a positional gene related to weight and carcass composition traits in chickens based on GWAS and differentially expressed gene studies. Other notable positional candidate genes in this region including *PKD2*, *LAP3*, *SPP1*, *MEPE*, *IBSP* and *MED28,* have been associated with carcass weight and growth traits (Lindholm-Perry et al., 2011; Weng et al., 2016; Roberts, 2018; Naserkheil et al., 2020), milk production (Olsen et al., 2005), and reproductive traits and puberty (Daetwyler et al., 2008; Cánovas et al., 2014) in beef and dairy cattle.

The other large-effect pleiotropic QTL identified in multiple traits (tenderloin, sirloin, brisket, top round, bottom round, shank, and flank) was located on chromosome 19 at position 26–27 Mb. This region explained between 1 and 2.07% of the genomic variance across the traits of interest, and harbors candidate genes, including *PITPNM3*, *WSCD1*, *MIS12*, and *RABEP1*. A recent study on linear type traits conducted by Doyle et al. (2020b) identified *PITPNM3* as a potential gene for chest depth in Angus cattle. The *WSCD1* gene encodes a protein with sulfotransferase activity involved in glucose metabolism, and is associated with udder depth in dairy cattle (Li et al., 2021), feed efficiency and feeding behaviors in pigs (Guo et al., 2015), and body size in beef cattle (An et al., 2020). The *MIS12* and *RABEP1* genes have also been recently linked to milk production traits in dairy cattle (Cai et al., 2020; Li et al., 2021). Two other pleiotropic genomic windows were located on chromosome 11 at 21–22 Mb, and on chromosome 14 at position 6–7 Mb, which the proportion of genomic variance explained by these windows ranged from 1.13 to 1.69 and from 0.87 to 1.9, respectively. The region on chromosome 14 that was found to be associated with sirloin, brisket, bottom round, and shank had no positional candidate genes in cattle, whereas its orthologous region on human chromosome 8 containing *KHDRBS3* gene that may play a role as a negative regulator of cell growth and inhibition of cell proliferation (Ma et al., 2017). For QTL region on chromosome 11, a total of 15 genes annotated were found to be related to brisket, top round, bottom round, and shank. Overall, 11 chromosomal regions identified in this study were trait-specific QTLs for six traits. Three trait-specific QTLs were identified for tenderloin, which are distributed on chromosomes 4 at 77–78 Mb, 6 at 37–38 Mb, and 10 at 32–33 Mb. The trait-specific QTL on chromosome 6 was responsible for 2.74% of the genomic variance in tenderloin and harbors the positional candidate

genes *PPM1K*, *ABCG2*, and *PIGY*. The *PPM1K* gene is involved in cellular survival, phosphorus metabolic process, amino acid dephosphorylation, and development by regulating mitochondrial permeability transition pore function, which have been shown to be associated with increased carcass weight, mid-point metabolic weight, reduction of residual feed intake, feed efficiency conversion ratio, and marbling score in crossbred beef cattle (McClure et al., 2010). The *ABCG2* gene is related to body weight (Weng et al., 2016) and milk yield and composition traits (Cohen-Zinder et al., 2005), which is involved in iron transport and metabolism. *PIGY* is a member of the PIG gene family, which encodes the glycosylphosphatidylinositol-N-acetylglucosaminyltransferase (GPI-GnT) complex and plays an important role in cell–cell interactions. A previous study reported that there were significant effects of copy number variation of the *PIGY* gene on growth traits in three Chinese sheep breeds (Feng et al., 2020). Brisket also had three trait-specific QTLs located on chromosomes 3 at 98–99, 10 at 49–50, and 25 at 32–33 Mb and explained 1.27, 0.92, and 0.92% of the genomic variance for this trait, respectively. The top round had two trait-specific QTLs. One was located on chromosome 4 at 6–7 Mb (GV% = 1.46), and the other on chromosome 25 at 4–5 Mb (GV% = 0.93). A QTL for the bottom round was identified on chromosome 16 at 25–26 Mb, which accounted for 1.13% of the genomic variance and harbors the positional candidate gene *DUSP10*. Previous studies have shown that the *DUSP10* gene is associated with growth traits (Ribeiro et al., 2021) and carcass weight (Chang et al., 2018) in beef cattle. The QTL located on chromosome 6 at position 39–40 Mb was associated with shank and had a greater proportion of genomic variance (4.89%) than other trait-specific QTLs. Only a QTL on chromosome 4 at 8–9 Mb, which was responsible for 3.23% of the genomic variance for the rib, was identified. The *CDK14* gene is located in this region and is associated with fatty acid profile (C18:1 *trans*-9) in the intramuscular fat of *Longissimus thoracis* muscle of Nellore cattle (Lemos et al., 2016). According to these results, most primal cut traits are probably controlled by several QTLs with large effects. Among these, genomic regions located on chromosomes 6 and 14 were considered as hot spots for several causal variants related to many economically important traits in beef cattle. A similar conclusion was given by Mehrban et al. (2021), who identified major QTLs on chromosomes 6 and 14 for EMA, yearling weight, and particularly for CW using the weighted single-step GWAS in Hanwoo cattle.

Gene ontology and pathway enrichment analyses were carried out to gain insight into the genes identified within QTL windows using g:Profiler and DAVID functional classification clustering tools (**Table 3**). Our analyses revealed the significant GO terms classified into the biological processes, cellular components, molecular functions, and seven KEGG pathways were enriched for the studied traits. It is interesting to note that the majority of the common genes identified for the primal cut traits are involved in growth-related processes: growth, positive regulation of growth, muscle structure development, regulation of cell development, muscle organ development, tissue development, skeletal system development, biomineral tissue development, and animal organ development. Among them, six genes, namely *HLX*, *SOS1*, *SDCBP*, *ANXA2*, *FZD1*, and *MEPE*, were highlighted as the main candidates for the traits under study in at least three biological pathways. The *HLX* gene, located on chromosome 16 at approximately 25–26 Mb, is a protein-coding gene that is involved in embryogenesis and hematopoiesis. It has also been shown to be associated with intramuscular fat in composite beef cattle (Roberts, 2018). The *SDCBP* gene is located on a conserved region on chromosome 14. It encodes a protein that binds to a variety of transmembrane proteins and plays a crucial role in carcass and meat quality traits. *ANXA2* gene is known to encode a member of a widely distributed, phospholipid-binding, calcium-regulated, peripheral membrane protein family known as annexins. This gene is involved in molecular functions related to calcium channel activity and lipid binding. It is not surprising then, that the role of calcium in meat tenderness and muscle contraction, and is a key regulator of muscle growth in beef cattle (Sadkowski et al., 2009). *ANXA2* knockout affects white adipose tissue hypotrophy due to reduced fatty acid uptake in mice (Salameh et al., 2016). Furthermore, this gene is associated with feed conversion efficiency in beef cattle (Al-Husseini et al., 2014). The GO terms related to the lipid metabolic process, cellular responses to the chemical stimulus, cell cycle, localization, and regulation of fat cell differentiation were also significantly represented in the pleiotropic QTL windows (**Table 3**). Pathway enrichment revealed that eight genes from six window regions (located on chromosomes 4, 6, 11, 14, and 25) were significantly associated with the metabolic pathway (bta01100). Among the genes harbored in this pathway, *CYP7A1* is involved in the transport, synthesis, and secretion of cholesterol, steroids, and other lipids (Zhao et al., 2013), which play a crucial role in digestion and absorption of dietary fat and contribute in maintaining the balance of cholesterol and lipid metabolism within the body (Monte et al., 2009). Interestingly, the extracellular matrix (ECM)-receptor interaction (bta04512) and focal adhesion (bta04510) pathways were enriched in the genes *SPP1* and *IBSP* from a pleiotropic QTL located on chromosome 6 at 38–39 Mb, which functions as a positive regulator in skeletal muscle cells and is involved in bone mineralization processes. The ECM-receptor interaction pathway plays a key role in tissue, organ morphogenesis, cell maintenance, and tissue structure and function (Mariman and Wang, 2010). It was previously reported that the ECM-receptor is upregulated in subcutaneous fat and intramuscular fat and appears to be involved in adipogenesis and meat tenderness (Taye et al., 2018). Focal adhesion also participates in important biological processes and serves as a mechanical link to ECM receptors and other molecules. These results will improve our understanding of enriched molecular processes, pathways, and genes associated with the primal cut traits and shed some light on how different pathways control these traits.

CONCLUSION

In the current study, 16 genomic regions (SNP windows) were found to be associated with 10 primal cut traits in Hanwoo cattle using a single-step Bayesian regression GWAS. Within these

windows, five QTLs had pleiotropic effects, with the most significant region located on chromosome 14 at position 26–27 Mb. Several candidate genes with potential functions in tissue development, regulation of growth, and lipid metabolism for the related traits were highlighted, among which *SPP1*, *IBSP*, *PKD2*, *SDCBP*, *PIGY*, *CYP7A1*, and *MEPE* were well-known. Moreover, our findings can contribute to a better understanding of the genetic basis and biological processes underlying the traits of interest; consequently, information on QTL regions can be used to search for causal mutations and marker-assisted or genomic selection in Hanwoo breeding schemes.

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.hknu.ac.kr/dhlee/2797/subview), (https://www.aiak.or.kr) and (https://www.mtrace.go.kr)].

ETHICS STATEMENT

The animal study was reviewed and approved by National Institute of Animal Science (NIAS), Rural Development Administration of South Korea.

AUTHOR CONTRIBUTIONS

HM and DL conceived and designed the study and contributed to the discussion of the results. MP provided genotype and phenotypic data for analysis and interpreted the results. MN and HM conceived the study and analyzed the data. MN drafted the manuscript. All authors have read and approved the final manuscript.

FUNDING

This study was supported by the Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ015987) from the Rural Development Administration (RDA), Republic of Korea. This study was supported by the 2021 RDA Research Associate Fellowship Program of the National Institute of Animal Science, Rural Development Administration, Republic of Korea.

ACKNOWLEDGMENTS

We are grateful to the staff of the Korean Hanwoo Improvement Center of the National Agricultural Cooperative Federation for supplying data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.752424/full#supplementary-material

Supplementary Figure S1 | Location of 10 carcass primal cut yields in Hanwoo cattle.

Supplementary Figure S2 | Manhattan plots of the percentage of additive genetic variance explained by 1-Mb windows for primal cut traits in Hanwoo cattle.

REFERENCES

Al-Husseini, W., Gondro, C., Quinn, K., Herd, R. M., Gibson, J. P., and Chen, Y. (2014). Expression of Candidate Genes for Residual Feed Intake in Angus Cattle. *Anim. Genet.* 45 (1), 12–19. doi:10.1111/age.12092

An, B., Xu, L., Xia, J., Wang, X., Miao, J., Chang, T., et al. (2020). Multiple Association Analysis of Loci and Candidate Genes that Regulate Body Size at Three Growth Stages in Simmental Beef Cattle. *BMC Genet.* 21 (1), 32–11. doi:10.1186/s12863-020-0837-6

Bedhane, M., van der Werf, J., Gondro, C., Duijvesteijn, N., Lim, D., Park, B., et al. (2019). Genome-wide Association Study of Meat Quality Traits in Hanwoo Beef Cattle Using Imputed Whole-Genome Sequence Data. *Front. Genet.* 10, 1235. doi:10.3389/fgene.2019.01235

Berry, D. P., Pabiou, T., Fanning, R., Evans, R. D., and Judge, M. M. (2019). Linear Classification Scores in Beef Cattle as Predictors of Genetic merit for Individual Carcass Primal Cut Yields1. *J. Anim. Sci.* 97 (6), 2329–2341. doi:10.1093/jas/skz138

Bhuiyan, M. S. A., Lim, D., Park, M., Lee, S., Kim, Y., Gondro, C., et al. (2018). Functional Partitioning of Genomic Variance and Genome-wide Association Study for Carcass Traits in Korean Hanwoo Cattle Using Imputed Sequence Level SNP Data. *Front. Genet.* 9, 217. doi:10.3389/fgene.2018.00217

Blasco, A., and Blasco, P. (2017). *Bayesian Data Analysis for Animal Scientists*. Springer.

Bongiorni, S., Mancini, G., Chillemi, G., Pariset, L., and Valentini, A. (2012). Identification of a Short Region on Chromosome 6 Affecting Direct Calving Ease in Piedmontese Cattle Breed. *PLoS One* 7 (12), e50137. doi:10.1371/journal.pone.0050137

Brunes, L. C., Baldi, F., Lopes, F. B., Lôbo, R. B., Espigolan, R., Costa, M. F. O., et al. (2021). Weighted Single-step Genome-wide Association Study and Pathway Analyses for Feed Efficiency Traits in Nellore Cattle. *J. Anim. Breed. Genet.* 138 (1), 23–44. doi:10.1111/jbg.12496

Cai, Z., Dusza, M., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2020). Distinguishing Pleiotropy from Linked QTL between Milk Production Traits and Mastitis Resistance in Nordic Holstein Cattle. *Genet. Sel Evol.* 52 (1), 19–15. doi:10.1186/s12711-020-00538-6

Cánovas, A., Reverter, A., DeAtley, K. L., Ashley, R. L., Colgrave, M. L., Fortes, M. R. S., et al. (2014). Multi-tissue Omics Analyses Reveal Molecular Regulatory Networks for Puberty in Composite Beef Cattle. *PloS one* 9 (7), e102551. doi:10.1371/journal.pone.0102551

Chang, T., Xia, J., Xu, L., Wang, X., Zhu, B., Zhang, L., et al. (2018). A Genome-wide Association Study Suggests Several Novel Candidate Genes for Carcass Traits in Chinese Simmental Beef Cattle. *Anim. Genet.* 49 (4), 312–316. doi:10.1111/age.12667

Cheng, H., Qu, L., Garrick, D. J., and Fernando, R. L. (2015). A Fast and Efficient Gibbs Sampler for BayesB in Whole-Genome Analyses. *Genet. Sel Evol.* 47 (1), 80–87. doi:10.1186/s12711-015-0157-x

Cheng, H., Fernando, R., and Garrick, D. (2018). "JWAS: Julia Implementation of Whole-Genome Analysis Software," in Proceedings Of the World congress on Genetics Applied to Livestock Production*)*, 859.

Choi, T. J., Alam, M., Cho, C. I., Lee, J. G., Park, B., Kim, S., et al. (2015). Genetic Parameters for Yearling Weight, Carcass Traits, and Primal-Cut Yields of Hanwoo Cattle1. *J. Anim. Sci.* 93 (4), 1511–1521. doi:10.2527/jas.2014-7953

Cohen-Zinder, M., Seroussi, E., Larkin, D. M., Loor, J. J., Everts-Van Der Wind, A., Lee, J.-H., et al. (2005). Identification of a Missense Mutation in the Bovine ABCG2 Gene with a Major Effect on the QTL on Chromosome 6 Affecting Milk Yield and Composition in Holstein Cattle. *Genome Res.* 15 (7), 936–944. doi:10.1101/gr.3806705

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M., and Robinson, J. A. B. (2008). A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map. *J. Dairy Sci.* 91 (8), 3225–3236. doi:10.3168/jds.2007-0333

de Camargo, G. M. F., Costa, R. B., de Albuquerque, L. G., Regitano, L. C. A., Baldi, F., and Tonhati, H. (2015). Polymorphisms in TOX and NCOA2 Genes and Their Associations with Reproductive Traits in Cattle. *Reprod. Fertil. Dev.* 27 (3), 523–528. doi:10.1071/rd13360

Doyle, J. L., Berry, D. P., Veerkamp, R. F., Carthy, T. R., Evans, R. D., Walsh, S. W., et al. (2020a). Genomic Regions Associated with Muscularity in Beef Cattle Differ in Five Contrasting Cattle Breeds. *Genet. Sel Evol.* 52 (1), 2–18. doi:10.1186/s12711-020-0523-1

Doyle, J. L., Berry, D. P., Veerkamp, R. F., Carthy, T. R., Walsh, S. W., Evans, R. D., et al. (2020b). Genomic Regions Associated with Skeletal Type Traits in Beef and Dairy Cattle Are Common to Regions Associated with Carcass Traits, Feed Intake and Calving Difficulty. *Front. Genet.* 11, 20. doi:10.3389/fgene.2020.00020

Feng, Z., Li, X., Cheng, J., Jiang, R., Huang, R., Wang, D., et al. (2020). Copy Number Variation of the PIGY Gene in Sheep and its Association Analysis with Growth Traits. *Animals* 10 (4), 688. doi:10.3390/ani10040688

Fernando, R. L., Cheng, H., Golden, B. L., and Garrick, D. J. (2016). Computational Strategies for Alternative Single-step Bayesian Regression Models with Large Numbers of Genotyped and Non-genotyped Animals. *Genet. Sel Evol.* 48 (1), 96–98. doi:10.1186/s12711-016-0273-2

Fernando, R. L., Dekkers, J. C., and Garrick, D. J. (2014). A Class of Bayesian Methods to Combine Large Numbers of Genotyped and Non-genotyped Animals for Whole-Genome Analyses. *Genet. Sel Evol.* 46 (1), 50–13. doi:10.1186/1297-9686-46-50

Fernando, R. L., and Garrick, D. (2013). "Bayesian Methods Applied to GWAS," in *Genome-wide Association Studies and Genomic Prediction* (Springer), 237–274. doi:10.1007/978-1-62703-447-0_10

Fernando, R., Toosi, A., Wolc, A., Garrick, D., and Dekkers, J. (2017). Application of Whole-Genome Prediction Methods for Genome-wide Association Studies: a Bayesian Approach. *Jabes* 22 (2), 172–193. doi:10.1007/s13253-017-0277-6

Fortes, M. R. S., Lehnert, S. A., Bolormaa, S., Reich, C., Fordyce, G., Corbet, N. J., et al. (2012). Finding Genes for Economically Important Traits: Brahman Cattle Puberty. *Anim. Prod. Sci.* 52 (3), 143–150. doi:10.1071/an11165

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

Guo, Y. M., Zhang, Z. Y., Ma, J. W., Ai, H. S., Ren, J., and Huang, L. S. (2015). A Genomewide Association Study of Feed Efficiency and Feeding Behaviors at Two Fattening Stages in a White Duroc × Erhualian F2 Population1. *J. Anim. Sci.* 93 (4), 1481–1489. doi:10.2527/jas.2014-8655

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian Alphabet for Genomic Selection. *BMC bioinformatics* 12 (1), 186–212. doi:10.1186/1471-2105-12-186

Hoshiba, H., Setoguchi, K., Watanabe, T., Kinoshita, A., Mizoshita, K., Sugimoto, Y., et al. (2013). Comparison of the Effects Explained by Variations in the bovinePLAG1andNCAPGgenes on Daily Body Weight Gain, Linear Skeletal Measurements and Carcass Traits in Japanese Black Steers from a Progeny Testing Program. *Anim. Sci. J.* 84 (7), 529–534. doi:10.1111/asj.12033

Jo, C., Cho, S. H., Chang, J., and Nam, K. C. (2012). Keys to Production and Processing of Hanwoo Beef: A Perspective of Tradition and Science. *Anim. Front.* 2 (4), 32–38. doi:10.2527/af.2012-0060

Judge, M. M., Pabiou, T., Murphy, J., Conroy, S. B., Hegarty, P. J., and Berry, D. P. (2019). Potential Exists to Change, through Breeding, the Yield of Individual Primal Carcass Cuts in Cattle without Increasing Overall Carcass Weight1. *J. Anim. Sci.* 97 (7), 2769–2779. doi:10.1093/jas/skz152

Kim, S., Alam, M., and Park, M. N. (2017). Breeding Initiatives for Hanwoo Cattle to Thrive as a Beef Industry–A Review Study. *J. Anim. Breed. Genomics* 1 (2), 103.

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

Lemos, M. V., Chiaia, H. L., Berton, M. P., Feitosa, F. L., Aboujaoud, C., Camargo, G. M., et al. (2016). Genome-wide Association between Single Nucleotide Polymorphisms with Beef Fatty Acid Profile in Nellore Cattle Using the Single Step Procedure. *BMC genomics* 17 (1), 213–216. doi:10.1186/s12864-016-2511-y

Li, B., VanRaden, P. M., Null, D. J., O'Connell, J. R., and Cole, J. B. (2021). Major Quantitative Trait Loci Influencing Milk Production and Conformation Traits in Guernsey Dairy Cattle Detected on *Bos taurus* Autosome 19. *J. Dairy Sci.* 104 (1), 550–560. doi:10.3168/jds.2020-18766

Lindholm-Perry, A. K., Sexten, A. K., Kuehn, L. A., Smith, T. P., King, D. A., Shackelford, S. D., et al. (2011). Association, Effects and Validation of Polymorphisms within the NCAPG - LCORL Locus Located on BTA6 with Feed Intake, Gain, Meat and Carcass Traits in Beef Cattle. *BMC Genet.* 12 (1), 103–113. doi:10.1186/1471-2156-12-103

Liu, R., Sun, Y., Zhao, G., Wang, H., Zheng, M., Li, P., et al. (2015). Identification of Loci and Genes for Growth Related Traits from a Genome-wide Association Study in a Slow- × Fast-Growing Broiler Chicken Cross. *Genes Genom* 37 (10), 829–836. doi:10.1007/s13258-015-0314-1

Ma, X., Ren, L., Zhang, N., Liu, C., Zhu, Y., and Xiao, J. (2017). Knock-down of KHDRBS3 Gene Inhibits Proliferation of Human Ovarian Cancer CAOV-3 Cells. *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi* 33 (8), 1062–1066.

Magalhães, A. F. B., de Camargo, G. M. F., Fernandes, G. A., Gordo, D. G. M., Tonussi, R. L., Costa, R. B., et al. (2016). Genome-wide Association Study of Meat Quality Traits in Nellore Cattle. *PLoS One* 11 (6), e0157845. doi:10.1371/journal.pone.0157845

Mariman, E. C. M., and Wang, P. (2010). Adipocyte Extracellular Matrix Composition, Dynamics and Role in Obesity. *Cell. Mol. Life Sci.* 67 (8), 1277–1292. doi:10.1007/s00018-010-0263-4

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., et al. (2009). Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PloS one* 4 (4), e5350. doi:10.1371/journal.pone.0005350

McClure, M. C., Morsci, N. S., Schnabel, R. D., Kim, J. W., Yao, P., Rolf, M. M., et al. (2010). A Genome Scan for Quantitative Trait Loci Influencing Carcass, post-natal Growth and Reproductive Traits in Commercial Angus Cattle. *Anim. Genet.* 41 (6), 597–607. doi:10.1111/j.1365-2052.2010.02063.x

Mehrban, H., Lee, D. H., Moradi, M. H., IlCho, C., Naserkheil, M., and Ibáñez-Escriche, N. (2017). Predictive Performance of Genomic Selection Methods for Carcass Traits in Hanwoo Beef Cattle: Impacts of the Genetic Architecture. *Genet. Sel Evol.* 49 (1), 1–13. doi:10.1186/s12711-016-0283-0

Mehrban, H., Naserkheil, M., Lee, D. H., Cho, C., Choi, T., Park, M., et al. (2021). Genomic Prediction Using Alternative Strategies of Weighted Single-step Genomic BLUP for Yearling Weight and Carcass Traits in Hanwoo Beef Cattle. *Genes* 12 (2), 266. doi:10.3390/genes12020266

Meyer, K. (2003). Secateurs software.[WWW Document]. Available at: http://agbu.une.edu.au/~kmeyer/PRUNE/secateurs.pdf.

Meyer, K., and Tier, B. (2012). "SNP Snappy": A Strategy for Fast Genome-wide Association Studies Fitting a Full Mixed Model. *Genetics* 190 (1), 275–277. doi:10.1534/genetics.111.134841

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing Procedures for Genetic Evaluation Including Phenotypic, Full Pedigree, and Genomic Information. *J. Dairy Sci.* 92 (9), 4648–4655. doi:10.3168/jds.2009-2064

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., and Vitezica, Z. (2015). *Manual for BLUPF90 Family of Programs*. Athens, Greece: University of Georgia, 199.

Monte, M. J., Marin, J. J., Antelo, A., and Vazquez-Tato, J. (2009). Bile Acids: Chemistry, Physiology, and Pathophysiology. *Wjg* 15 (7), 804. doi:10.3748/wjg.15.804

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genet.* 11 (4), e1004969. doi:10.1371/journal.pgen.1004969

Naserkheil, M., Bahrami, A., Lee, D., and Mehrban, H. (2020). Integrating Single-step GWAS and Bipartite Networks Reconstruction Provides Novel Insights into Yearling Weight and Carcass Traits in Hanwoo Beef Cattle. *Animals* 10 (10), 1836. doi:10.3390/ani10101836

Nishimura, S., Watanabe, T., Mizoshita, K., Tatsuda, K., Fujita, T., Watanabe, N., et al. (2012). Genome-wide Association Study Identified Three Major QTL for Carcass Weight Including the PLAG1-CHCHD7 QTN for Stature in Japanese Black Cattle. *BMC Genet.* 13 (1), 40–11. doi:10.1186/1471-2156-13-40

Olsen, H. G., Lien, S., Gautier, M., Nilsen, H., Roseth, A., Berg, P. R., et al. (2005). Mapping of a Milk Production Quantitative Trait Locus to a 420-kb Region on Bovine Chromosome 6. *Genetics* 169 (1), 275–283. doi:10.1534/genetics.104.031559

Peters, S. O., Kizilkaya, K., Garrick, D. J., Fernando, R. L., Reecy, J. M., Weaber, R. L., et al. (2012). Bayesian Genome-wide Association Analysis of Growth and Yearling Ultrasound Measures of Carcass Traits in Brangus Heifers1. *J. Anim. Sci.* 90 (10), 3398–3409. doi:10.2527/jas.2011-4507

Purfield, D. C., Evans, R. D., and Berry, D. P. (2019). Reaffirmation of Known Major Genes and the Identification of Novel Candidate Genes Associated with Carcass-Related Metrics Based on Whole Genome Sequence within a Large Multi-Breed Cattle Population. *BMC genomics* 20 (1), 720–817. doi:10.1186/s12864-019-6071-9

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* 47 (W1), W191–W198. doi:10.1093/nar/gkz369

Ribeiro, V. M. P., Gouveia, G. C., Moraes, M. M. d., Araújo, A. E. M. d., Raidan, F. S. S., Fonseca, P. A. d. S., et al. (2021). Genes Underlying Genetic Correlation between Growth, Reproductive and Parasite burden Traits in Beef Cattle. *Livestock Sci.* 244, 104332. doi:10.1016/j.livsci.2020.104332

Roberts, A. (2018). Genome-wide Association Study for Carcass Traits in a Composite Beef Cattle Breed. *Livestock Sci.* 213, 35–43.

Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Large-effect Pleiotropic or Closely Linked QTL Segregate within and across Ten US Cattle Breeds. *BMC genomics* 15 (1), 442–517. doi:10.1186/1471-2164-15-442

Sadkowski, T., Jank, M., Zwierzchowski, L., Oprządek, J., and Motyl, T. (2009). Comparison of Skeletal Muscle Transcriptional Profiles in Dairy and Beef Breeds Bulls. *J. Appl. Genet.* 50 (2), 109–123. doi:10.1007/bf03195662

Salameh, A., Daquinag, A. C., Staquicini, D. I., An, Z., Hajjar, K. A., Pasqualini, R., et al. (2016). Prohibitin/annexin 2 Interaction Regulates Fatty Acid Transport in Adipose Tissue. *JCI insight* 1 (10). doi:10.1172/jci.insight.86351

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A New Approach for Efficient Genotype Imputation Using Information from Relatives. *BMC genomics* 15 (1), 478–512. doi:10.1186/1471-2164-15-478

Sarti, F. M., Pieramati, C., Lubricchio, E., Giontella, A., Lasagna, E., and Panella, F. (2013). Genetic Parameters for the Weights and Yields of Carcass Cuts in Chianina Cattle1. *J. Anim. Sci.* 91 (9), 4099–4103. doi:10.2527/jas.2012-6105

Seabury, C. M., Oldeschulte, D. L., Saatchi, M., Beever, J. E., Decker, J. E., Halley, Y. A., et al. (2017). Genome-wide Association Study for Feed Efficiency and Growth Traits in U.S. Beef Cattle. *BMC genomics* 18 (1), 386–425. doi:10.1186/s12864-017-3754-y

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 Cattle12. *J. Anim. Sci.* 88 (3), 837–848. doi:10.2527/jas.2009-2257

Sorensen, D., Fernando, R., and Gianola, D. (2001). Inferring the Trajectory of Genetic Variance in the Course of Artificial Selection. *Genet. Res.* 77 (1), 83–94. doi:10.1017/s0016672300004845

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

Srivastava, S., Srikanth, K., Won, S., Son, J.-H., Park, J.-E., Park, W., et al. (2020). Haplotype-Based Genome-wide Association Study and Identification of Candidate Genes Associated with Carcass Traits in Hanwoo Cattle. *Genes* 11 (5), 551. doi:10.3390/genes11050551

Strömberg, U. (2009). Empirical Bayes and Semi-bayes Adjustments for a Vast Number of Estimations. *Eur. J. Epidemiol.* 24 (12), 737–741. doi:10.1007/s10654-009-9393-0

Taye, M., Yoon, J., Dessie, T., Cho, S., Oh, S. J., Lee, H.-K., et al. (2018). Deciphering Signature of Selection Affecting Beef Quality Traits in Angus Cattle. *Genes Genom* 40 (1), 63–75. doi:10.1007/s13258-017-0610-z

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide Association Mapping Including Phenotypes from Relatives without Genotypes. *Genet. Res.* 94 (2), 73–83. doi:10.1017/s0016672312000274

Wang, T., Chen, Y. P., Bowman, P. J., Goddard, M. E., and Hayes, B. J. (2016). A Hybrid Expectation Maximisation and MCMC Sampling Algorithm to Implement Bayesian Mixture Model Based Genomic Prediction and QTL Mapping. *BMC genomics* 17 (1), 744–821. doi:10.1186/s12864-016-3082-7

Wang, Z., Chapman, D., Morota, G., and Cheng, H. (2020). A Multiple-Trait Bayesian Variable Selection Regression Method for Integrating Phenotypic Causal Networks in Genome-wide Association Studies. *G3: Genes, Genomes, Genet.* 10 (12), 4439–4448. doi:10.1534/g3.120.401618

Weikard, R., Altmaier, E., Suhre, K., Weinberger, K. M., Hammon, H. M., Albrecht, E., et al. (2010). Metabolomic Profiles Indicate Distinct Physiological Pathways Affected by Two Loci with Major Divergent Effect on *Bos taurus* Growth and Lipid Deposition. *Physiol. genomics* 42A (2), 79–88. doi:10.1152/physiolgenomics.00120.2010

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. *Livestock Sci.* 183, 4–11. doi:10.1016/j.livsci.2015.11.011

Wickham, H. (2009). Elegant Graphics for Data Analysis. *Media* 35 (211), 10–1007.

Xia, J., Fan, H., Chang, T., Xu, L., Zhang, W., Song, Y., et al. (2017). Searching for New Loci and Candidate Genes for Economically Important Traits through Gene-Based Association Analysis of Simmental Cattle. *Sci. Rep.* 7 (1), 42048–42049. doi:10.1038/srep42048

Yi, N., and Shriner, D. (2008). Advances in Bayesian Multiple Quantitative Trait Loci Mapping in Experimental Crosses. *Heredity* 100 (3), 240–252. doi:10.1038/sj.hdy.6801074

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, G.-X., Liu, Y., Li, Z.-X., Lv, C.-Z., Traboulsee, A., Sadovnick, A. D., et al. (2013). Variants in the Promoter Region of CYP7A1 Are Associated with Neuromyelitis Optica but Not with Multiple Sclerosis in the Han Chinese Population. *Neurosci. Bull.* 29 (5), 525–530. doi:10.1007/s12264-013-1347-6

Zhu, B., Guo, P., Wang, Z., Zhang, W., Chen, Y., Zhang, L., et al. (2019). Accuracies of Genomic Prediction for Twenty Economically Important Traits in Chinese Simmental Beef Cattle. *Anim. Genet.* 50 (6), 634–643. doi:10.1111/age.12853

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 in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.704118

Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data

Tianyu Deng[1†], Pengfei Zhang[1†], Dorian Garrick[2], Huijiang Gao[1], Lixian Wang[1] and Fuping Zhao[1*]*

[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, [2]A. L. Rae Centre of Genetics and Breeding, Massey University, Hamilton, New Zealand

OPEN ACCESS

Edited by:
Lingzhao Fang, University of Edinburgh, United Kingdom

Reviewed by:
Oscar Gonzalez-Recio, Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain
Peipei Ma, Shanghai Jiao Tong University, China
Goutam Sahana, Aarhus University, Denmark

***Correspondence:**
Lixian Wang
iaswlx@263.net
Fuping Zhao
zhaofuping@caas.cn

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

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

Received: *01 May 2021*
Accepted: *03 December 2021*
Published: *03 January 2022*

Citation:
Deng T, Zhang P, Garrick D, Gao H, Wang L and Zhao F (2022) Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. Front. Genet. 12:704118. doi: 10.3389/fgene.2021.704118

Genotype imputation is the term used to describe the process of inferring unobserved genotypes in a sample of individuals. It is a key step prior to a genome-wide association study (GWAS) or genomic prediction. The imputation accuracy will directly influence the results from subsequent analyses. In this simulation-based study, we investigate the accuracy of genotype imputation in relation to some factors characterizing SNP chip or low-coverage whole-genome sequencing (LCWGS) data. The factors included the imputation reference population size, the proportion of target markers /SNP density, the genetic relationship (distance) between the target population and the reference population, and the imputation method. Simulations of genotypes were based on coalescence theory accounting for the demographic history of pigs. A population of simulated founders diverged to produce four separate but related populations of descendants. The genomic data of 20,000 individuals were simulated for a 10-Mb chromosome fragment. Our results showed that the proportion of target markers or SNP density was the most critical factor affecting imputation accuracy under all imputation situations. Compared with Minimac4, Beagle5.1 reproduced higher-accuracy imputed data in most cases, more notably when imputing from the LCWGS data. Compared with SNP chip data, LCWGS provided more accurate genotype imputation. Our findings provided a relatively comprehensive insight into the accuracy of genotype imputation in a realistic population of domestic animals.

Keywords: genotype imputation, SNP density, reference population size, imputation accuracy, SNP chip, sequencing

INTRODUCTION

The availability of next-generation sequencing technologies has made it possible to take account of whole-genome sequencing (WGS) data for genome-wide association studies (GWASs) or genomic prediction (GP) (Koboldt et al., 2013; Ni et al., 2017). However, whole genome resequencing is typically more expensive than SNP chip genotyping in most species, precluding deep sequencing of every individual in a population. Accordingly, over the past decade, the application of GWAS and GP has mainly been based on SNP chip data. The content of SNP arrays have typically been chosen from a database comprising relatively small numbers of sequenced individuals, which can result in ascertain bias (Lachance and Tishkoff, 2013). Although SNP chips tend to be cost-effective compared to sequencing, they cannot capitalize on all the genomic information if the SNPs on the chip array

have incomplete linkage disequilibrium with the causal mutations. Furthermore, they do not provide the understanding of the causal mutation that can be obtained by annotation of highly significant sequence variants. One option is to impute SNP array genotypes to sequence resolution based on a reference population of a small number of deeply sequenced relatives. Another option is imputation from a large number of sparsely sequenced individuals, obtained from low-coverage whole-genome sequencing (LCWGS). Compared to SNP chip data, LCWGS can expose the segregating sequence variants and mitigate the ascertainment bias from SNP array.

Regardless of whether SNP arrays or LCWGS are used to characterize genotypes, imputation is an essential step in a GWAS or as a precursor to genomic prediction (Li et al., 2009; Al Kalaldeh et al., 2019). Imputation can infer unobserved genotypes in a sample of individuals that have higher genotyping density from an SNP array, LCWGS, or WGS. Since WGS data should contain all genomic variants including causal mutations, it can increase the probability that causal variants can be directly identified. Accordingly, imputation can boost the power of GWAS analyses, improve the accuracy of GEBV in genomic prediction, be the basis for fine mapping, and facilitate meta-analysis that combines multiple studies based on different types of marker sets (Druet et al., 2014; Al-Tassan et al., 2015; Song et al., 2019).

Orho-Melander et al. (2008) imputed untyped HapMap SNPs to carry out fine-mapping and consequently found that GCKR rs780094 was associated with opposite effects on fasting plasma triglyceride concentrations. Many novel loci that increased the risk of type 2 diabetes were identified using high-density imputation (Mahajan et al., 2018). Association statistics obtained using imputed data from ultra low-coverage (0.24x) sequencing data attained similar *p*-values at known associated variants to those which had been obtained using an SNP chip (Pasaniuc et al., 2012). Huang et al. (2015) used imputation to construct a genome map for 1,495 elite hybrid rice varieties and their inbred parental lines and investigated 38 agronomic traits. They identified 130 associated loci which proved that the accumulation of numerous rare superior alleles with positive dominance was an important contributor to the heterotic phenomena.

The advent of low-cost next-generation sequencing has led to a rapid increase in the size of publicly available reference data sets. For example, the 1,000 Bull Genomes Project (http://www.1000bullgenomes.com/) has now sequenced thousands of animals and obtained about 155 million genetic variants representing many of the world's cattle breeds, providing a high-quality reference population (Georges, 2014; Hayes and Daetwyler, 2019). Many studies have used the variants in that reference population for imputation to new datasets to improve the accuracy of genomic prediction or to identify new candidate genes (Ibeagha-Awemu et al., 2016; Aliloo et al., 2018).

However, using low-quality imputed data may not lead to reliable GWAS or higher accuracy in genomic predictions (van Binsbergen et al., 2014). Multiple factors can affect the imputation accuracy, including size of the imputation reference panel, the imputation method, the minor allele frequency of the variant being imputed, the accuracy of phasing that constructs haplotypes in the reference and the study samples, and the sequencing coverage of the reference panel (Das et al., 2018). Although some of the effects of these factors have been analyzed separately, a comprehensive analysis that jointly considered these factors would help users design more powerful datasets for GWAS or genomic prediction.

METHODS AND MATERIALS

Simulation

In this study, we employed simulations based on coalescence theory using *msprime* software to simulate sequence resolution data that are compatible with our knowledge of the demographic history of pigs (Pérez-Enciso, 2014). Pig populations experienced genetic mutation, migration, and bottleneck effects (Giuffra et al., 2000; Kim et al., 2002; Frantz et al., 2015), and the detailed parameters used are shown in **Table 1**. Following 58,000 simulated generations, four separate but related populations were simulated, which we refer to as P_1, P_2, P_3, or P_4 according to their genetic distance (**Figure 1**). In these four populations, there were a total of 20,000 diploid samples with 10 Mb of simulated sequence data. The P_1 population included 11,000 individuals, while each of the other three populations had 3,000 samples. The first 1,000 individuals from P_1 represented the target population for imputation. We randomly selected biallelic variants with MAF $\geq$ 0.01 in the target population to generate LCWGS data, and then selected evenly spaced markers at various densities to represent SNP chip data.

We used the WGS data to calculate the average kinship coefficients in a pair-wise fashion between individuals in these populations, as in **Table 2**. The kinship coefficients between P_1 and P_2–P_4 decrease successively, reflecting the increases in the genetic distances separating them.

Factors Influencing Imputation Accuracy

We took four factors affecting imputation accuracy for LCWGS and SNP chip data into account. These were the proportion of SNP markers relative to target sequence variants (i.e., SNP chip density), imputation reference population size, genetic distance between target and imputation reference population, and the methods of imputation. **Table 3** lists the levels of each factor considered. A total of 336 scenarios representing all the factorial combinations of these levels were analyzed. In terms of SNP density, we set six levels where 1, 5, 10, 30, 50, or 90% of genomic biallelic variants were present on the SNP chip or LCWGS, the target marker number or density in reference populations are shown in **Table 4**. In the P_1 population, we selected 100, 1 k, 3 k, 5 k, or 10 k simulated individuals to represent the imputation reference population but did not include any of the target individuals. For each of the other three populations, we set three levels of 100, 1 k, and 3 k of imputation reference samples.

Imputation for every scenario was undertaken using Beagle5.1 (20Nov19.573) in comparison to Minimac4 v1.0.0,

TABLE 1 | Parameters used of the simulation with *msprime*.

Population history structural factors	Parameters				
Chromosome length			10,000 000 bp (10 Mb)		
Mutation rate			1×10^{-7}		
Recombination rate			1×10^{-7}		
Number of generations back to the population history event	T_{ori} = 58,000	T_0 = 9,000	T_1 = 3,000	T_2 = 200	T_3 = 20
Migration rate	$m_{01} = 2.1 \times 10^{-5}$	$m_{12} = 1.1 \times 10^{-3}$	$m_{14} = 3.7 \times 10^{-4}$	$m_{23} = 5.2 \times 10^{-5}$	$m_{34} = 1.6 \times 10^{-3}$
Effective population size	N_0 = 10,873	N_1 = 1,600	N_2 = 1,200	N_3 = 1,000	N_4 = 1,400

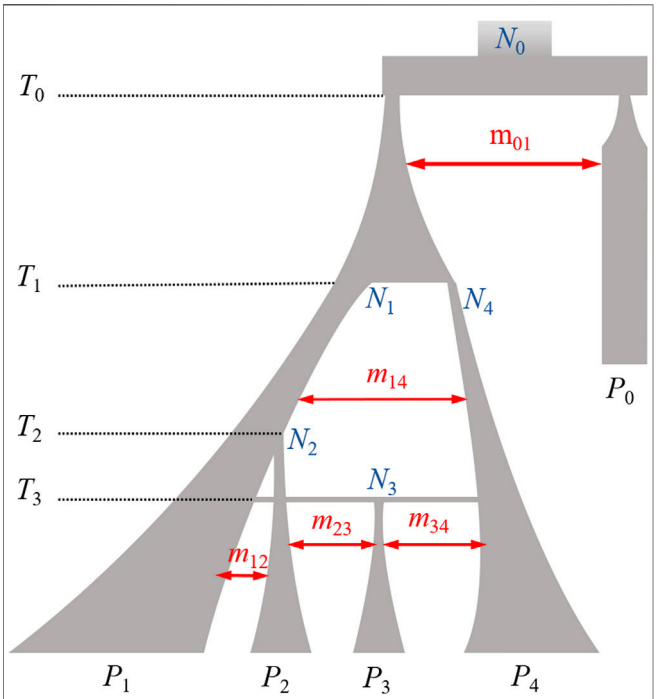

FIGURE 1 | Illustration of the simulation demographic model. Branch width corresponding to population size and time flowing from the top to the bottom. The width of the double arrows is proportional to the migration intensity. T_i, number of generations back to the population history event; N_j, effective population size of P_j population; m_{xy}, migration rate between P_x and P_y. Model details are provided in **Table 1** and **Supplementary Text S1**.

both with default parameter settings. Each program was run using its specific formats for reference panel data (bref3 for Beagle5.1 and m3vcf for Minimac4). We used Minimac3 to construct the m3vcf files. All imputation analyses were run on a dedicated 24-core 2.1-GHz workstation with an Intel Xeon Silver 4116 CPU and 128 GB of memory, and we evaluated one program at a time using five computational threads.

Assessment of Imputation Accuracy

Imputation reliability and the error rate were used as the two criteria to assess imputation accuracy. The imputation reliability is the squared Pearson correlation coefficient between the imputed genotypes and the true genotypes at a specific locus. The genotypes were coded as 0, 1, or 2, corresponding to the homozygous reference allele, heterozygous alternative allele, or homozygous alternative allele. The equation can be written as follows:

$$r_i^2 = \frac{(Cov(X_i, Y_i))^2}{Var(X_i)Var(Y_i)}$$

where r_i^2 is the imputation reliability for locus i; X_i is a vector of the imputed genotypes at locus i and Y_i is a vector of the true genotypes of imputed individuals at locus i.

The error rate refers to the percentage of loci that have wrongly imputed alleles:

$$er\,(\%) = \frac{n_{imputed \neq true}}{n_{imputed}} \times 100$$

where $er\,(\%)$ = the allelic imputation error rate, $n_{imputed \neq true}$ is the number of imputed alleles not equal to the true alleles, and $n_{imputed}$ is the number of alleles imputed.

We allocated the markers into several bins according to their MAFs and reported the average values of the imputation reliability and the error rate for all the markers within each bin. Furthermore, we calculated the regression of the imputation reliability or the error rate on the levels of each factor to determine if the factor had a significant effect ($p < 0.05$). We also report the correlation between the levels of each factor with the imputation reliability or the error rate. We used coefficients of

TABLE 2 | Genetic relationship between pair-wise populations.

Population	Average kinship coefficient		
	P_1	P_2	P_3
P_2	0.0070 (−0.065~0.522)[a]		
P_3	0.0027 (−0.077~0.394)	0.0030 (−0.070~0.510)	
P_4	0.0011 (−0.083~0.217)	0.0013 (−0.080~0.270)	0.0184 (−0.059~0.519)

[a]*Range of kinship coefficients, with minimum to maximum in parentheses.*

TABLE 3 | Levels of each factor to define the imputation scenarios.

Factors	Levels					
Reference population size	100	1,000	3,000	5,000	10,000	
Proportion of target markers/SNP density	1%	5%	10%	30%	50%	90%
Reference population	P_1	P_2	P_3	P_4		
Imputation method	Beagle5.1		Minimac4			
Data type	Chip data		Sequencing data			

TABLE 4 | Number of segregating genetic variants in four simulated populations.

Proportion of target markers[a]	Reference population				Marker density (SNPs/kb)
	P_1	P_2	P_3	P_4	
Total[b]	212,696	214,899	216,366	213,389	21.4
1%	2,126	2,148	2,163	2,133	0.2
5%	10,634	10,744	10,818	10,669	1.1
10%	21,269	21,489	21,636	21,338	2.1
30%	63,808	64,469	64,909	64,016	6.4
50%	106,348	107,449	108,183	106,694	10.7
90%	191,426	193,409	194,729	192,050	19.3

[a]*Represents the relative density of the pre-imputation marker panel.*
[b]*Total reflects the number of sequence variants targeted for imputation.*

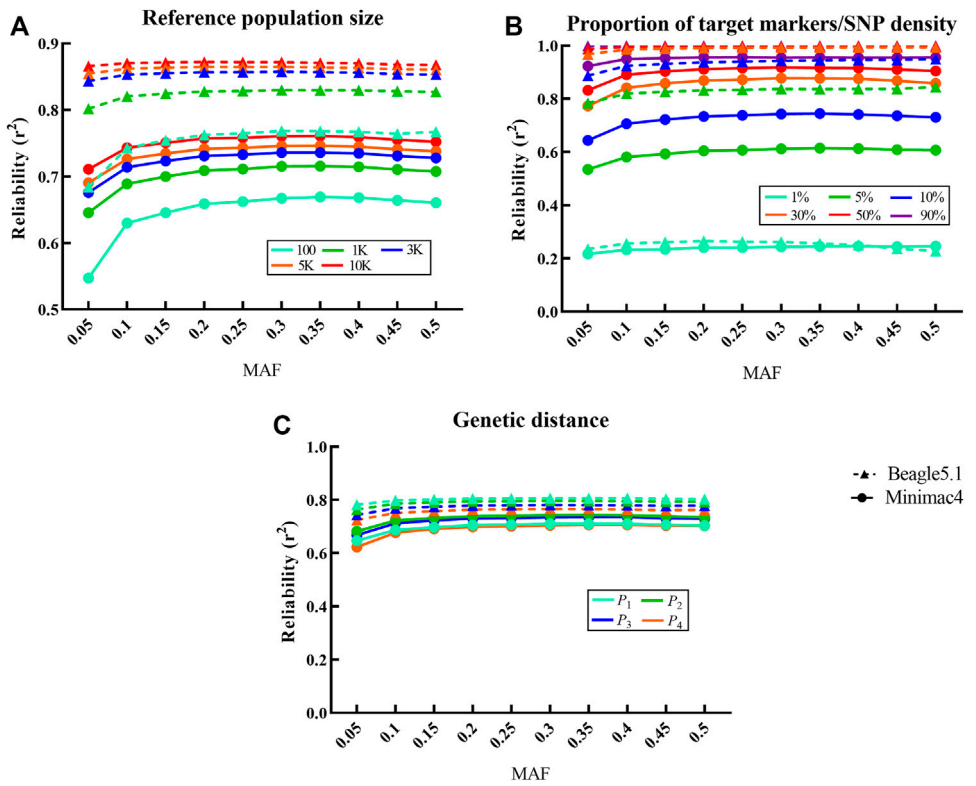

FIGURE 2 | Influence of different factors on imputation reliability in LCWGS data. For each fixed level of the factors under each scenarios, the average at different levels of all other factors is taken as the reliability. Imputed alleles are binned according to their MAF count in each scenarios. Dotted line with a triangle sign represents Beagle5.1, while the solid line with a round sign represents Minimac4. Different colored signs represent different levels. **(A)** Influence of reference population size on imputation reliability. **(B)** Influence of the proportion of target markers or SNP density on imputation reliability. **(C)** Influence of genetic distance between reference population and target population on imputation reliability.

TABLE 5 | Imputation reliability for different levels of imputation reference population size and SNP density.

Software	Reference population size	Proportion of target markers or SNP density/%					
		1	5	10	30	50	90
Beagle5.1	100	0.21	0.56	0.80	0.97	0.99	1.00
	1,000	0.25	0.78	0.94	0.99	1.00	1.00
	3,000	0.26	0.90	0.97	0.99	1.00	1.00
	5,000	0.26	0.94	0.98	1.00	1.00	1.00
	10,000	0.27	0.96	0.99	1.00	1.00	1.00
Minimac4	100	0.14	0.47	0.63	0.82	0.88	0.94
	1,000	0.20	0.58	0.72	0.86	0.90	0.95
	3,000	0.25	0.63	0.74	0.87	0.91	0.95
	5,000	0.28[a]	0.64	0.75	0.87	0.91	0.95
	10,000	0.33[a]	0.67	0.77	0.87	0.91	0.95

[a]*The imputation reliability of Minimac4 is higher than Beagle5.1 only for these two scenarios.*

variation (CVs) of the imputation reliability and the error rate to characterize imputation accuracy. The imputation computing time taken is reported for each scenario.

RESULTS

Factors Affecting Imputation Reliability

Significant differences in imputation reliabilities when imputing the sequence data were observed with regard to reference population size. Beagle5.1 typically outperformed Minimac4 with regression coefficients for reliability on reference population size being β = 0.783 and 0.756, respectively (**Figure 2A**). As seen in **Figure 2A**, as the reference population size increased from 100 to 10,000, the average imputation reliabilities of Beagle5.1 increased from 0.75 to 0.87, whereas the average reliabilities of Minimac4 increased from 0.65 to 0.75.

Changes in SNP density in the target population significantly affect the reliability of Beagle5.1 and Minimac4 ($p < 10^{-4}$, β = 0.785 and 0.925). When SNP density increased from 1 to 90%, the average imputation reliabilities increased from 0.25 to 0.99 in Beagle5.1 and from 0.24 to 0.95 in Minimac4 (**Figure 2B**).

The genetic distance between the target population and the reference population had a very significant impact on the reliability for Beagle5.1 ($p < 10^{-4}$, β = -0.852), but not for Minimac4 (p = 0.43). When the reference population changed from P_1 to P_4, the average imputation reliabilities with Beagle5.1 decreased from 0.80 to 0.69 (**Figure 2C**). A similar trend was shown in SNP chip data (**Supplementary Figure S1**; **Supplementary Table S1**). In addition, imputation reliability showed a trend of first increasing and then slightly decreasing with an increase in MAF, which was more obvious when the reference population was small or genetically distant.

CVs of imputation reliability varied at different levels for the above factors. For Beagle5.1, the CV of reference population size, proportion of target markers/SNP density, and genetic distance were 0.051, 0.320, and 0.021, respectively, while the CV of reference population size and proportion of target markers/SNP density in Minimac4 were 0.051 and 0.340. These indicate that proportion of target markers/SNP density is the most important factor affecting the imputation reliability in both methods.

The imputation reliabilities (**Table 5**) of Beagle5.1 ranged from 0.21 to 1.00 under different levels of SNP density and reference population size, while the imputation reliabilities of Minimac4 ranged from 0.14 to 0.95. In most cases, the reliabilities of Beagle5.1 were higher than those of Minimac4, except when SNP density was 1% and the reference population size was greater than 5,000. To obtain $r^2 \geq 0.8$ with at least 100 individuals in a reference population, Beagle5.1 required an SNP density of 10%, but Minimac4 required an SNP density of around 30%. Minimac4 could not achieve imputation accuracies of 100%. The performance of Beagle5.1 in reliability was better than that of Minimac4.

Factors Affecting Imputation Error Rate

The reference population size (**Figure 3**) had a very significant effect on the imputation error rate of Beagle5.1 with a negative correlation (β = -0.431, $p < 10^{-2}$), but not with Minimac4. As shown in **Figure 3A**, when the number of reference samples increased from 100 to 10,000, the average error rates of Beagle5.1 decreased from 6.42 to 3.31%, while the average imputation error rate of Minimac4 hardly changed. As shown in **Figure 3B**, SNP density has a very significant impact on the imputation error rate in both Beagle5.1 and Minimac4 ($p < 10^{-4}$, β = -0.687 and −0.530), and the error rate declined with the increase in SNP density. When the SNP density increased from 1 to 90%, the error rates in Beagle5.1 decreased from 18.43 to 0.07%; the error rates in Minimac4 decreased from 16.22 to 7.35%, corresponding to SNP density increasing from 1% to 50%. Although the genetic distance between the target population and the reference panel has no significant effects on the average imputation error rates of Beagle5.1 or Minimac4 (p = 0.36 and p = 0.74), it was observed that the lowest average error rates were 4.61 and 9.97% only when the reference population was P_1 (**Figure 3C**), and similar results are seen when imputing chip data (**Supplementary Figure S2**; **Supplementary Table S2**). In addition, the influence of MAF on the imputation error rate was significant and positively correlated in both methods ($p \leq 0.04$, $0.268 < \beta < 0.975$). But when the

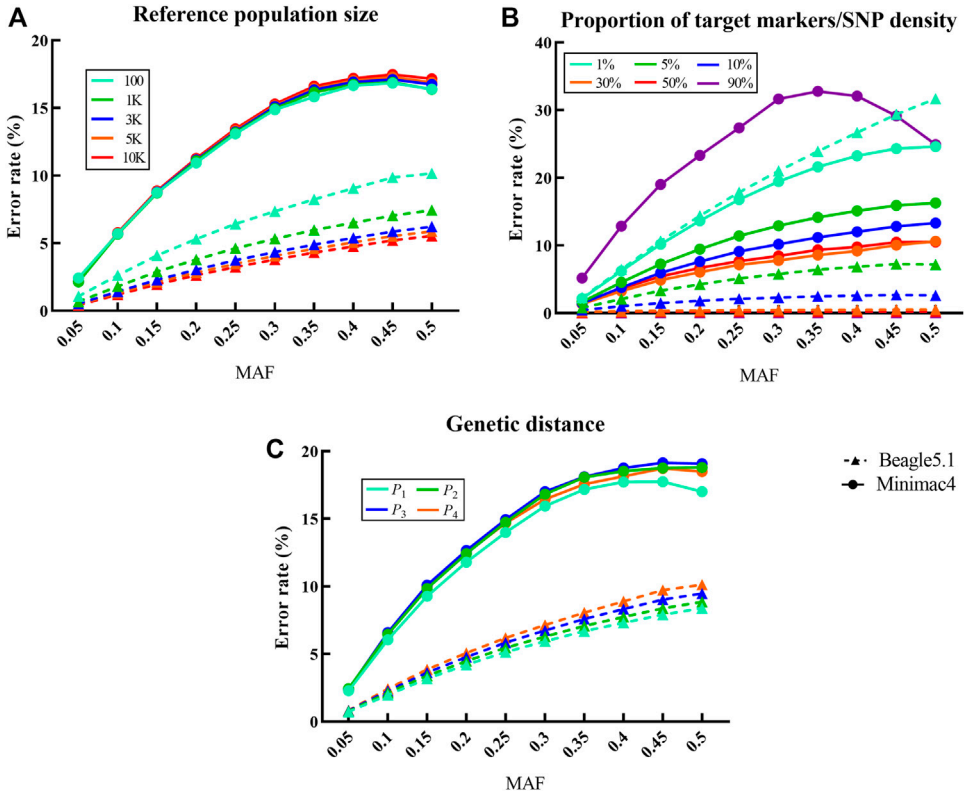

FIGURE 3 | Influence of different factors on the imputation error rate in LCWGS data. For each fixed level of the factors under each scenarios, the average at different levels of all other factors is taken as the error rate. Imputed alleles are binned according to their MAF count in each scenarios. Dotted line with a triangle sign represents Beagle5.1, while the solid line with a round sign represents Minimac4. Different colored signs represent different levels. **(A)** Influence of reference population size on the imputation error rate. **(B)** Influence of proportion of target markers or SNP density on the imputation error rate. **(C)** Influence of genetic distance between reference population and target population on the imputation error rate.

TABLE 6 | Imputation error rate (%) in the different levels of reference population size and SNP density.

Software	Reference population size	Proportion of target markers or SNP density/%					
		1	5	10	30	50	90
Beagle5.1	100	19.15	11.76	5.77	1.15	0.53	0.16
	1,000	18.44	6.55	1.91	0.43	0.23	0.09
	3,000	18.27	2.99	0.98	0.22	0.12	0.05
	5,000	18.20	2.03	0.68	0.15	0.08	0.04
	10,000	18.10	1.17	0.41	0.09	0.05	0.02
Minimac4	100	17.25[a]	11.41[a]	9.01	6.83	7.09	21.26
	1,000	16.36[a]	10.80	8.64	6.79	7.24	23.31
	3,000	16.00[a]	10.72	8.65	6.88	7.39	24.33
	5,000	15.85[a]	10.70	8.68	6.93	7.46	24.80
	10,000	15.67[a]	10.70	8.74	7.02	7.58	25.46

[a]*The imputation error rate of Minimac4 is lower than Beagle5.1 only for these six scenarios.*

conditions are conducive to imputation (such as a larger reference population, higher SNP density, or a closer genetic distance between populations), this effect will be less pronounced.

In Beagle5.1, the CVs of the imputation error rate for reference population size and SNP density were 0.262 and 1.508, respectively, while the CV of the imputation error rate affected by the SNP density in Minimac4 is 0.339. This indicated that SNP density was the most important factor affecting the error rate in both imputation methods. In addition, the uncontrollable factor MAF also has a considerable impact on the error rate.

As seen in **Table 6**, the imputation error rate ranges of Beagle5.1 and Minimac4 were 0.02–19.15% and 6.79–17.25%, respectively. Only when the SNP density was at the extreme low of 1% did Minimac4 exhibit its advantage. In order to achieve an

TABLE 7 | Runtime (min) to impute 10 Mb low-coverage whole-genome sequencing data with regard to software, reference population size, and proportion of target markers/SNP density.

Software	Reference population size	Proportion of target markers or SNP density/%					
		1	5	10	30	50	90
Beagle5.1	100	107.28	114.60	114.52	110.37	106.45	103.09
	1,000	108.88	120.42	122.82	121.04	105.25	104.76
	3,000	106.18	119.28	116.05	116.43	102.54	100.29
	5,000	106.48	122.42	122.37	118.86	112.88	103.36
	10,000	110.80	123.02	122.37	120.22	112.03	106.18
Minimac4	100	5.22	7.31	7.00	6.75	5.45	4.59
	1,000	8.37	10.06	9.76	9.25	9.36	9.41
	3,000	11.39	13.47	13.91	14.29	15.52	16.5
	5,000	15.20	17.40	17.11	19.15	21.25	24.95
	10,000	21.23	24.63	24.95	29.55	33.40	35.43

TABLE 8 | Coefficient of variation of imputation reliability and imputation error rates.

Software	Accuracy criterion	Data type	Coefficient of variation		
			Proportion of target markers / SNP density	Reference population size	Genetic distance
Beagle5.1	Reliability	SNP chip	0.164[a]	0.083	0.051
		LCWG sequencing	0.320[a]	0.051	0.021
	Error rate	SNP chip	0.393	0.541[a]	---
		LCWG sequencing	1.508[a]	0.262	---
Minimac4	Reliability	SNP chip	0.313[a]	0.056	---
		LCWG sequencing	0.340[a]	0.051	---
	Error rate	SNP chip	0.490[a]	---	---
		LCWG sequencing	0.339[a]	---	---

[a]*The most important factor affecting the imputation in each scenario.*

A dash (---) indicates that the factor has no significant effect on imputation accuracy in this scenario.

imputation error rate <10%, the imputation of Beagle required SNP density over 5% or to appropriately reduce the SNP density when increasing reference population size, while Minimac4 required the SNP density above 10% but was less dependent on the size of the reference panel. When the reference sample size was 100 and SNP density was slightly higher than 10%, the error rate was less than 5% for Beagle5.1 but not for Minimac4. The performance of Beagle5.1 was better than that of Minimac4 in most cases in terms of the error rate.

Imputation Runtime

The runtimes to impute to the sequence level taken by the two methods in the 1,000-target sample under all scenarios are summarized in **Table 7**. As seen in **Table 7**, both the reference population size and SNP density affected the imputation times. Minimac4 was always faster than Beagle5.1. Reference population size and SNP density hardly affected the imputation times taken for Beagle5.1 (**Supplementary Table S3**). The imputation time of Beagle5.1 only increased with an increase in the proportion of target markers. Beagle5.1 was only faster than Minimac4 when the percentage of target markers was 1% and the reference population sample was more than 1,000 individuals or when the proportion of target markers was 5% and the reference population sample was 10,000. However, considering the trend that the size of the reference population has little effect on time consumed in Beagle5.1, it is likely that Beagle5.1 will eventually be faster than Minimac4 as the reference population size continues to increase.

Comparison of Imputation Accuracies of LCWGS and Chip Array Data

We have calculated the CV of the two imputation accuracy standards in all scenarios. The CV is defined as the ratio of the standard deviation to the mean, and it can indicate the extent of the impact of factors considered on the imputation accuracy. Each row in **Table 8** represents a different imputation scenario with the asterisked ones being the most important factor affecting imputation in each situation. It can be seen that the SNP density (the proportion of target markers) was the most important in most scenarios. Compared to Minimac4, the imputation accuracies of Beagle5.1 were affected by more factors under the same condition.

Although the changes of various factors in this study have almost the same influence on imputation of either LCWGS or chip data, when the level of each factor is the same, there is a difference in imputation between chip data and LCWGS data. Therefore, we directly compared the imputation of the two

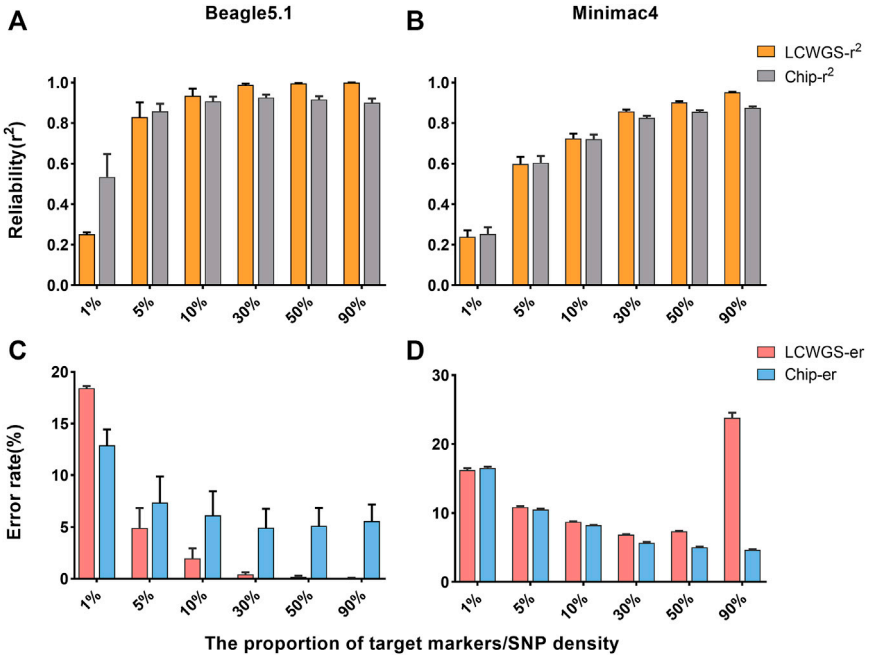

FIGURE 4 | Comparison of imputation accuracy using two types of data. **(A)** Comparison of reliability in Beagle5.1. **(B)** Comparison of reliability in Minimac4. **(C)** Comparison of the error rate in Beagle5.1. **(D)** Comparison of the error rate in Minimac4. LCWGS-r^2, imputation reliability use LCWG sequence data; Chip-r^2, imputation reliability use chip data; LCWGS-er, imputation error rate use LCWG sequence data; Chip-er, imputation error rate use chip data.

methods based on the two types of data. The imputation reliability of the two types of data in Beagle5.1 is shown in **Figure 4A**. When the proportion of target markers is 1%, the average imputation reliability using chip data is 0.51, which is higher than the 0.25 using sequencing data. When the proportion of target markers is greater than 5% (the reliability of the two data types is equal to 0.83), the imputation reliability of LCWGS data completely surpasses that of chip data, and when the proportion of target makers is 30%, the average reliability using LCWGS data can reach the extremely high level of 0.99. Using the Minimac4 method, the reliability with chip data is not less than that with LCWGS data except when the proportion of target markers is 1% and both are 0.24. At other levels, higher imputation reliability can be obtained with LCWGS data (**Figure 4B**).

Figure 4C shows the error rate of imputation with two types of data in Beagle5.1. When the target marker proportion is ≥5%, the error rate with LCWGS data was lower than that with chip data and can reach at best 4.9%. In contrast, when imputation was with chip data, the error rate in all cases was higher. In Minimac4 (**Figure 4D**), there was no significant difference in the error rate between imputation with the two types of data, but neither reached the best achieved by Beagle5.1. These showed that in most cases, compared to imputation with chip data, imputation with LCWGS data can achieve higher accuracy imputation, especially in terms of the imputation error rate.

DISCUSSION

In previous studies (van Binsbergen et al., 2014; Kreiner-Møller et al., 2015; Schurz et al., 2019), the imputation reliability and the imputation error rate were used to assess imputation accuracy. Imputation reliability appears to be a more useful measure with respect to genomic prediction because the nature of imputation reliability coincides with the definition of reliability used for breeding values, and it does not depend on minor allele frequency (MAF). The imputation error rate depends on MAF, which makes it difficult to select the imputed SNPs used for subsequent predictions (Calus et al., 2014).

Imputation accuracy is more problematic for rare variants. Rare variants mean that the locus is almost mono-allelic. The correlation is not defined when one or other of the vectors of true and imputed variants are mono-allelic. Many rare variants will be excluded in subsequent analyses (Pook et al., 2020). Therefore, both imputation reliability and error rate were used to evaluate the accuracy of imputation in this study to consider different applications of the imputed data.

With the development of sequencing technology and the reduction of sequencing costs, choosing SNP chip or LCWGS data has become blurred. In this study, the imputation accuracies of two types of genomic data were different, but under the same scenario, these two types of genetic data have similar influences on significance for each factor considered. That is, the imputation process was not affected by the data type to impute. In the case of the SNP density or proportion of target markers being ≥5%, the

imputation performance of Beagle5.1 for the LCWGS data was better than that for the SNP array data, especially in terms of the error rate. This was consistent with findings by Rubinacci et al. (2021), who reported that the reliability of imputation of human sequencing data was the highest in ultrahigh-density chip data, sequencing data, and chip data. Moreover, VanRaden et al. (2015) compared the imputation of low- or medium-density chip data with low-coverage sequencing data with similar costs and found that 1× and 2× deep sequencing data performed better than 10 and 60 K chip data in terms of the imputation error rate and reliability. All these results suggest that low-coverage whole-genome sequencing data has great potential for imputing to whole-genome sequencing resolution. It should be noted that in the case of the proportion of target marker/SNP density being ≤1%, the imputation accuracy of Minimac4 for LCWGS was better than that of Beagle5.1. This might be because SNP markers evenly distributed in the genome can capture more genetic information than LWGS data with a limited number of genetic variants.

Apart from the choice of imputation reference panel, the software used affects the imputation accuracy. In this study, we only compared two software products including Beagle5.1 and Minimac4. Both packages are based on a 'state-space reduction' of the hidden Markov models (HMMs) describing haplotype sharing, but the specific simplification methods are different. In Beagle5.1, genotype imputation is based on identity by descent (IBD) and uses the genotypes at the target markers to identify long IBD segments that a target haplotype shares with the reference haplotypes before imputation. It integrates the identified IBD fragments of different lengths into a subset that contains almost the same information as the complete reference haplotypes (Browning et al., 2018). While Minimac4's model first divides the whole genome into consecutive blocks and iterates only over the unique haplotypes in each genomic block (for imputation with a fixed chromosome length, the length of these blocks is fixed). It uses a reversible mapping function that can reconstruct exactly the state space used by Minimac4 (Das et al., 2016). This will also change the length and number of IBDs in the subset. This is the reason why Beagle5.1 is more sensitive to reference population size. The flexible and computationally intensive method makes Beagle5.1 more suitable for imputing sequencing data in a large reference population size. Under most scenarios, the imputation accuracies of Beagle5.1 were better than those of Minimac4. When the reference population was small, Minimac4 had better performance in the error rate than Beagle5.1. This was consistent with the results of Korkuć et al. (2019). It should be mentioned that when the proportion of target markers was 90%, the imputation error rate of Minimac4 increased abnormally. This was due to the over-correction that caused the error rate of some alleles to be greater than 100% during imputation. To further explain this phenomenon, we rerun our script using Minimac4 when proportions of target markers were 70 and 80%. We still found that the results were similar to that of the density of 90%, and the numbers of alleles with over-correction increased with the increase in density (**Supplementary Table S4**). This may be a bug of Minimac4.

In the present study, increasing the reference population size led to more accurate imputation, which agreed with other studies (Delaneau et al., 2013; García-Ruiz et al., 2015). A larger reference population can provide more reference haplotypes and the target markers can be more easily matched to the haplotypes, making the reliability higher. Our results are similar to the findings of Hozé et al. (2013), that is, changes of reference population size in Beagle5.1 has a significant impact on the error rate. However, Zhang and Druet (2010) reported that compared with the number of SNPs and genetic distance between populations, the size of the reference population had a relatively small effect on the imputation error rate, which is similar to our findings for Minimac4. This also reflects the differences in calculations between the methods.

In order to obtain high reliability and low error rate imputation, in addition to choosing target markers that more easily match reference haplotypes, we can increase the proportion of target markers or SNP density or select individuals closely related to the target population as the reference population. Another factor that affected the imputation error rate was the difference in MAF, which at first sight may be an unexpected indicator for imputation, especially since haplotypes are used for imputation. However, as shown in other studies (Huang et al., 2009; Oliveira Júnior et al., 2017), since the process of imputation first calculated correlation between reference and target haplotypes and then considered the consistency between the haplotypes, when imputing markers with a higher allele frequency can maintain high correlations, if the frequency between the two genotypes were similar, the marker may not be imputed correctly.

In general, SNP density/the proportion of target markers should be considered first. In this study, when the proportion of target markers was less than 1%, the imputation results in all cases were very poor except that the reliability of imputing chip data with Beagle5.1 could be more than 0.5. An alternative method was a two-step method that has been proven to improve imputation reliability which first imputed the target marker with low-density to a medium-density chip or high-density chip data and then further imputed to sequence resolution (Kreiner-Møller et al., 2015; Wang et al., 2015). A large number of high-coverage sequencing individuals as the reference population data will significantly increase the cost. When the total sequencing depth is fixed (e.g., constrained by budget), balancing the number and depth of sequencing individuals can effectively improve the imputation accuracy, such as using 1,000 individuals with depths of 8× as a reference population have higher imputation reliability than a reference population composed of 500 individuals with 16× (VanRaden et al., 2015). On the other hand, the development and progress of the network database and cloud server technologies also provide opportunities for solving this issue (Das et al., 2018). For instance, the 1,000 Genomes Project and Haplotype Reference Consortium (HRC) public dataset in human research greatly facilitates the application of genotype imputation (Rubinacci et al., 2021). However, in the animal domain, except for the 1,000 Bull Genomes Project (Hayes and Daetwyler, 2019), data sharing channels are still very limited. The use of multiple populations to form a mixed reference population can effectively reduce genetic distance and improve imputation accuracy (Schurz et al., 2019).

CONCLUSION

In this study, we have comprehensively analyzed the influence of several factors on the accuracy of genotype imputation. The proportion of target marker/SNP density has a very significant impact on the imputation reliability and the error rate under all imputation situations, which indicate that it is the most important factor in genotype imputation. The imputation performance of Beagle5.1 was better than Minimac4 in most cases, but when the reference population was small, SNP density was low, or genetic distance was large; the imputation accuracy of Beagle5.1 was more easily affected than that of Minimac4. Compared with Minimac4, Beagle5.1 can achieve better imputation performance with relatively relaxed conditions, which was more obvious when the LCWG sequencing data was used to impute to sequence data. Except in the case of extremely low SNP density, the imputation accuracy based on sequencing data is usually better than that based on chip data. Our results provided a reference for the application of genotype imputation in domestic animals.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

FZ and LW conceived this research and designed the experiments. TD and PZ conducted the research and drafted the manuscript. DG and HG participated in its design and participated in drafting the manuscript. All authors contributed to the article and approved the final manuscript.

FUNDING

This research was funded by the Natural Science Foundations of China (No. 31572357) to FZ, the China Agriculture Research System of MOF and MARA (CARS-35) and the Agricultural Science and Technology Innovation Program (ASTIP-IAS02) to LW, and the Science and Technology Project of the Inner Mongolia Autonomous Region (No. 2020GG0210) to HG.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.704118/full#supplementary-material

REFERENCES

Al Kalaldeh, M., Gibson, J., Duijvesteijn, N., Daetwyler, H. D., MacLeod, I., Moghaddar, N., et al. (2019). Using Imputed Whole-Genome Sequence Data to Improve the Accuracy of Genomic Prediction for Parasite Resistance in Australian Sheep. *Genet. Sel Evol.* 51 (1), 32. doi:10.1186/s12711-019-0476-4

Al-Tassan, N. A., Whiffin, N., Hosking, F. J., Palles, C., Farrington, S. M., Dobbins, S. E., et al. (2015). A New GWAS and Meta-Analysis with 1000Genomes Imputation Identifies Novel Risk Variants for Colorectal Cancer. *Sci. Rep.* 5 (1), 10442. doi:10.1038/srep10442

Aliloo, H., Mrode, R., Okeyo, A. M., Ni, G., Goddard, M. E., and Gibson, J. P. (2018). The Feasibility of Using Low-Density Marker Panels for Genotype Imputation and Genomic Prediction of Crossbred Dairy Cattle of East Africa. *J. Dairy Sci.* 101 (10), 9108–9127. doi:10.3168/jds.2018-14621

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103 (3), 338–348. doi:10.1016/j.ajhg.2018.07.015

Calus, M. P. L., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F., and Mulder, H. A. (2014). Evaluation of Measures of Correctness of Genotype Imputation in the Context of Genomic Prediction: a Review of Livestock Applications. *Animal* 8 (11), 1743–1753. doi:10.1017/s1751731114001803

Das, S., Abecasis, G. R., and Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annu. Rev. Genom. Hum. Genet.* 19, 73–96. doi:10.1146/annurev-genom-083117-021602

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation Genotype Imputation Service and Methods. *Nat. Genet.* 48 (10), 1284–1287. doi:10.1038/ng.3656

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies. *Nat. Methods* 10 (1), 5–6. doi:10.1038/nmeth.2307

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward Genomic Prediction from Whole-Genome Sequence Data: Impact of Sequencing Design on Genotype Imputation and Accuracy of Predictions. *Heredity* 112 (1), 39–47. doi:10.1038/hdy.2013.13

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 (10), 1141–1148. doi:10.1038/ng.3394

García-Ruiz, A., Ruiz-Lopez, F. J., Wiggans, G. R., Van Tassell, C. P., and Montaldo, H. H. (2015). Effect of Reference Population Size and Available Ancestor Genotypes on Imputation of Mexican Holstein Genotypes. *J. Dairy Sci.* 98 (5), 3478–3484. doi:10.3168/jds.2014-9132

Georges, M. (2014). Towards Sequence-Based Genomic Selection of Cattle. *Nat. Genet.* 46 (8), 807–809. doi:10.1038/ng.3048

Giuffra, E., Kijas, J. M. H., Amarger, V., Carlborg, Ö., Jeon, J.-T., and Andersson, L. (2000). The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. *Genetics* 154 (4), 1785–1791. doi:10.1093/genetics/154.4.1785

Hayes, B. J., and Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* 7 (1), 89–102. doi:10.1146/annurev-animal-020518-115024

Hozé, C., Fouilloux, M.-N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., et al. (2013). High-density Marker Imputation Accuracy in Sixteen French Cattle Breeds. *Genet. Sel Evol.* 45 (1), 33. doi:10.1186/1297-9686-45-33

Huang, L., Wang, C., and Rosenberg, N. A. (2009). The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations. *Am. J. Hum. Genet.* 85 (5), 692–698. doi:10.1016/j.ajhg.2009.09.017

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., et al. (2015). Genomic Analysis of Hybrid rice Varieties Reveals Numerous superior Alleles that Contribute to Heterosis. *Nat. Commun.* 6 (1), 6258. doi:10.1038/ncomms7258

Ibeagha-Awemu, E. M., Peters, S. O., Akwanji, K. A., Imumorin, I. G., and Zhao, X. (2016). High Density Genome Wide Genotyping-By-Sequencing and Association Identifies Common and Low Frequency SNPs, and Novel Candidate Genes Influencing Cow Milk Traits. *Sci. Rep.* 6 (1), 31109. doi:10.1038/srep31109

Kim, K.-I., Lee, J.-H., Li, K., Zhang, Y.-P., Lee, S.-S., Gongora, J., et al. (2002). Phylogenetic Relationships of Asian and European Pig Breeds Determined by Mitochondrial DNA D-Loop Sequence Polymorphism. *Anim. Genet.* 33 (1), 19–25. doi:10.1046/j.1365-2052.2002.00784.x

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The Next-Generation Sequencing Revolution and its Impact on Genomics. *Cell* 155 (1), 27–38. doi:10.1016/j.cell.2013.09.006

Korkuć, P., Arends, D., and Brockmann, G. A. (2019). Finding the Optimal Imputation Strategy for Small Cattle Populations. *Front. Genet.* 10, 52. doi:10.3389/fgene.2019.00052

Kreiner-Møller, E., Medina-Gomez, C., Uitterlinden, A. G., Rivadeneira, F., and Estrada, K. (2015). Improving Accuracy of Rare Variant Imputation with a Two-step Imputation Approach. *Eur. J. Hum. Genet.* 23 (3), 395–400. doi:10.1038/ejhg.2014.91

Lachance, J., and Tishkoff, S. A. (2013). SNP Ascertainment Bias in Population Genetic Analyses: Why it Is Important, and How to Correct it. *Bioessays* 35 (9), 780–786. doi:10.1002/bies.201300014

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annu. Rev. Genom. Hum. Genet.* 10, 387–406. doi:10.1146/annurev.genom.9.081307.164242

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., et al. (2018). Fine-mapping Type 2 Diabetes Loci to Single-Variant Resolution Using High-Density Imputation and Islet-specific Epigenome Maps. *Nat. Genet.* 50 (11), 1505–1513. doi:10.1038/s41588-018-0241-6

Ni, G., Cavero, D., Fangmann, A., Erbe, M., and Simianer, H. (2017). Whole-genome Sequence-Based Genomic Prediction in Laying Chickens with Different Genomic Relationship Matrices to Account for Genetic Architecture. *Genet. Sel Evol.* 49 (1), 8. doi:10.1186/s12711-016-0277-y

Oliveira Júnior, G. A., Chud, T. C. S., Ventura, R. V., Garrick, D. J., Cole, J. B., Munari, D. P., et al. (2017). Genotype Imputation in a Tropical Crossbred Dairy Cattle Population. *J. Dairy Sci.* 100 (12), 9623–9634. doi:10.3168/jds.2017-12732

Orho-Melander, M., Melander, O., Guiducci, C., Perez-Martinez, P., Corella, D., Roos, C., et al. (2008). Common Missense Variant in the Glucokinase Regulatory Protein Gene Is Associated with Increased Plasma Triglyceride and C-Reactive Protein but Lower Fasting Glucose Concentrations. *Diabetes* 57 (11), 3112–3121. doi:10.2337/db08-0516

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., et al. (2012). Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-wide Association Studies. *Nat. Genet.* 44 (6), 631–635. doi:10.1038/ng.2283

Pérez-Enciso, M. (2014). Genomic Relationships Computed from Either Next-Generation Sequence or Array SNP Data. *J. Anim. Breed. Genet.* 131 (2), 85–96. doi:10.1111/jbg.12074

Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C. C., et al. (2020). Improving Imputation Quality in BEAGLE for Crop and Livestock Data. *G3 (Bethesda, Md.* 10 (1), 177–188. doi:10.1534/g3.119.400798

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., and Delaneau, O. (2021). Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels. *Nat. Genet.* 53 (1), 120–126. doi:10.1038/s41588-020-00756-0

Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., et al. (2019). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Front. Genet.* 10, 34. doi:10.3389/fgene.2019.00034

Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019). Using Imputation-Based Whole-Genome Sequencing Data to Improve the Accuracy of Genomic Prediction for Combined Populations in Pigs. *Genet. Sel Evol.* 51 (1), 58. doi:10.1186/s12711-019-0500-8

van Binsbergen, R., Bink, M. C., Calus, M. P., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., et al. (2014). Accuracy of Imputation to Whole-Genome Sequence Data in Holstein Friesian Cattle. *Genet. Selection Evol.* 46 (1), 41. doi:10.1186/1297-9686-46-41

VanRaden, P. M., Sun, C., and O'Connell, J. R. (2015). Fast Imputation Using Medium or Low-Coverage Sequence Data. *BMC Genet.* 16 (1), 82. doi:10.1186/s12863-015-0243-7

Wang, Y., Wylie, T., Stothard, P., and Lin, G. (2015). Whole Genome SNP Genotype Piecemeal Imputation. *BMC bioinformatics* 16, 340. doi:10.1186/s12859-015-0770-2

Zhang, Z., and Druet, T. (2010). Marker Imputation with Low-Density Marker Panels in Dutch Holstein Cattle. *J. Dairy Sci.* 93 (11), 5487–5494. doi:10.3168/jds.2010-3501

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership