# HAPLOTYPE ANALYSIS APPLIED TO LIVESTOCK GENOMICS

EDITED BY: Gábor Mészáros, Marco Milanesi, Paolo Ajmone Marsan and Yuri Tani Utsunomiya

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers



ISSN 1664-8714
ISBN 978-2-88966-968-4
DOI 10.3389/978-2-88966-968-4

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# HAPLOTYPE ANALYSIS APPLIED TO LIVESTOCK GENOMICS

Topic Editors:
**Gábor Mészáros,** University of Natural Resources and Life Sciences, Vienna, Austria
**Marco Milanesi,** University of Tuscia, Italy
**Paolo Ajmone Marsan,** Catholic University of the Sacred Heart, Italy
**Yuri Tani Utsunomiya,** São Paulo State University, Brazil

**Citation:** Mészáros, G., Milanesi, M., Marsan, P. A., Utsunomiya, Y. T., eds. (2021). Haplotype Analysis Applied to Livestock Genomics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-968-4

# Table of Contents

frontiers in Genetics

EDITORIAL
published: 18 May 2021
doi: 10.3389/fgene.2021.660478

# Editorial: Haplotype Analysis Applied to Livestock Genomics

*Gábor Mészáros*[1]*, *Marco Milanesi*[2,3], *Paolo Ajmone-Marsan*[4] and *Yuri Tani Utsunomiya*[3,5,6]

[1] Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Vienna, Austria, [2] Department for Innovation in Biological, Agro-food and Forest Systems-DIBAF, Università della Tuscia, Viterbo, Italy, [3] Department of Production and Animal Health, School of Veterinary Medicine, São Paulo State University (Unesp), São Paulo, Brazil, [4] Department of Animal Science Food and Nutrition - DIANA, Università Cattolica del Sacro Cuore, Milan, Italy, [5] International Atomic Energy Agency (IAEA) Collaborating Centre on AnimalGenomics and Bioinformatics, São Paulo, Brazil, [6] AgroPartners Consulting, São Paulo, Brazil

**Keywords: haplotypes, genome architecture, phasing, recombination, mutation**

## Editorial on the Research Topic

## Haplotype Analysis Applied to Livestock Genomics

The recent availability of dense panels of single nucleotide polymorphism (SNP) markers has permitted a finer investigation of genome architecture, a deeper understanding of biology and evolution, and the implementation of marker-assisted and genomic selection in livestock species. Paradigmatic examples of the use of SNP panels include understanding domestication, population diversity, inbreeding, admixture, demographic trajectories, identification of loci associated with economically important traits, and accurate prediction of breeding values. The common denominator of the vast majority of the research conducted in livestock to date has relied on analytical tools that treat genetic markers as individual and independent variables. We know, however, that genetic inheritance is driven by segments of closely interlinked nucleotides. Thus, utilizing phased multi-marker segments (i.e., haplotypes) holds the potential of improving existing models. This is particularly true in genome-wide association studies (GWAS) and genomic predictions, which are analyses that rely on the concept that information of unobserved causal variants is captured by correlation (linkage disequilibrium—LD) with nearby (observed) markers.

The potential utilization of haplotypes in genetic analysis is highly varied. Haplotypes are used in the imputation process. Imputation is the *in silico* procedure that allows us to expand upon our information on sparse SNP markers produced by existing microarray data up to the whole-genome sequence level without additional genotyping and sequencing.

Since haplotypes may serve as better proxies for causal variants than single SNP markers, the incorporation of haplotype data in genomic predictions seems promising in the absence of information on functional alleles. In extensive conditions, e.g., in the tropics, haplotypes could be used to select favorable combinations of variants in crossbreds and advanced backcross programs to retain those important for adaptation to local environmental conditions as well as those for improved production. Also, the models applied to the characterization of livestock genetic diversity could be re-designed to better estimate relationship and inbreeding, facilitate the investigation of difficult traits, as those involved in adaptation to different production systems and ecosystems, and extend the investigation of genotype-by-environment interaction. Future developments in animal breeding and genetics will be strongly based on the increasing availability of data, both molecular and phenotypic. However, our ability to dissect and understand livestock complex traits is still limited. The use of haplotypes instead of single markers and of more correct inheritance models may contribute to a better understanding of the genetics underlying livestock trait complexity and biology.

The "Haplotype Analysis Applied to Livestock" Research Topic is intended to collect empirical studies and theoretical papers exploring, evaluating, and improving the use of haplotype analysis

OPEN ACCESS

***Edited by:***
*Joanna Szyda, Wroclaw University of Environmental and Life Sciences, Poland*

***Reviewed by:***
*Mario Calus, Wageningen University and Research, Netherlands*

****Correspondence:***
*Gábor Mészáros*
*gabor.meszaros@boku.ac.at*

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

***Received:*** *29 January 2021*
***Accepted:*** *25 March 2021*
***Published:*** *18 May 2021*

***Citation:***
*Mészáros G, Milanesi M, Ajmone-Marsan P and Utsunomiya YT (2021) Editorial: Haplotype Analysis Applied to Livestock Genomics. Front. Genet. 12:660478. doi: 10.3389/fgene.2021.660478*

in livestock. After its conclusion, it managed to collect 12 articles from 89 authors, with subjects ranging from relatively straightforward diversity analyses to complex applications to unravel the genetic architecture of quantitative traits. Data used in the research studies were SNP microarray data, whole-genome sequences, or a combination of both.

Haplotype size is influenced by recombination, and consequently by the level of linkage disequilibrium (LD) existing in a population. In livestock, LD has been largely influenced by human decisions since domestication, as humans have ruled livestock demography and recent selection intensity and direction. The extent of LD in livestock is reviewed by Qanbari, with a focus on cattle and chicken populations. The study provides insights into pair-wise allelic correlations and haplotype structure in the genomes of livestock.

The concept of LD was also utilized in the development of hierarchical clustering methods for haplotype-based genomic predictions by Won et al. Their study showed increased accuracies when haplotypes, rather than single SNPs, were used to predict genomic breeding values. Importantly, the authors found that not all traits benefit from the use of haplotype data equally, and that haplotype size should be optimized on a case-by-case basis. Therefore, their results suggest a need for further improvements in methods for haplotype size selection that can consider both population structure and trait architecture.

Haplotypes are also used to detect selection signatures. An example of their utilization is shown by Aliloo et al., who sought genomic regions influencing milk production and carrying selection signals related to variation in environment, climate, and disease challenges on the African continent. The study focused on highly admixed populations of exotic and local cattle breeds. Finding selection signatures related to tolerance to African animal trypanosomiasis in Sheko cattle was the goal of Mekonnen et al. The identified genomic regions were further investigated to find promising candidate genes and over-represented genomic pathways influencing trypanotolerance. A promising regulator appears to be Caspase protease, which could play a role in the design of future intervention strategies to improve the health of cattle populations.

A wider diversity study by Luzuriaga-Neira et al. described the population structure and the relationships among South American chicken populations. Understanding the origin and assessing the extent of genetic diversity is pivotal in safeguarding and valuating animal genetic resources. Unfortunately, local chicken populations are often neglected in this respect. Admixture studies revealed the strong influence of commercial populations but also discovered unusual gene flows within the continent.

The correct identification of haplotypes based on reference genomic data is one of the cornerstones of imputation techniques. Butty et al. compared different methods designed to optimize the selection of samples to compose a reference haplotype library supporting routine imputation. In summary, if the reference set is empty, key ancestors and animals carrying common haplotypes should be the first to be included in the library. Identification of the latter can be conducted with the new Highly Segregating Haplotype method presented by the authors. As the reference set grows, rare alleles become more important, in which case newer reference samples should be selected using the Inverse Selection Method. Faux et al. presented a method for automatically matching haplotypes used for imputation, which utilizes extremely randomized trees in a random forests method. The approach holds great potential in improving imputation accuracy, as well as in developing new applications that rely on haplotype matching, such as identification of deleterious haplotypes or prediction of carriers of complex structural variants.

The power of haplotypes in capturing information about unobserved sequence variants has vast applications. For example, Meier et al. analyzed casein variants in German Black Pied cattle, Xu et al. provided insight into the genomic architecture of scrotal hernia in pigs, Oyelami et al. revealed new candidate genes and QTLs for meat quality and disease resistance in pigs, and Zhang et al. investigated hip-height and muscle development in beef cattle. The overarching theme in these studies was the use of haplotypes, instead of single SNPs, to identify relevant regions of the genome to be used in DNA-assisted breeding programs.

From a general perspective, however, the natural hereditary processes and the genetic architecture of economic traits are profoundly complex. This is in contrast with the need for the development of reliable models. The hierarchical modeling technique developed by Selle et al. combined simulation studies and cattle data with the goal to improve estimates of haplotype effects on traits of interest, especially in cases of limited data availability for rare haplotypes.

The papers included in the Research Topic "Haplotype Analysis Applied to Livestock" are examples of how these genomic segments could be utilized in a wide variety of ways. We invite you to browse and read them with the hope that they widen your overview and give you new ideas for future investigations.

## AUTHOR CONTRIBUTIONS

All authors contributed to the writing and revision of the manuscript to an equal extent.

## ACKNOWLEDGMENTS

As guest editors, we wish to thank all authors to consider our Research Topic for their publication. We also want to express our gratitude to reviewers for the contribution of their time and expertise. MM was supported by grant 2016/05787-7, São Paulo Research Foundation (FAPESP).

**Conflict of Interest:** By the time this research topic was completed, YTU was a member of the scientific board of AgroPartners Consulting.

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: 31 May 2019
doi: 10.3389/fgene.2019.00510

# Optimizing Selection of the Reference Population for Genotype Imputation From Array to Sequence Variants

*Adrien M. Butty[1], Mehdi Sargolzaei[1,2], Filippo Miglior[1], Paul Stothard[3], Flavio S. Schenkel[1], Birgit Gredler-Grandl[4,5] and Christine F. Baes[1,6]**

[1] Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada, [2] Select Sires Inc., Plain City, OH, United States, [3] Department of Agricultural, Food & Nutritional Science, University of Alberta, Edmonton, AB, Canada, [4] Qualitas AG, Zug, Switzerland, [5] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, Netherlands, [6] Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, Switzerland

Imputation of high-density genotypes to whole-genome sequences (WGS) is a cost-effective method to increase the density of available markers within a population. Imputed genotypes have been successfully used for genomic selection and discovery of variants associated with traits of interest for the population. To allow for the use of imputed genotypes for genomic analyses, accuracy of imputation must be high. Accuracy of imputation is influenced by multiple factors, such as size and composition of the reference group, and the allele frequency of variants included. Understanding the use of imputed WGSs prior to the generation of the reference population is important, as accurate imputation might be more focused, for instance, on common or on rare variants. The aim of this study was to present and evaluate new methods to select animals for sequencing relying on a previously genotyped population. The Genetic Diversity Index method optimizes the number of unique haplotypes in the future reference population, while the Highly Segregating Haplotype selection method targets haplotype alleles found throughout the majority of the population of interest. First the WGSs of a dairy cattle population were simulated. The simulated sequences mimicked the linkage disequilibrium level and the variants' frequency distribution observed in currently available Holstein sequences. Then, reference populations of different sizes, in which animals were selected using both novel methods proposed here as well as two other methods presented in previous studies, were created. Finally, accuracies of imputation obtained with different reference populations were compared against each other. The novel methods were found to have overall accuracies of imputation of more than 0.85. Accuracies of imputation of rare variants reached values above 0.50. In conclusion, if imputed sequences are to be used for discovery of novel associations between variants and traits of interest in the population, animals carrying novel information should be selected and, consequently, the Genetic Diversity Index method proposed here may be used. If sequences are to be used to impute the overall genotyped population, a reference population consisting of common haplotypes carriers selected using the proposed Highly Segregating Haplotype method is recommended.

**Keywords: dairy cattle, sequencing, imputation, haplotypes, accuracy, selection**

OPEN ACCESS

***Edited by:***
*Gábor Mészáros, University of Natural Resources and Life Sciences, Vienna, Austria*

***Reviewed by:***
*Mario Calus, Wageningen University & Research, Netherlands*
*Tom Druet, University of Liège, Belgium*
*Joanna Jadwiga Ilska, The University of Edinburgh, United Kingdom*

***Correspondence:**
*Christine F. Baes*
*cbaes@uoguelph.ca*

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

***Received:*** *15 January 2019*
***Accepted:*** *10 May 2019*
***Published:*** *31 May 2019*

***Citation:***
*Butty AM, Sargolzaei M, Miglior F, Stothard P, Schenkel FS, Gredler-Grandl B and Baes CF (2019) Optimizing Selection of the Reference Population for Genotype Imputation From Array to Sequence Variants. Front. Genet. 10:510. doi: 10.3389/fgene.2019.00510*

## INTRODUCTION

Globally, over 2.6 million cattle have been genotyped to date and the number of genotyped animals is expected to further grow in the coming years[1]. Dairy cattle genotyping is typically performed using genotype arrays of low or medium densities. Variants on genotype arrays are not selected randomly, rather they are evenly distributed over the whole genome and selected for their high level of segregation across multiple breeds (Boichard et al., 2012). Such a selection of variants has the advantage of enabling the application of the same array for multiple breeds, thus simplifying comparison between breeds. A disadvantage, however, is that they show an ascertainment bias, and variants with a low minor allele frequency (MAF) are underrepresented in genotype array data. The term "rare variants" henceforth refers to variants with a MAF lower than 0.05. Depending on the number of animals included and the alleles they carry, each genomic dataset contains its share of rare variants.

The lack of knowledge about rare variants hinders the discovery of quantitative trait loci (QTL) that, for example, appeared recently in a population through mutation (Fritz et al., 2013). Observed low MAF of variants can also be due to natural or artificial selection against an allele that has a negative impact on animal fitness or performance, thus indicating that a rare variant could be linked to a trait of interest or even a lethal malformation. An example of a rare variant associated with a disease can be found in a study by Drögemüller et al. (2009) in which a variant with a MAF of 0.03 is associated to arachnomelia (a calf malformation also called spider legs) in Brown Swiss cattle. Errors during genotyping or sequencing can also lead to wrongly identified variants with low MAF (Zhang et al., 2016).

Whole-genome sequencing can help provide better insight about rare variants (Daetwyler et al., 2014) but the costs of Next-Generation Sequencing technologies are still too high for mass sequencing of animals (Fraser et al., 2018). Imputation allows inference of whole-genome sequence (WGS) information for animals genotyped with various arrays based on complete WGS information of a reference population. The *in silico* creation of WGS from the readily available high number of genotypes enables a drastic increase in genotypic information for a large number of animals. High levels of imputation accuracy, however, are needed to allow use of the predicted genotypes for genomic evaluation or GWAS as demonstrated by Marchini and Howie (2010). The imputation from 50K to HD has been widely studied, and accurate HD genotypes are routinely imputed in dairy cattle genetic evaluation centers (e.g., Hozé et al., 2013; Ma et al., 2013; Pausch et al., 2013). Imputation to WGS variants, however, still needs to be improved. Accuracy of imputation is influenced by: (a) the size of the reference population; (b) the imputation method; (c) the relatedness between the reference and the target population; (d) the genotyping densities used, the difference in the number of variants and the linkage disequilibrium between SNP of both low- and high-density panels; (e) the MAF of the variants considered; and (f) the genetic diversity of the reference population. A thorough review of the factors influencing accuracy of imputation in livestock species was written by Calus et al. (2014). The selection of animals to include in reference populations influences many of these parameters and is thus of high importance. Druet et al. (2014) stated that as the MAF of variants becomes lower, the method used to select animals to be included in the reference population becomes more important.

The international dataset created under the scope of the 1,000 Bull Genomes Project (Daetwyler et al., 2014) is a possible reference set for imputation of cattle array genotypes to WGS. Up to Run 5 of this project, most animals sequenced were selected for their high genetic contribution to the population of their breed (Goddard and Hayes, 2009). These key ancestors carry most of the common variants for the populations they were selected from but lack information on rare variants. Pausch et al. (2017) showed that overall average imputation accuracy of array genotypes to the variant list from the 1,000 Bull Genomes Project was greater than 90%, but that the imputation accuracy of rare variants did not reach 70%. Low imputation accuracy of rare variants hinders the discovery of causal variants, not only for highly polygenic traits, but also for recent mutations that lead to malformations or loss of fitness (Li et al., 2011). Zhang et al. (2017) showed that the lack of accuracy in imputation of variants with low MAF also limits the success of genomic selection, particularly for health traits. Improved accuracy of WGS imputation will not only increase the probability of discovering causal variants for newly recognized diseases or malformations, but will also enable more precise categorization and selection of variants for routine genomic selection programs for traditional and novel traits.

Various methods have been proposed to select animals for sequencing, the first of which relied solely on pedigree information and targeted influential ancestors of the population of interest. Boichard et al. (1997) developed a method to identify animals that have the greatest genetic contribution to a population based on its pedigree information. This method was implemented and widely distributed using the software PEDIG (Boichard, 2002). The key ancestors method, developed thereafter, relied on the numerator relationship matrix of the genotyped population of interest and also aimed to maximize the proportion of genes of the population captured by the selected animals (Goddard and Hayes, 2009). As the number of genotyped animals increased, selection methods have been adapted to consider genomic information. Methods were proposed which emphasize selection of animals carrying common haplotypes. Druet et al. (2014) presented a method maximizing the number of haplotypes selected. The key contributors method presented by Neuditschko et al. (2017) defines animals as informative based on the genomic relationship matrix of the population and aims to select individuals within possible subpopulations. Another selection method developed by Gonen et al. (2017) involved the algorithm AlphaSeqOpt that not only selects individuals that, together, represent the maximum haplotype diversity of a population, but also suggests different sequencing coverages in situations where the sequencing costs are predetermined. An optimized version of AlphaSeqOpt was proposed by Ros-Freixedes et al. (2017), similarly considering situations where the sequencing costs were predetermined, but

[1] https://queries.uscdcb.com/Genotype/cur_ctry.html, last accessed 2018-09-23

additionally targeting haplotypes instead of individuals. This method was shown to improve the phasing accuracy of the reference population it formed, even if it still maximizes the proportion of the total haplotypes included. In contrast to the previously described methods, which target representative animals of a population, the Inverse Weighted Selection Method (Bickhart et al., 2016) was developed to prioritize individuals for their higher genetic diversity at the haplotype level, classifying animals based on the rarity of their haplotypes. The Inverse Selection Methods was shown to allow sequencing of the maximum number of haplotypes with the fewest number of animals. In this study, two new selection methods are presented: the optimized Genetic Diversity Index (GDI), which targets animals carrying more rare haplotype alleles than the average individuals and the Selection of Highly Segregating Haplotype (HSH), which aims at selecting animals whose haplotypes are highly segregating, but not selected yet. The GDI method aims to improve the accuracy of imputation of rare variants through selection of animals that, together, carry the most different haplotypes, whereas the HSH should help to improve overall accuracy through selection of animals that carry the highest segregating haplotypes not previously sequenced.

The objectives of this study were: (1) to describe two innovative methods to select animals for sequencing from a population, and (2) to compare these methods to two previously described selection methods: the key ancestors method and the Inverse Weighted Selection method.

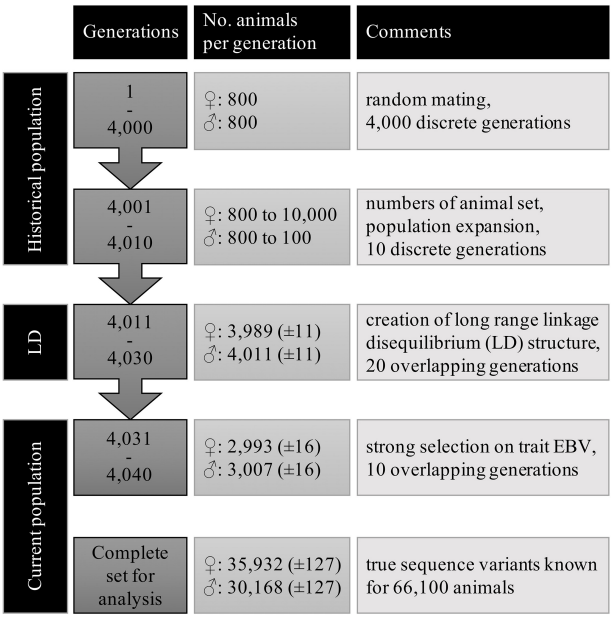


**FIGURE 1 |** Structure and number of animals of the simulated populations.

## MATERIALS AND METHODS

Firstly, the WGS and high-density array genotypes of a dairy cattle population were simulated. Secondly, reference populations were created by selecting animals based on four different methods. Thirdly, a set of simulated target animals were imputed using the different reference populations. Finally, the imputation accuracies of the different methods were compared to each other considering sets of variants, defined depending on their MAF (**Figure 1**).

### Simulation

#### Population Structure

Large scale WGS data was simulated with the QMsim program (Sargolzaei and Schenkel, 2009) using three subsequent populations. First, a historical population was simulated to create linkage disequilibrium (LD) between the variants. Then, a second population, termed LongRangeLD, was simulated to increase long-range LD between variants. Finally, a third population (CurrentPop) was simulated for downstream analysis. CurrentPop simulated the latest years of dairy cattle breeding, in which few selected sires were used heavily in the breeding population.

The historical population considered an equal number of individuals from both sexes, discrete generations, random mating at the gametic level, no selection, and no migration. A total of 800 males and 800 females were simulated for 4,000 generations to achieve mutation-drift equilibrium. Ten further historical generations were generated expanding the population to 10,100 animals. In the last generation of the historical population, there were 100 males and 10,000 females.

The founders of LongRangeLD were all animals of the last generation in the historical population, after which each generation was composed of 8,000 animals. Through using different replacement rates, the 20 generations of this population overlapped. The total LongRangeLD population was composed of 168,100 animals. Founders of CurrentPop were 100 males and 4,000 females from the last generation of LongRangeLD and also 4,000 more females from the second-last generation of LongRangeLD. The 10 generations of this population had 6,000 animals and overlapped too. Finally, the complete population for downstream analysis had 66,100 animals, of which 30,168 (±127) were males. Migration was not simulated in any scenario. Further parameters used for both LongRangeLD and CurrentPop are presented in **Table 1**. The complete simulation process was replicated 10 times and the results reported are averages of the replicates.

#### Genome

Gene-dropping simulation was completed using QMSim (Sargolzaei and Schenkel, 2009). The same genome was simulated for all populations. Cattle autosomal chromosomes were simulated with a length that followed the results presented by Bohmanova et al. (2010) and summed up to a total of 2,496 cM. Bi-allelic markers and QTL were randomly distributed over all chromosomes with equal MAF in the first historical generation. The QTL effects were sampled from a gamma distribution with a shape parameter of 0.4, following the results obtained by Hayes and Goddard (2001). The number of crossovers per chromosome was sampled from a Poisson distribution with mean equal to

**TABLE 1 |** Parameters used for the simulation of the populations LongRangeLD and CurrentPop.

| Parameter | LongRangeLD | CurrentPop |
|---|---|---|
| Number of generations | 20 | 10 |
| Litter size | 1.0 | 1.0 |
| Sire replacement rate | 0.5 | 0.5 |
| Dam replacement rate | 0.3 | 0.3 |
| Mating design | Positive assortative on phenotypes | Positive assortative on EBV |
| Selection design | On phenotypes | On EBV |
| Culling design | Age | Low EBV |
| EBV estimation method | None | BLUP using the true additive genetic variance |
| Number of traits | 1.0 | 1.0 |
| Heritability | 0.3 | 0.3 |
| Phenotypic variance of trait | 1.0 | 1.0 |

the chromosome length in centimorgans. The probability of a second crossover within 25 cM of a first recombination event was, therefore, lower depending on the proximity of crossovers. The mutation rate of the markers and the QTL was assumed to be $10^{-4}$. For each replicate, 8,622,767 markers and 4,000 QTL were generated.

## Introduction of Genotyping Error and Selection of Variant Subsets

Selection of markers in the simulated data was performed to ensure that the MAF distribution followed that observed in the real data, described below. From all simulated variants, a first subset representing WGS was selected that contained all QTL. Then two subsets of the WGS were selected, which simulated high-density (HD) and medium density (50K) array genotype variant panels. In contrast to the WGS set, no QTL were allowed in the HD and 50K variant panels. Minor allele frequencies considered at this stage were computed considering a random sample of 30,000 animals from CurrentPop. Those animals represented 45% of the total population.

Real data comprised 425 Holstein (HOL) and 25 Red-Holstein animals from Run 5 of the 1,000 Bull Genomes Project (Daetwyler et al., 2014), 2,946 HOL animals (males and females) from the Canadian Dairy Network database (as of August 2017), and 36,157 HOL bulls with a North American identification tag born after 2010 for the WGS, HD, and 50K panels, respectively. The real WGS set was filtered for a minor allele count of 1 and was composed of 31,787,016 bi-allelic variants. Variants with a MAF lower than 0.1% were filtered out from the HD dataset. The real HD genotypes contained information for 587,817 bi-allelic variants. The same filter for variants with a MAF lower than 0.1% was applied to the 50K panel leading to 44,347 bi-allelic variants.

The number of selected variants per chromosome was proportional to the number of variants found in the real data. Variants were distributed by MAF in 50 bins. The sampling of the variants occurred randomly within the bin-by-chromosome groups with the function *sample()* in R, version 3.4.3 (R Core Team, 2017). The final simulated data was composed of 3,235,171 (±155,117), 571,661 (±6) and 44,288 (±0) variants for WGS, HD, and 50K, respectively. Genotyping error was introduced in the WGS based on error rates observed by Baes et al. (2014) using the HaplotypeCaller function of the Genome Analysis Toolkit with a multi-sample approach (McKenna et al., 2010; **Table 2**). Missing data was also added at this stage. Inclusion of genotyping errors and missing data in the genotypes was done using snp1101 (Sargolzaei, 2014).

## Creation of the Reference Populations and the Validation Set

Groups of 50, 100, 200, 400, 800, and 1,200 animals were created from one pool of candidates using four selection methods. This pool of candidates was composed by all males of the CurrentPop and contained 30,027 (±108) bulls. As the 50K chip represents the preferred SNP chip for bull genotyping, animals were selected on their 50K haplotypes. The groups of selected bulls were later used as the reference populations for imputation from HD to WGS genotype density. Although imputation was done from HD to WGS genotype densities, selection of animals, when performed based on genotypes, was run on the 50K array panel to mimic again real situations, where the majority of the individuals would have only 50K genotype information. Haplotypes were defined as non-overlapping segments of 20 contiguous SNP of the 50K SNP panel throughout the study and had an average length of 1,082,875 bp (±264,426 bp). The same candidate pool was available for each method, so the same animal could be selected by multiple methods.

The selection methods were: (1) the key ancestors method, which used the additive genetic relationship matrix; (2) a combination of the newly developed Genetic Diversity Index and the simulated annealing algorithm (Kirkpatrick et al., 1983; Černý, 1985); (3) the Inverse Weighted Selection method (Bickhart et al., 2016); and (4) a second novel method aiming to select highly segregating haplotypes in the genotyped population that are not carried by any animal of the population of interest already sequenced. These methods are described in more details next. The 5,000 youngest animals (males and

**TABLE 2 |** Rate of genotyping change as introduced in the simulated whole-genome sequence genotypes.

| | | Simulated genotypes including genotyping error and missing values | | | |
|---|---|---|---|---|---|
| | | AA | AB | BB | –/– |
| True genotypes | AA | 0.639 | 0.004 | 0.001 | 0.356 |
| | AB | 0.011 | 0.970 | 0.000 | 0.019 |
| | BB | 0.002 | 0.004 | 0.976 | 0.018 |

*As an example, 1.1% of the simulated AB genotypes were changed to AA genotypes. Values were retrieved from the study by Baes et al. (2014).*

females) from CurrentPop that were not selected during the creation of the reference groups composed the target population of the imputation.

## Selection Methods

Selection of key ancestors was the method of choice to select the first animals sequenced in populations, as a representative genotyped group of animals from the population of interest was not needed (Daetwyler et al., 2014). This key ancestor method (AMAT) was chosen for comparison because of its frequent use and because it had indirectly a similar aim than the novel Selection of Highly Segregating Haplotype (HSH) method proposed here, i.e., selection of carriers of commonly found variants. Shortly after the first draft of the optimized Genetic Diversity Index (GDI) proposed here was designed, the paper of Bickhart et al. (2016) was published that presented the Inverse Selection Method (IWS). As GDI, this method aimed at selecting animals that are genetically more diverse in the pool of candidates. IWS seemed thus to be fairly comparable to GDI and was chosen to be included in this study. Other methods of animal selection for sequencing considered other objectives such as sequencing some animals at different coverages or combination of genotyping and sequencing, given a limited budget. In contrast, this study only considers situations where a given number of animals to sequence is given. Focusing on methods with similar aims than the novel methods proposed here seemed a way to allow for an in-depth analysis of them, as for example, differentiating accuracies of imputation of variants with different MAF.

### Selection of Key Ancestors

The AMAT method aimed to identify animals explaining most of the genetic variation of a population following the equation $p_n = A_n^{-1*} c_n$ where $p_n$ was a vector of the proportion of gene pool captured by the $n$ selected animals, $A_n^{-1}$ was the inverse of the numerator relationship matrix of the $n$ selected animals, and $c_n$ was a vector of the average relationships of the $n$ selected animals with the entire population (Goddard and Hayes, 2009).

### Inverse Selection Method

The IWS method developed by Bickhart et al. (2016) prioritized sequencing of rare haplotypes following the equation $Index = \sum_{i=1}^{NHAP} f_i^2 - 2f_i + 1$ where $NHAP$ was the number of haplotypes and $f_i$ was the frequency of haplotype $i$ in the population. This inverted parabolic function gave a high index value to individuals carrying haplotype alleles with low frequencies, as higher frequencies led to higher penalization (through the term $-2f_i$). The computation of this index was iterative: (1) select the animal with the highest index; (2) recalculate the index of the remaining candidates without considering the haplotypes present in the genotypes of selected animals; and (3) pick out the next animal with the best new index. This method was used as it is implemented in the software program snp1101 (Sargolzaei, 2014).

### Optimized Genetic Diversity Index

Relying on a probabilistic optimization algorithm –simulated annealing (Kirkpatrick et al., 1983; Černý, 1985) – the proposed GDI method optimized the count of unique haplotypes of a group of animals composed of all previously sequenced animals and a defined number of sequencing candidates. The simulated annealing algorithm was developed to find the global optimum of a dataset with multiple local optima. The GDI of the whole group of animals was optimized with the simulated annealing algorithm permuting one candidate at a time and recalculating the index. The GDI was computed by summing the count of unique haplotype alleles present within a group of animals following the equation $Index = \sum_{i=1}^{NHAP} unique(HAP_i)$, where $NHAP$ was the number of haplotype blocks and $HAP_i$ were the haplotype variants in block $i$. **Figure 2** gives an example of the index calculation based on five animals and four haplotypes. This method was also used as implemented in the program snp1101 (Sargolzaei, 2014).

### Selection of Highly Segregating Haplotypes

To identify animals with the highest contribution to the population, the novel HSH method based on haplotype diversity was developed. The method had the following steps: (1) a haplotype library was created for all selection candidates using non-imputed genotypes. Haplotypes that appeared in less than 10 animals were discarded to reduce errors in the computation of their frequencies due to phasing error or haplotypes from other breeds; (2) contribution of each animal to the haplotype library based on the haplotypes' frequency was calculated following the equation, $Index = \sum_{i=1}^{NHAP} f_i$, where $NHAP$ was the number of haplotypes and $f_i$ was the frequency of haplotype $i$ in the population. The animal with the highest Index value was then selected and; (3) frequencies of all haplotypes present in the selected animal were multiplied by a factor of 0.75 to penalize these already captured haplotypes. The factor for penalization is decided based on haplotypes frequency distribution in Holstein. Then the second most influential animal was selected based on highest contribution from the penalized haplotypes frequencies of all haplotypes it carries were multiplied again by the same factor of 0.75. After selecting an influential animal, total haplotype coverage (i.e., prevalence) was calculated for the new group of selected candidates. The process was repeated until the desired number of animals was selected, increasing the number of unique haplotypes selected with each animal, but avoiding selection of possible outliers (which carry many low-frequency haplotypes from another breed), for example from crossbred individuals as long as any non-outlier animals were still in the selection pool. Because the most frequent haplotypes were penalized first, the next animal chosen tended to carry haplotypes that are less frequent in the library or population. This method was also used as it is implemented in the software program snp1101 (Sargolzaei, 2014). The HSH method could accommodate any situation where some animals were previously sequenced, as the choice of the next influential animal is a

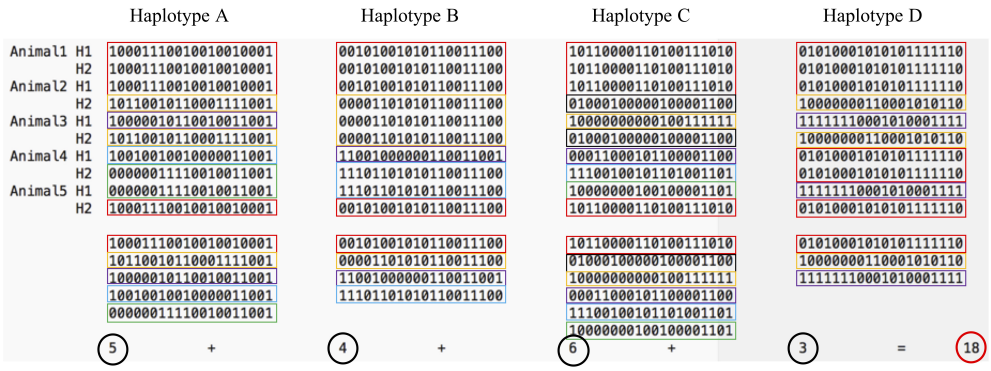


**FIGURE 2 |** The Genetic Diversity Index is the sum of the unique haplotypes found in a group of animals. On this figure, five animals carry in total 18 unique haplotypes (5 variants of haplotypes A, 4 variants of haplotype B, 6 variants of haplotype C, and 3 variants of haplotype D). Colors highlight the unique haplotype alleles of each haplotype block.

function of already selected animals. Therefore, although the selected candidates may be different depending on which initial list of sequenced animals is used, the overall contribution to the population haplotypes should change only minimally.

## Measures of Diversity in the Reference Population

The level of genetic diversity was compared between reference populations. Next to the number of segregating variants as presented by Pluzhnikov and Donnelly (1996), the proportion of the total number of unique haplotypes alleles found in the candidate groups that were also found in the individuals selected for sequencing were used to compare the level of genetic diversity of the reference population of each scenario. The proportion of the rare haplotypes found in differently selected individuals was computed using the R package GHap (Utsunomiya et al., 2016). First, all haplotypes found within the candidates were identified. Second, the frequencies of the haplotypes within the candidates were computed. Finally, the proportions of haplotypes found in different groups were calculated. Following the construction of haplotypes when the animals were selected, haplotypes were built here again with 20-SNP windows and without overlap.

## Principal Component Analysis

Principal components analysis (PCA) is a statistical method that, when applied to genotypic data, allows detection of its structure (Ely et al., 2010). PCA was run on 50K genotypes of the candidate pool to determine the structure of the simulated population. This analysis was conducted using the implementation presented by Abraham and Inouye (2014) and available in snp1101 (Sargolzaei, 2014) with the following parameters: a maximum of 50 iterations were allowed, 40 principal components were computed and only variants with a MAF equal or higher than 0.01 were considered.

## Imputation

Following results presented by Whalen et al. (2018), the combination of the phasing software Eagle version 2.3.5 (Loh et al., 2016) and the imputation software Minimac3 (Das et al., 2016) – two programs developed for analysis of human data for which little to no family information is available – was used without pedigree information on the differently created reference populations to impute one set of target animals. Both software programs were used in their default mode. A linear genetic map of 1 cM per Mb was used to approximate the average recombination rate at phasing. From this step onward, all genotypes were reduced to the 10 first simulated chromosomes to reduce computation time and memory load. Imputed genotype calls only were used, not the genotype probabilities.

## Measure of Imputation Accuracy

Imputation accuracy was computed on multiple sets of variants for each scenario. Variants were distributed over multiple bins, depending on the MAF observed in the true genotypes of the target population of each simulation replicate. Two non-overlapping subsets containing common (MAF $>$ 0.05) or rare (MAF = 0.05) variants were created, as well as a set of adjacent SNP bins. Variants were distributed following their MAF in the bins with boundaries at 0.00, 0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.50. The bins were created to allow for the higher bound MAF to be included but not the lower bound. The composition of all bins is represented in **Figure 3**.

Imputation was evaluated at a per SNP basis by the squared correlation between the true and imputed genotypes and average. This accuracy measure, called allelic $R^2$ by Browning and Browning (2009), is advantageous, as it is independent of the MAF of the variants imputed. Correlations between true and imputed genotypes were checked to ensure that negative correlations were not present so that no variants were filtered out at this stage. Accuracies of variants that were not segregating anymore after imputation were set to zero. Genotype concordance rates between all variants of the true and imputed genotypes were also computed on all variants. This measure represents the proportion of genotypes that are correctly imputed and allowed for evaluation of the imputation on a per animal basis.

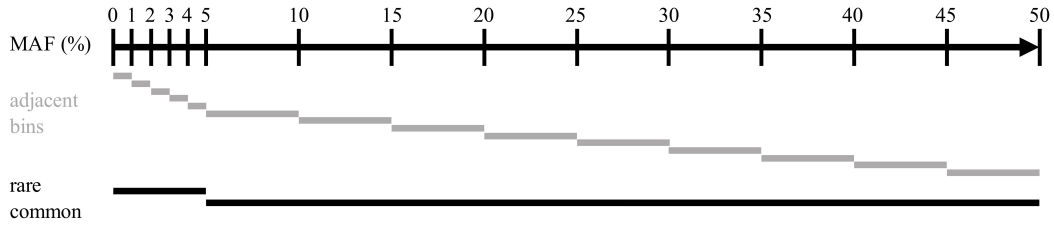


**FIGURE 3 |** Representation of the distribution of the intervals of minor alleles frequencies used for assessing imputation accuracy.

## Performance of the Haplotype-Based Selection Methods With Crossbred Animals in the Candidate Pool

Selection of animals for sequencing is often run in one population at a time. Depending on the quality of the data recording, a proportion of the animals declared to be purely from one population may be crossbred or from another population. It is important that the method of selection avoids selecting individuals that are not part of the population of interest. BovineSNP50 genotypes of 16,420 Holstein and 2,920 Jersey (JE) males born after 2011 were retrieved from the Canadian Dairy Network database to create pools of 5,840 selection candidates with different degrees of admixture as presented on the horizontal axis of **Figure 4**. From the complete dataset, animals were randomly selected to enter each pool. The IWS, GDI and HSH methods were then used to select 100 animals out of each pool and the number of JE animals that were picked were counted.

## Statistical Tests of Average Differences Between Scenarios

After testing for the normality of the replicates within methods-by-reference size scenarios with Shapiro–Wilk tests, Kruskal–Wallis Rank Sum tests, and Wilcoxon Rank Sum statistical tests were performed for each MAF category to determine significant differences in accuracies among all methods or pairwise, respectively. The Bonferroni correction was used to adjust for multiple comparisons for an experimental-wise significance level of 0.05.

# RESULTS

## LD Structure, MAF Distribution and Structure of the Simulated Population

A rapid decrease in LD over increasing genomic distance was observed in both real and simulated genomic data (**Figure 5**). The high level of LD at distances shorter than 100kb in the real Holstein population already described by Sargolzaei et al. (2008) is mimicked in the simulation. Rare variants comprised 52.43% (±2.2%) of the WGS variants over the replicates. Principal component analysis showed a compactly distributed population on the two first components, which explained 6.11% of the total genomic variance (**Figure 6**). Spearman's rank correlation between the first principal component and the generation of the animals was 0.87 (data not shown). Density curves of the MAF over the generations of the simulated population showed that an increasing number of variants became rare (**Figure 7**).

## Haplotype Coverage in the Reference Population

The number of segregating variants and the proportion of unique haplotype alleles found in each reference population had a correlation of 0.68 ($P < 0.0001$). Increasing the number of animals in the reference population led to an increased proportion of unique haplotypes covered (**Figure 8**). Overall, haplotypes coverage ranged from 8.6% of the total haplotypes from the scenario with 50 animals selected on the basis of HSH, to 35.5% in the scenario including 1,200 animals selected through GDI. The reference groups created following the AMAT and HSH methods captured a lower proportion of the total haplotypes than reference populations created following the IWS and GDI methods. The proportion of haplotypes with a frequency equal or below 5% that were selected in each reference group followed the proportion of total haplotype selected.

## Overlap in Selection

The same pool of candidates was made available for selection for each method and reference size so that the same animals could be selected by multiple methods. The proportions of animals present in two groups for each reference size are shown in **Table 3**. Overlaps were higher between AMAT and HSH and between IWS and GDI. Small reference population sizes led to a higher proportion of animals found in multiple reference populations, with a maximum of 26% of animals found in common between the reference groups of AMAT and HSH that contained 50 animals in total. GDI did not have any overlap with AMAT or HSH for groups containing 50 and 100 animals. The overlap in the selected references of 100 individuals can be observed in **Figure 6** where plusses, representing the animals selected with IWS, and crosses, representing the animals selected with GDI, are superposed.

## Selection With Possible Crossbred Animals

With a pool composed of animals from two populations in a 50:50 ratio, no genotype-based method could avoid selecting at least half of them from the JE population (**Figure 4**). Differences were observed between methods in the more realistic scenarios

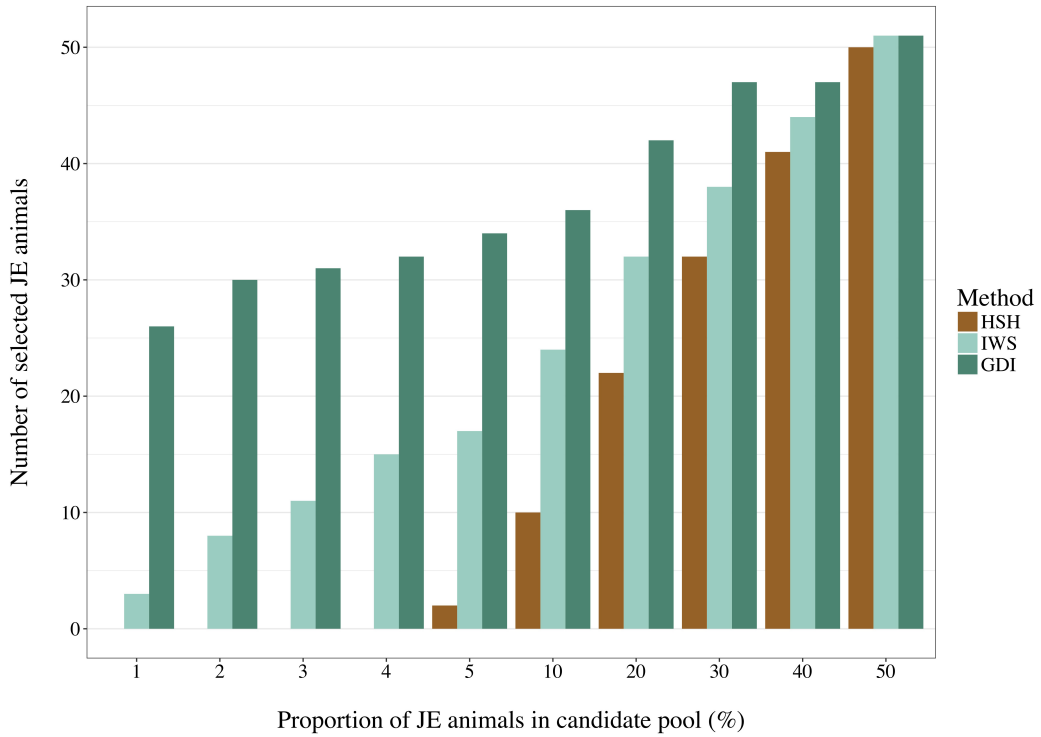


**FIGURE 4 |** Number of Jersey (JE) animals selected by the Highly Segregating Haplotype selection (HSH), the Inverse Weighted Selection (IWS), and the Genetic Diversity Index (GDI) methods from candidate pools with different proportion of JE animals.

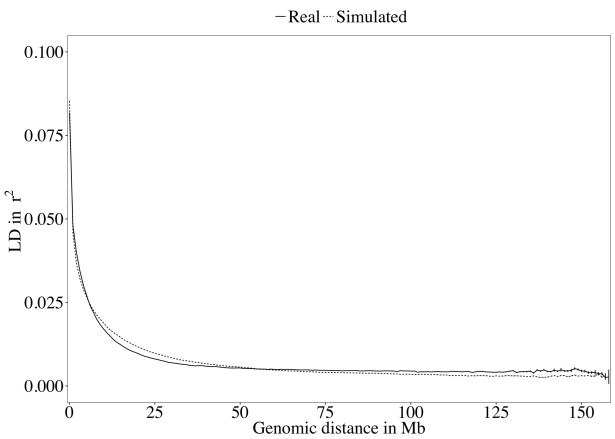


**FIGURE 5 |** Decay of linkage disequilibrium (LD) over genomic distance in the real and the simulated sequences.

with a proportion of 5% or less non-target animals. HSH did not select any JE animals until they comprised 5% of the candidate pool. In contrast, GDI already selected 58 JE animals when JE comprised 1% of the candidate pool. The 58 JE selected in this scenario were 45% of all JE animals present in the pool. In the scenarios with a candidate pool composed of 5% or less JE animals, IWS consistently selected only 5% of the JE animals.

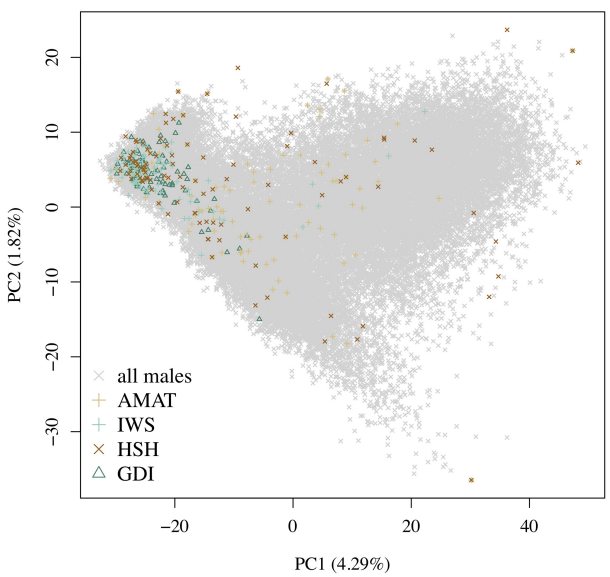


**FIGURE 6 |** Distribution of the different groups of animals on the first and second principal components. The variance explained by the components is given in brackets. Gray crosses represent all the candidates, the green triangles are the animals selected by the key ancestors (AMAT) method, the purple plusses are the animals selected by the Inverse Weighted Selection (IWS) method, the green plusses are the animals selected by the Highly Segregating Haplotypes selection (HSH) method, and the blue crosses are the animals selected with the Genetic Diversity Index (GDI) method.

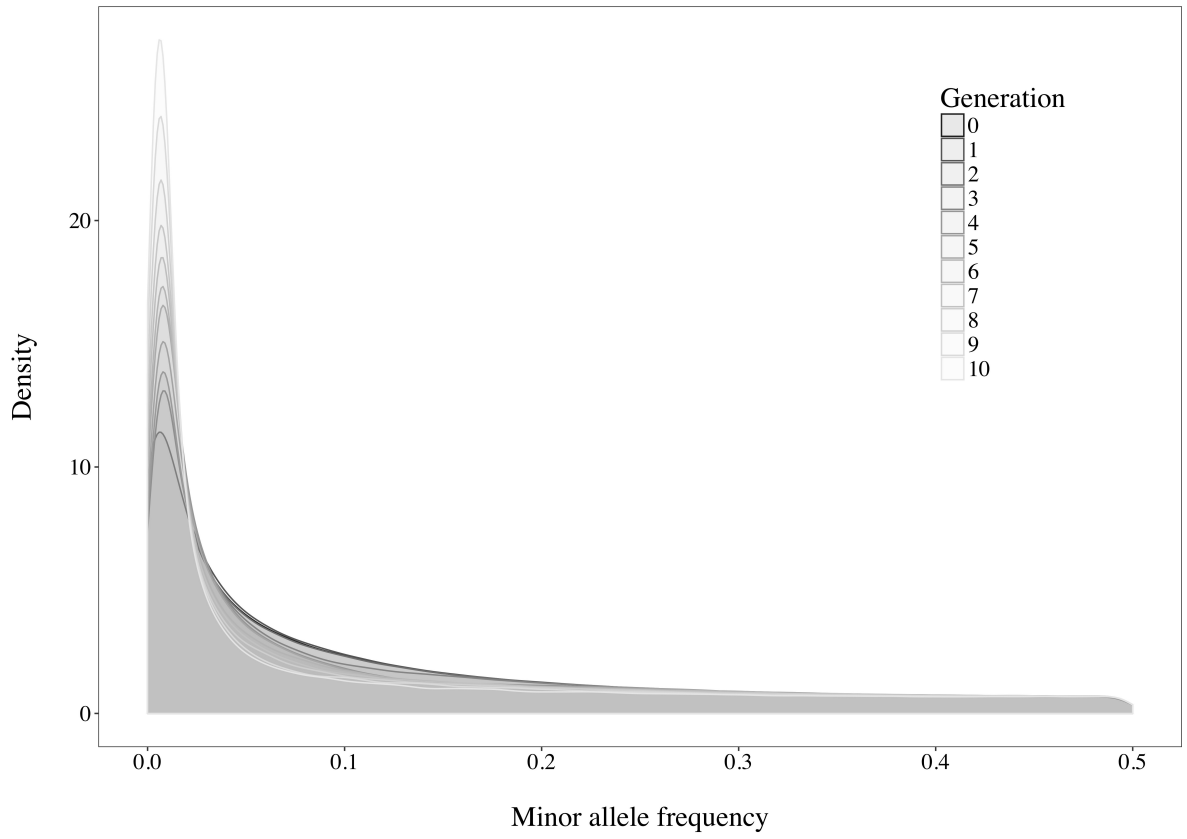


**FIGURE 7 |** Density curves of the minor allele frequencies observed in the simulated population over the generations.

## Accuracy of Imputation

Accuracies of imputation were observed on all variants and on two non-overlapping subsets: the rare variants with a MAF below 0.05 and the common variants with a MAF equal or above 0.05. Results are presented about these sets in the following order: first, all variants, then the rare variants and finally the common variants as the later showed re-ranking in comparison with the two other groups.

Considering all variants, accuracy of imputation reached values between 0.55 and 0.85, depending on the method used to create the reference groups and their sizes. Increasing the number of animals in the reference population led to corresponding increases in accuracies. **Table 4** shows the accuracies of imputation reached in scenarios with 50, 200, and 1,200 reference animals selected by the four methods and across all adjacent MAF bins. In general, AMAT and HSH reached lower accuracies than IWS and GDI. The differences in accuracies, however, were smaller when the reference population size increased (**Figure 9**). In the scenario in which only 50 animals composed the reference population, IWS and GDI had the highest accuracies and were not significantly different ($P > 0.05$). AMAT had a significantly lower accuracy and the accuracy of HSH was even lower than that of AMAT ($P < 0.0001$) (**Table 4**). By increasing the size of the reference population to 100, 200, or 400 animals, differences in accuracy between AMAT and HSH were small, so that only two groups of methods, AMAT/HSH and IWS/GDI, could be differentiated. With reference groups of 800 and 1,200 individuals, only GDI and AMAT were significantly different ($P < 0.0001$), where GDI had the highest accuracy (0.944). Genotype concordance rates reached values above 0.96 in all cases (**Figure 10**). Significant differences between methods were only observed with reference populations comprising 50, 100, or 200 animals. Concordance rates were higher when animals were selected with HSH or AMAT than with IWS or GDI for reference sizes of 50 or 100 animals ($P < 0.0001$). Only the concordance rate of GDI for a reference population comprised of 200 animals was significantly lower than any other ($P < 0.0001$).

When the accuracies of imputation were estimated on rare variants only, accuracies reached values between 0.33 and 0.76, but the rank of the methods from best to worst was consistent with results based on all variants (IWS/GDI > AMAT/HSH), and significant differences were also observed at any reference populations size ($P < 0.0001$). With reference size of 1,200 individuals, differences were only found between AMAT vs. HSH and AMAT vs. GDI, where AMAT had lower accuracy in both contrasts. In contrast, when only common variants were considered, the ranking was reversed: AMAT and HSH produced significantly higher accuracies than IWS and GDI ($P < 0.0001$). Accuracies took values as high as 0.99 and were

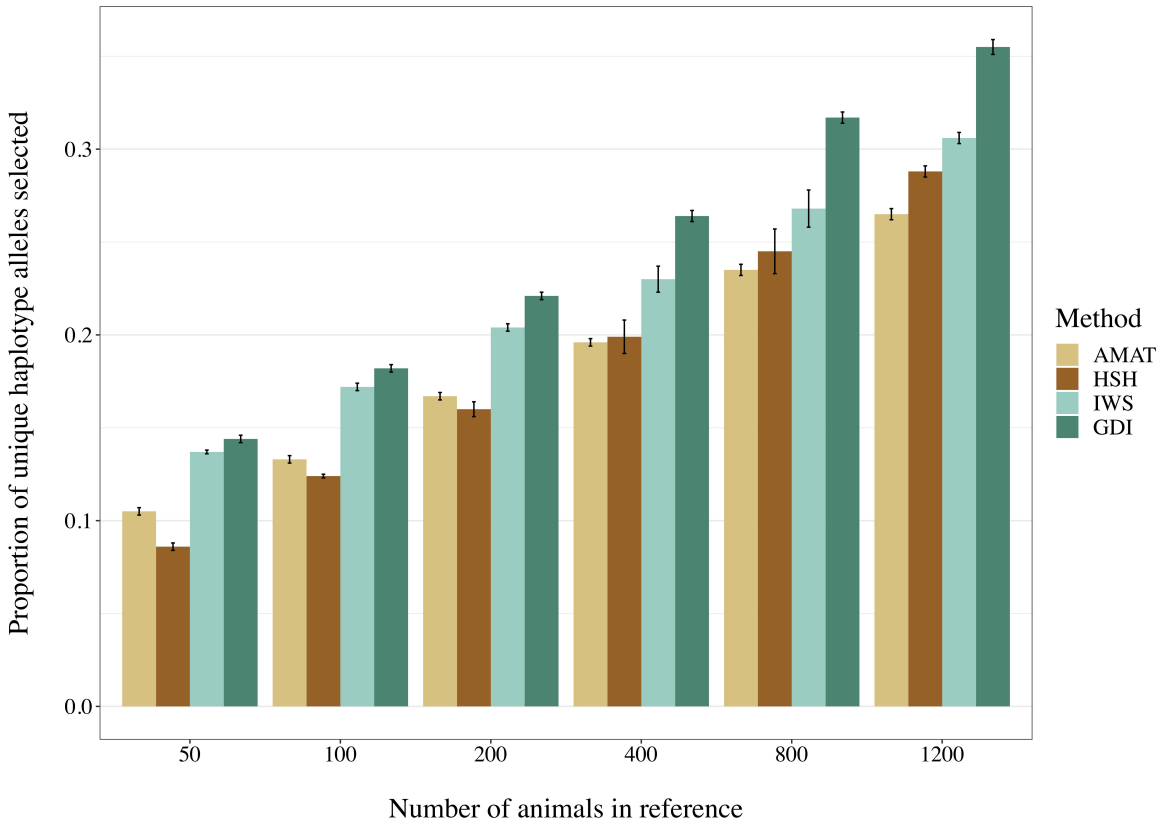


**FIGURE 8 |** Selected proportion of unique haplotypes from the total haplotype library found in the reference group created with different selection methods. The methods compared are the key ancestors (AMAT), the Highly Segregating Haplotype selection (HSH), the Inverse Weighted Selection (IWS), and the Genetic Diversity Index (GDI).

never below 0.84. With a reference population of 50 animals, HSH reached a greater accuracy than AMAT and both were better than IWS and GDI. With reference sizes of 100 and 200, significant differences were again observed between the groups of methods AMAT/HSH and IWS/GDI ($P < 0.0001$). With 400 animals as reference, the accuracy reached by GDI was significantly lower than the other methods. Scenarios where 800 and 1,200 animals composed the reference population did not show difference in accuracy value before the fourth decimal. Although no change in the values was observed for these scenarios (**Table 4**), variances between replicates were very small (standard deviation $< 0.004$), therefore testing the methods against each other still led to significant results after correction for multiple testing.

Distribution of the variants into 14 adjacent bins allowed a more precise evaluation of the effect of the reference composition on the imputation accuracy. With no exception, increased MAF led to increased accuracy values. For example, accuracies increased from 0.21 to 0.94 when using a reference group of 50 individuals selected with AMAT (**Figure 11**). Only pairs of contiguous MAF bins were analyzed and no significant differences within reference size-by-method scenario were found in imputation accuracy of variants with a MAF greater than 0.3 ($P > 0.05$).

## DISCUSSION

In the first step of this study, the WGSs of a dairy cattle population were simulated. They were compared to currently available real Holstein sequence data to ensure they mimicked observed levels of LD and MAF distribution. In the second step, reference populations of different size were created with animals selected by both proposed novel methods as well as two other methods presented in previous studies. The selection methods were assessed with respect to their propensity to select animals that might not be of the population of interest, the genetic diversity of the groups of animals picked, and the distribution of those over generations. Finally, accuracies of imputation were compared for imputation runs with the different reference populations. The differentiation of imputation accuracy of variants with specific MAF allowed comparison between the strengths and weaknesses of each method of selection.

Different software programs were developed to simulate genomic information such as AlphaSim (Faux et al., 2016), ms2gs (Pérez-Enciso and Legarra, 2016), and QMsim (Sargolzaei and Schenkel, 2009). With its highly flexible genome and population configuration system, QMsim allowed for simulation of a great number of WGSs that had a LD structure properly following the parameters of the real data available. With the aim of

**TABLE 3 |** Proportion of animals overlapping between selection methods in reference populations of different sizes.

| Size | | Method | | |
|---|---|---|---|---|
| | | **AMAT** | **HSH** | **IWS** |
| **50** | | | | |
| | HSH | 0.26 | | |
| | IWS | 0.04 | 0.00 | |
| | GDI | 0.00 | 0.00 | 0.08 |
| **100** | | | | |
| | HSH | 0.23 | | |
| | IWS | 0.03 | 0.01 | |
| | GDI | 0.00 | 0.00 | 0.1 |
| **200** | | | | |
| | HSH | 0.20 | | |
| | IWS | 0.05 | 0.03 | |
| | GDI | 0.01 | 0.02 | 0.14 |
| **400** | | | | |
| | HSH | 0.13 | | |
| | IWS | 0.04 | 0.06 | |
| | GDI | 0.02 | 0.03 | 0.09 |
| **800** | | | | |
| | HSH | 0.09 | | |
| | IWS | 0.03 | 0.12 | |
| | GDI | 0.03 | 0.06 | 0.12 |
| **1,200** | | | | |
| | HSH | 0.08 | | |
| | IWS | 0.03 | 0.16 | |
| | GDI | 0.04 | 0.09 | 0.12 |

*The methods were the key ancestors (AMAT), the selection of Highly Segregating Haplotypes (HSH), the Inverse Weighted Selection (IWS), and the Genetic Diversity Index (GDI).*

simulating a Holstein population, only sequences of Holstein animals from the 1,000 Bull Genomes Project Run 5 were retrieved. These animals were mostly sequenced because they had a great genetic contribution to their population (Daetwyler et al., 2014). Although they are considered representative, these animals became influential as they were used heavily for breeding in their population and probably had a high genetic merit. They may, in fact, carry alleles at frequencies different from those in the overall population. This difference between the influential animals and the complete population limits the possible true closeness of the simulation with the whole real Holstein population. The LD level of the simulated sequences followed the real observed LD decay (**Figure 5**). Similarly, the distribution of the variants used in this work in MAF bins followed the distribution observed in real datasets.

Once the sequence was simulated, multiple reference populations were created by selecting animals using methods of selection that can be divided into two groups: AMAT and HSH, which mainly target animals that are carriers of commonly found haplotypes, whereas IWS and GDI are methods aiming to maximize the selection of animals carrying more haplotype alleles. Moreover, although AMAT keeps on searching for commonly found haplotypes, the penalization of those implemented in HSH leads to a shift from the search of commonly found haplotypes to rare ones. Through this shift, not only selection of common, but also of rare haplotypes is optimized. This shift, however, is highly dependent on the size of the candidate pool and the number of animals to be selected, as the increasing ratio of selected animals over the candidate pool facilitates the capture of more different haplotypes. A disadvantage of the haplotype-based selection method is that candidates must all be genotyped. In this sense, selection of animals for genotyping or sequencing in populations in which only a small proportion of individuals are genotyped should be done with AMAT, as long as a correct and complete pedigree is available.

Candidate pools for animal selection are often composed of individuals belonging to more than one population due to errors at the time of data recording, and thus crossbred animals could be erroneously selected. Testing methods for their tendency to pick crossbred animals revealed that methods targeting rare variants selected more animals that were not from the population of interest, which was expected. HSH was the only method in which no animal of the JE population was selected before their proportion in the candidate pool reached 5%, which can be considered a usual proportion of crossbred animals wrongly declared as purebreds (**Figure 4**). If GDI or IWS are used on real datasets, population structure analysis and analysis of the relationships between the candidates is essential to ensure that crossbred animals are removed prior to selection.

Following the control of the non-target animals selected with each method, a principal component analysis was used. This allowed for comparison of the distribution and overlap of the selected animals over the complete candidate pool. Methods targeting rare haplotypes picked the same animals more often (**Table 3**). The concentration of points representing the animals selected by IWS and GDI or the superposed dark and light brown points on **Figure 6** follows the same idea. The animals selected for their higher genetic diversity were mostly of generation 1 and 2 out of the 10 simulated generations. Selection applied without allowing for migration in the simulated population led to a reduction of the MAF of the variants under selection pressure (**Figure 7**). Accordingly, the number of combinations of SNP alleles at the haplotype level was reduced, and less unique haplotypes alleles could be found in animals in generation three or more. Fewer unique haplotype alleles also led to higher haplotype frequencies of the remaining ones. Finally, carrying less unique haplotype and haplotypes alleles of higher frequencies, individuals of generation three or more were less likely to be selected by GDI and IWS.

It is of interest to assess the genetic diversity within and between the created reference populations. The proportion of selected haplotypes alleles increased with the number of animals selected, which was expected, as more animals can collectively carry additional different haplotypes. Similarly, when looking at the overlap of picked haplotypes alleles between methods, the methods presented more overlap if they targeted the common (AMAT, HSH) or the rare variants (IWS, GDI). Notably, when reference populations were smaller, animals selected with AMAT carried a greater number of different haplotypes than HSH.

**TABLE 4 |** Accuracies for reference populations of 50, 200, and 1,200 individuals and increasing MAF of the variants considered, all variants, the rare variants (MAF $< 0.05$), or the common variants (MAF $\geq 0.05$).

| | 50 | | | | 200 | | | | 1,200 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MAF bin** | **AMAT** | **IWS** | **HSH** | **GDI** | **AMAT** | **IWS** | **HSH** | **GDI** | **AMAT** | **IWS** | **HSH** | **GDI** |
| 0.00–0.01 | $0.212^{a}$ | $0.282^{b}$ | $0.146^{c}$ | $0.281^{b}$ | $0.471^{a}$ | $0.527^{b}$ | $0.468^{a}$ | $0.540^{b}$ | $0.625^{a}$ | $0.641^{a,b}$ | $0.643^{a,b}$ | $0.647^{b}$ |
| 0.01–0.02 | 0.624 | 0.640 | 0.557 | 0.629 | 0.903 | 0.912 | 0.908 | 0.906 | 0.960 | 0.962 | 0.961 | 0.961 |
| 0.02–0.03 | $0.712^{a}$ | $0.704^{b}$ | $0.678^{a}$ | $0.691^{b}$ | 0.931 | 0.936 | 0.935 | 0.929 | 0.970 | 0.972 | 0.971 | 0.970 |
| 0.03–0.04 | $0.761^{a}$ | $0.732^{b}$ | $0.746^{a}$ | $0.725^{b}$ | $0.944^{a,b}$ | $0.948^{a}$ | $0.948^{a,b}$ | $0.940^{b}$ | 0.975 | 0.976 | 0.975 | 0.975 |
| 0.04–0.05 | $0.793^{a}$ | $0.754^{b}$ | $0.790^{a}$ | $0.745^{b}$ | 0.952 | 0.954 | 0.956 | 0.946 | 0.979 | 0.979 | 0.978 | 0.978 |
| 0.05–0.10 | $0.837^{a}$ | $0.786^{b}$ | $0.848^{a}$ | $0.776^{b}$ | 0.964 | 0.966 | 0.966 | 0.957 | 0.983 | 0.984 | 0.983 | 0.983 |
| 0.10–0.15 | 0.887 | 0.834 | 0.898 | 0.825 | 0.974 | 0.975 | 0.976 | 0.969 | 0.987 | 0.988 | 0.987 | 0.987 |
| 0.15–0.20 | $0.911^{a}$ | $0.865^{b}$ | $0.920^{a}$ | $0.855^{b}$ | 0.979 | 0.980 | 0.980 | 0.974 | 0.990 | 0.990 | 0.989 | 0.989 |
| 0.20–0.25 | $0.925^{a}$ | $0.878^{b}$ | $0.932^{a}$ | $0.870^{b}$ | 0.982 | 0.982 | 0.983 | 0.978 | 0.991 | 0.991 | 0.990 | 0.990 |
| 0.25–0.30 | $0.933^{a}$ | $0.892^{b}$ | $0.940^{a}$ | $0.881^{b}$ | $0.984^{a,b}$ | $0.984^{a}$ | $0.984^{a,b}$ | $0.980^{b}$ | 0.991 | 0.992 | 0.991 | 0.991 |
| 0.30–0.35 | $0.939^{a}$ | $0.904^{b}$ | $0.944^{a}$ | $0.896^{b}$ | $0.985^{a,b}$ | $0.985^{a}$ | $0.985^{a,b}$ | $0.982^{b}$ | 0.992 | 0.993 | 0.992 | 0.992 |
| 0.35–0.40 | $0.942^{a}$ | $0.908^{b}$ | $0.948^{a}$ | $0.900^{b}$ | 0.986 | 0.986 | 0.986 | 0.982 | 0.992 | 0.993 | 0.992 | 0.992 |
| 0.40–0.45 | $0.944^{a}$ | $0.911^{b}$ | $0.949^{a}$ | $0.904^{b}$ | 0.986 | 0.986 | 0.986 | 0.983 | 0.993 | 0.993 | 0.992 | 0.992 |
| 0.45–0.50 | $0.945^{a,b}$ | $0.914^{b,c}$ | $0.949^{a}$ | $0.908^{b,c}$ | $0.986^{a}$ | $0.986^{b}$ | $0.986^{a}$ | $0.983^{b}$ | $0.993^{a}$ | $0.993^{b}$ | $0.992^{b}$ | $0.993^{b}$ |
| All | $0.580^{a}$ | $0.589^{b}$ | $0.549^{c}$ | $0.583^{b}$ | $0.765^{a}$ | $0.789^{b}$ | $0.765^{a}$ | $0.790^{b}$ | $0.840^{a}$ | $0.847^{a,b}$ | $0.847^{a,b}$ | $0.849^{b}$ |
| Common | $0.894^{a}$ | $0.848^{b}$ | $0.903^{c}$ | $0.839^{b}$ | $0.976^{a}$ | $0.976^{b}$ | $0.977^{a}$ | $0.970^{b}$ | $0.988^{a}$ | $0.989^{a,b}$ | $0.988^{b}$ | $0.988^{b}$ |
| Rare | $0.387^{a}$ | $0.429^{b}$ | $0.331^{c}$ | $0.425^{b}$ | $0.635^{a}$ | $0.673^{b}$ | $0.635^{a}$ | $0.679^{b}$ | $0.749^{a}$ | $0.759^{a,b}$ | $0.761^{b}$ | $0.763^{b}$ |

*MAF bins "x-y" stands for "x $<$ MAF $\leq$ y".* $^{a,b,c}$ *Different letters represent significant differences in accuracies among the methods within a bin-by-size set of values (pairwise Wilcoxon Rank Sum Test significant with p-value after experimental-wise Bonferroni correction).*

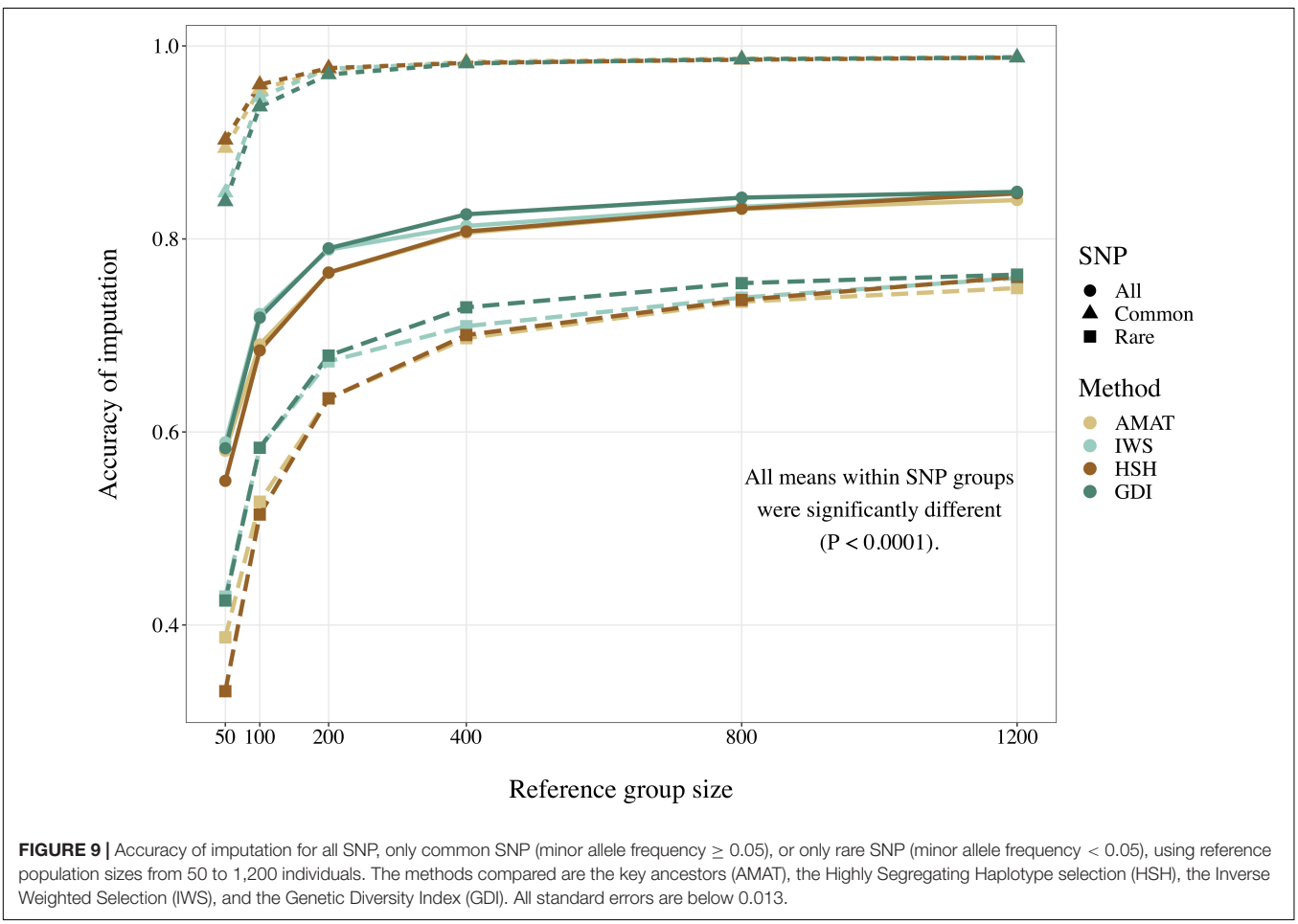


**FIGURE 9 |** Accuracy of imputation for all SNP, only common SNP (minor allele frequency $\geq 0.05$), or only rare SNP (minor allele frequency $< 0.05$), using reference population sizes from 50 to 1,200 individuals. The methods compared are the key ancestors (AMAT), the Highly Segregating Haplotype selection (HSH), the Inverse Weighted Selection (IWS), and the Genetic Diversity Index (GDI). All standard errors are below 0.013.

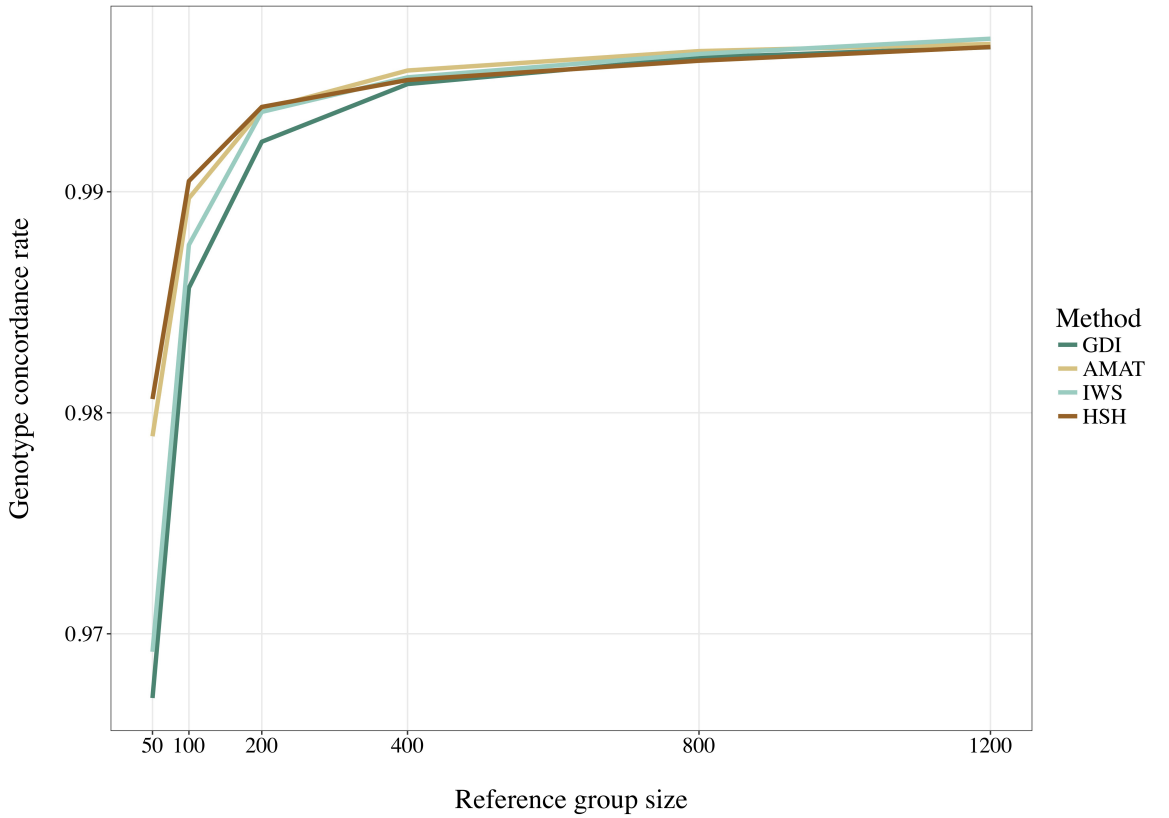


**FIGURE 10 |** Genotype concordance rates for all SNP using reference population sizes from 50 to 1,200 individuals. The methods compared are the key ancestors (AMAT), the Highly Segregating Haplotype selection (HSH), the Inverse Weighted Selection (IWS), and the Genetic Diversity Index (GDI). All standard errors are below 0.009.

This can be explained by the following arguments. The HSH method makes sure that all commonly found haplotypes are selected before animals carrying rare variants get targeted, while AMAT relies solely on pedigree and thus has no possibility to consider the Mendelian sampling happening over generations. This limitation of AMAT, when compared to haplotype-based methods, was observed in our study when 1,200 animals comprised the reference group and only rare variants were considered. In this case, AMAT had a significantly lower accuracy than both HSH and GDI (**Table 4**). Moreover, it is likely that a real pedigree would contain errors that would not allow for a better haplotype coverage using AMAT than HSH as missing and incorrect information would impeach correct computation of the kinship among animals and thus the probable proportion of haplotypes they share. The pedigree-based method AMAT also showed a limitation once the number of animals increased, as redundancy of the added haplotype in the selected group was not directly avoided and effective Mendelian sampling could not be evaluated, which was in contrast to the results obtained for HSH. GDI consistently obtained greater haplotype coverage in the selected group of animals (**Figure 8**). This shows that the targeted optimization at the group level of the number of rare haplotypes was also achieved. Therefore, GDI and IWS seem to be the methods of choice when the objective is to select animals for their propensity to carry novel, rare or deleterious variants. The influence of the selection of genetically more diverse animals on the accuracy of selection, however, must be carefully assessed.

Overall, accuracies of imputation from HD to WGS were similar to previous results observed in real dairy cattle datasets (e.g., Pausch et al., 2017), although rare variants were kept throughout the whole analysis in the current study. Differences between scenarios were significant ($P < 0.0001$), however, the accuracies were mostly similar between methods. This was probably due to very low variance between the replicates, as the simulation algorithm is highly stable. All methods of selection avoided redundancy of the haplotypes selected, thus only minor differences between methods were observed after enough animals were selected. The greatest differences in accuracy of imputation between the methods were found when the reference population was small. Moreover, when observing genotype concordance rates, no differences were found when the reference populations comprised more than 200 animals. In contrary to the allelic $r^2$ and as demonstrated in the review by Calus et al. (2014), the genotype concordance is dependent on the MAF of the variants considered and increases artificially with lower MAF. Differences in the distribution of the MAF of the rare variants between the reference population led to the observed re-ranking. Considering that animals selected with IWS and GDI were mainly from

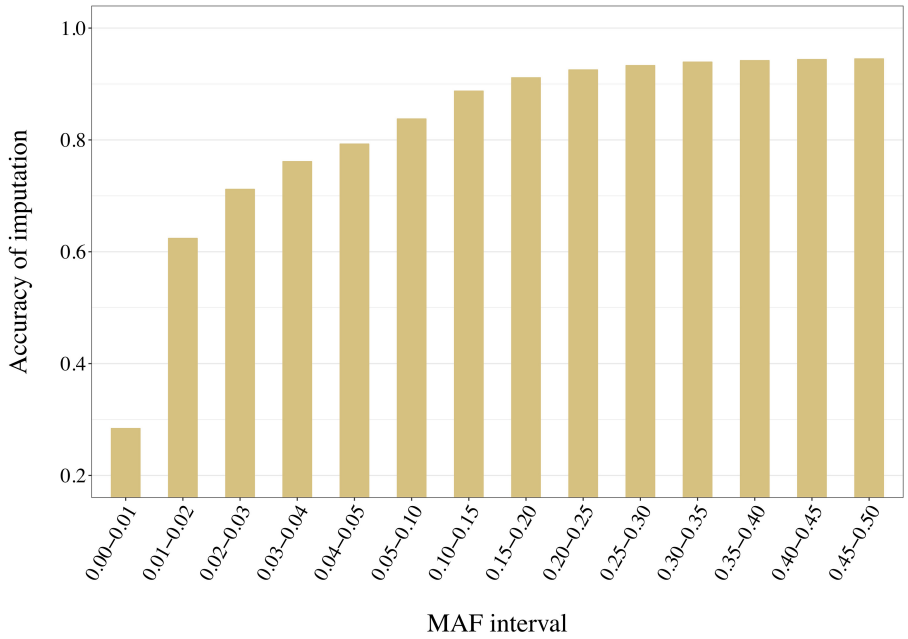


**FIGURE 11 |** Accuracy of imputation increases with higher minor allele frequency (MAF) of the variants. Here the accuracies reached with 50 animals in a reference group of key ancestors (AMAT) are presented. MAF bins "x-y" stands for "x < MAF ≤ y". All standard errors are below 0.008.

generation 1 and 2 of the simulated population (**Figure 6**), and that the MAF distribution of the rare variants shifted toward zero generation after generation (**Figure 7**), the MAF distribution within the rare variants category might be different between reference populations selected for high coverage of rare or common haplotypes. More different haplotype alleles were present in the reference populations selected with GDI and IWS (**Figure 8**), whereas animals selected with AMAT and HSH carried, as intended, more common variants. Animals selected with AMAT and HSH, however, still carried some rare variants but those had more often a MAF below 0.01. **Figure 11** shows a distinctly bigger change in accuracy between monomorphic or rare variants with a MAF lower than 0.01 and rare variants with MAF between 0.01 and 0.05. It is this difference in the distribution of the MAF of the rare variants that explain the re-ranking of the methods between the genotype concordance and the allelic $r^2$ values. Targeting rare haplotypes at selection (GDI and IWS) led to the creation of a reference population with more rare variants, but most of the added rare variants had a MAF between 0.01 and 0.05, whereas targeting common haplotypes led to the creation of a reference population carrying mainly common variants, but also some rare variants that mainly had a MAF below 0.01. Those variants with a MAF below 0.01 artificially increased the genotype concordance so that a re-ranking was observed.

Considering the re-ranking observed between method group, i.e., HSH/AMAT and IWS/GDI when looking at either rare or common variants, the method to select animals should be chosen using one of two principles: if the future imputed genotypes will be used as full genotypes and the imputation needs to be specially accurate for variants that will explain most of the genetic variation of a trait, animals should be selected using AMAT or HSH. In contrast, if future analysis will focus on the discovery of novel functional rare variants animals should be selected using IWS or GDI. Genotype concordance is the measure of imputation accuracy of choice when common variants that explain most of the genetic variance of most traits of interest for the dairy industry, are of interest for future analyses. Our results showed that genotype concordances with small reference populations were higher when the individuals were selected with AMAT or HSH. The first line of **Table 4**, where only the segregating variants with a MAF below 1% were considered, is a good example of the differences in accuracy of imputation for rare variants, variants that could have a novel deleterious effect. In this example, when the reference population only contained 50 animals, the difference in accuracy of imputation reached 0.18 points between the best (IWS) and the worst (HSH) methods. The accuracy of imputation increased with the MAF of the variants, but this increase stopped once segregation reached a level of 30% (**Figure 11**).

## CONCLUSION

Selection of animals for sequencing is an important task, as it greatly impacts the information gained about a population of interest, especially in populations with limited effective population size. Different selection methods are available that either rely solely on pedigree or that utilize information on previously genotyped individuals. In the first case, selecting key ancestors is highly recommended. Otherwise, the best method

depends on the use of the future set of sequences. If the newly selected animals will be the first sequenced animals in their population and should allow for the overall imputation of the rest of the population, it is better to select animals carrying common haplotypes using the new HSH method instead of any of the other methods described in this study. If the resulting sequences of the selection of animals in a population will be used for discovery of new variants or should allow annotation of possible deleterious ones, animals carrying novel information should be selected and, consequently, the GDI method proposed here may be used.

## AUTHOR CONTRIBUTIONS

AB, MS, and CB designed the simulation study and developed the new methods. FM, PS, and FS provided the data. AB, MS, and BG-G performed the data editing and the analysis. AB, MS, PS, and CB interpreted the results. AB wrote the manuscript. All authors read, commented on, and approved the final manuscript.

## FUNDING

We gratefully acknowledge funding by the Efficient Dairy Genome Project, funded by Genome Canada (Ottawa, ON, Canada), Genome Alberta (Calgary, AB, Canada), Ontario Genomics (Toronto, ON, Canada), Alberta Ministry of Agriculture (Edmonton, AB, Canada), Ontario Ministry of Research and Innovation (Toronto, ON, Canada), Ontario Ministry of Agriculture, Food and Rural Affairs (Guelph, ON, Canada), Canadian Dairy Network (Guelph, ON, Canada), GrowSafe Systems (Airdrie, AB, Canada), Alberta Milk (Edmonton, AB, Canada), Victoria Agriculture (Melbourne, VIC, Australia), Scotland's Rural College (Edinburgh, United Kingdom), USDA Agricultural Research Service (Beltsville, MD, United States), Qualitas AG (Zug, Switzerland), Aarhus University (Aarhus, Denmark). We also acknowledge financial support from the National Science and Engineering Research Council of Canada (NSERC).

## ACKNOWLEDGMENTS

The 1,000 Bulls Genome Project is acknowledged for providing the whole-genome sequence genotypes of the Holstein animals. The Canadian Dairy Network provided the array genotypes. Computations were done at the server facilities provided by the Centre for Genetic Improvement of Livestock, Department of Animal Biosciences at the University of Guelph, Guelph, ON, Canada. AB is especially grateful to the Qualitas AG team and the Swiss Association for Animal Science for their support.

## REFERENCES

Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS One* 9:e93766. doi: 10.1371/journal.pone.0093766

Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., et al. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics* 15:948. doi: 10.1186/1471-2164-15-948

Bickhart, D. M., Hutchison, J. L., Null, D. J., VanRaden, P. M., and Cole, J. B. (2016). Reducing animal sequencing redundancy by preferentially selecting animals with low-frequency haplotypes. *J. Dairy Sci.* 99, 5526–5534. doi: 10.3168/jds.2015-10347

Bohmanova, J., Sargolzaei, M., and Schenkel, F. S. (2010). Characteristics of linkage disequilibrium in north american holsteins. *BMC Genomics* 11:421. doi: 10.1186/1471-2164-11-421

Boichard, D. (2002). "Pedig?: a Fortran Package for Pedigree Analysis Suited for Large Populations," in *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production*, Montpellier, 28–29.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., et al. (2012). Design of a bovine low-density snp array optimized for imputation. *PLoS One* 7:e34130. doi: 10.1371/journal.pone.0034130

Boichard, D., Maignel, L., and Verrier, É. (1997). The value of using probabilities of gene origin to measure genetic variability in a population. *Genet. Sel. Evol.* 29, 5–23. doi: 10.1186/1297-9686-29-1-5

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

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, 1743–1753. doi: 10.1017/S1751731114001803

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.* 45, 41–51. doi: 10.1007/BF00940812

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

Drögemüller, C., Rossi, M., Gentile, A., Testoni, S., Jörg, H., Stranzinger, G., et al. (2009). Arachnomelia in brown swiss cattle maps to chromosome 5. *Mamm. Genome* 20, 53–59. doi: 10.1007/s00335-008-9157-2

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

Ely, J. J., Bishop, M. A., Lammey, M. L., Sleeper, M. M., Steiner, J. M., and Lee, D. R. (2010). Use of biomarkers of collagen types I and III fibrosis metabolism to detect cardiovascular and renal disease in chimpanzees (Pan troglodytes). *Comp. Med.* 60, 154–158. doi: 10.1371/journal.pgen.0020190

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., et al. (2016). AlphaSim: software for breeding program simulation. *Plant Genome* 9, 1–14.

Fraser, R. S., Lumsden, J. S., and Lillie, B. N. (2018). Identification of polymorphisms in the bovine collagenous lectins and their association with infectious diseases in cattle. *Immunogenetics* 70, 533–546. doi: 10.1007/s00251-018-1061-7

Fritz, S., Capitan, A., Djari, A., Rodriguez, S. C., Barbat, A., Baur, A., et al. (2013). Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. *PLoS One* 8:e65550. doi: 10.1371/journal.pone.0065550

Goddard, M. E., and Hayes, B. (2009). Genomic selection based on dense genotypes inferred from sparse genotypes. *Proc. Assoc. Advmt. Anim. Breed. Genet.* 18, 26–29.

Gonen, S., Ros-Freixedes, R., Battagin, M., Gorjanc, G., and Hickey, J. M. (2017). An exact method for optimal allocation of sequencing resources in genotyped livestock populations. *Genet. Sel. Evol.* 49:47. doi: 10.1186/1297-9686-43-7

Hayes, B., and Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33:209. doi: 10.1186/1297-9686-33-3-209

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.* 45:33. doi: 10.1186/1297-9686-45-33

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Li, L., Li, Y., Browning, S. R., Browning, B. L., Slater, A. J., Kong, X., et al. (2011). Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* 6:e24945. doi: 10.1371/journal.pone.0024945

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, A. Y., Finucane, K. H., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. doi: 10.1038/ng.3679

Ma, P., Brøndum, R., Zhang, Q., Lund, M., and Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J. Dairy Sci.* 96, 4666–4677. doi: 10.3168/jds.2012-6316

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi: 10.1038/nrg2796

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Neuditschko, M., Raadsma, H. W., Khatkar, M. S., Jonas, E., Steinig, E. J., Flury, C., et al. (2017). Identification of key contributors in complex population structures. *PLoS One* 12:e0177638. doi: 10.1371/journal.pone.0177638

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

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

Pérez-Enciso, M., and Legarra, A. (2016). A combined coalescence gene-dropping tool for evaluating genomic selection in complex scenarios (ms2gs). *J. Anim. Breed. Genet.* 133, 85–91. doi: 10.1111/jbg.12200

Pluzhnikov, A., and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144, 1247–1262.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Available at: https://www.r-project.org/ (accessed March 26, 2019).

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: 10.1101/188896

Sargolzaei, M. (2014). *SNP1101 User's Guide. Version* 1.0. Guelph: HiggsGene Solut. Inc.

Sargolzaei, M., and Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics* 25, 680–681. doi: 10.1093/bioinformatics/btp045

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., and Schaeffer, L. R. (2008). Extent of linkage disequilibrium in holstein cattle in North America. *J. Dairy Sci.* 91, 2106–2117. doi: 10.3168/jds.2007-0553

Utsunomiya, Y. T., Milanesi, M., Utsunomiya, A. T. H., Ajmone-Marsan, P., and Garcia, J. F. (2016). GHap: an R package for genome-wide haplotyping. *Bioinformatics* 32, 2861–2862. doi: 10.1093/bioinformatics/btw356

Whalen, A., Gorjanc, G., Ros-Freixedes, R., and Hickey, J. M. (2018). Assessment of the performance of hidden Markov models for imputation in animal breeding. *Genet. Sel. Evol.* 50:44. doi: 10.1186/s12711-018-0416-8

Zhang, Q., Calus, M. P. L., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2017). Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. *Genet. Sel. Evol.* 49:60. doi: 10.1186/s12711-017-0336-z

Zhang, Q., Guldbrandtsen, B., Calus, M. P. L., Lund, M. S., and Sahana, G. (2016). Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. *Genet. Sel. Evol.* 48:60. doi: 10.1186/s12711-016-0238-5

**Conflict of Interest Statement:** MS was employed by HiggsGene Solutions Inc. and BG-G was employed by Qualitas AG.

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

The reviewer MC declared a shared affiliation, with no collaboration, with one of the authors BG-G to the handling Editor at the time of review.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 June 2019
doi: 10.3389/fgene.2019.00562

# A Random Forests Framework for Modeling Haplotypes as Mosaics of Reference Haplotypes

*Pierre Faux[1]*, Pierre Geurts[2] and Tom Druet[1]*

[1] *Unit of Animal Genomics, GIGA-R, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium,* [2] *Department of Electrical Engineering and Computer Science, Montefiore Institute, University of Liège, Liège, Belgium*

Many genomic data analyses such as phasing, genotype imputation, or local ancestry inference share a common core task: matching pairs of haplotypes at any position along the chromosome, thereby inferring a target haplotype as a succession of pieces from reference haplotypes, commonly called a mosaic of reference haplotypes. For that purpose, these analyses combine information provided by linkage disequilibrium, linkage and/or genealogy through a set of heuristic rules or, most often, by a hidden Markov model. Here, we develop an extremely randomized trees framework to address the issue of local haplotype matching. In our approach, a supervised classifier using extra-trees (a particular type of random forests) learns how to identify the best local matches between haplotypes using a collection of observed examples. For each example, various features related to the different sources of information are observed, such as the length of a segment shared between haplotypes, or estimates of relationships between individuals, gametes, and haplotypes. The random forests framework was fed with 30 relevant features for local haplotype matching. Repeated cross-validations allowed ranking these features in regard to their importance for local haplotype matching. The distance to the edge of a segment shared by both haplotypes being matched was found to be the most important feature. Similarity comparisons between predicted and true whole-genome sequence haplotypes showed that the random forests framework was more efficient than a hidden Markov model in reconstructing a target haplotype as a mosaic of reference haplotypes. To further evaluate its efficiency, the random forests framework was applied to imputation of whole-genome sequence from 50k genotypes and it yielded average reliabilities similar or slightly better than IMPUTE2. Through this exploratory study, we lay the foundations of a new framework to automatically learn local haplotype matching and we show that extra-trees are a promising approach for such purposes. The use of this new technique also reveals some useful lessons on the relevant features for the purpose of haplotype matching. We also discuss potential improvements for routine implementation.

**Keywords: random forests, supervised classification, haplotype mosaic, imputation, extra-trees**

OPEN ACCESS

***Edited by:***
*Marco Milanesi,*
*São Paulo State University, Brazil*

***Reviewed by:***
*Fabyano Fonseca Silva,*
*Universidade Federal de Viçosa, Brazil*
*Filippo Biscarini,*
*Italian National Research Council (CNR), Italy*

****Correspondence:***
*Pierre Faux*
*pierrefaux@gmail.com*

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

***Received:*** *08 October 2018*
***Accepted:*** *29 May 2019*
***Published:*** *27 June 2019*

***Citation:***
*Faux P, Geurts P and Druet T (2019) A Random Forests Framework for Modeling Haplotypes as Mosaics of Reference Haplotypes. Front. Genet. 10:562. doi: 10.3389/fgene.2019.00562*

## INTRODUCTION

Modeling a target haplotype as a succession of segments from other haplotypes (referred to as *reference* or *template* haplotypes) is a common issue and a primary step in various genotype data analyses such as genotype imputation (e.g., in Burdick et al., 2006; Li et al., 2006; Marchini et al., 2007; Howie et al., 2009; Daetwyler et al., 2011; Sargolzaei et al., 2014) often coupled with phase reconstruction, local ancestry inference (e.g., in Price et al., 2009; Baran et al., 2012; Maples et al., 2013), estimation of identity-by-descent between segments (Druet and Farnir, 2011), or even clustering (e.g., in Su et al., 2009; Lawson et al., 2012). To describe this modeling procedure, it is commonly written that target haplotypes are modeled as a mosaic of reference haplotypes (e.g., Burdick et al., 2006; Baran et al., 2012). At any map position along the chromosome, the issue is to find which reference haplotype matches the target haplotype best (**Figure 1A**). Answering this question, for instance in the particular case of genotype imputation, allows to infer the target haplotype on a higher density map, on which the reference haplotypes were observed. Several sources of information are useful to address this question. Many methods (Li et al., 2006; Scheet and Stephens, 2006; Howie et al., 2009; Price et al., 2009) only take into consideration the linkage disequilibrium information. Family information can also be a trustful source, when available at large scale, for instance in livestock (Daetwyler et al., 2011; Sargolzaei et al., 2014). Linkage information (Burdick et al., 2006; Druet and Farnir, 2011; Sargolzaei et al., 2014) is a third potential source of information to locally match haplotypes. Common methods to address this question are usually either based on hidden Markov models (HMM-based methods; see Scheet and Stephens, 2006 for a general model) or rely on a set of deterministic rules (heuristic methods, e.g., based on long-range segments shared between individuals as in Kong et al., 2008).

The development of the latter type of methods, heuristics, could be described as the iterative repetition of two main steps. First, during a conception step, the human operator identifies relevant variables and uses them in a set of rules. Then, during a validation step, the proposed heuristic is tested. If the validation does not return the desired efficiency, then the human operator adjusts the heuristic in the conception step and validates it again. Conception and validation steps would therefore be repeated back and forth until enough efficiency is reached. Defining in these terms the development of a heuristic method for the issue of local haplotype matching makes it an attractive problem for a class of machine learning methods known as *supervised classification*. In such a learning framework, the classifier is fed with data containing both explicative variables (hereafter referred to as *features*, as this denomination prevails in the machine learning community) and their classification (variable to explain, also referred to as *labels*). Then, the data is repeatedly partitioned between a learning sample, on which the classifier performs the conception step, and an independent testing sample, on which the classifier assesses the efficiency of the method. We recommend to readers the review by Libbrecht and Noble (2015) for a detailed glossary as well as clear explanations about the terms used in machine learning.

Additionally, supervised classification also allows combining automatically different sources of information with flexibility. Such aspects make it interesting for locally matching haplotypes: although most of the HMM-based methods (using models similar to Scheet and Stephens, 2006) only rely on haplotype similarity, other methods (e.g., Druet and Georges, 2010) can reach higher efficiency by integrating linkage information. Also, supervised classification returns the importance of any explicative variable as a useful by-product for improving other methods. Because of these advantages, Maples et al. (2013) have already used supervised classification to address a specific problem of local haplotype matching – local ancestry inference. In their approach (RFMix), these authors implemented a random forests (RF) classifier which uses positions along the genetic map as the features.

Here, our main objective is to describe a new learning framework to locally match haplotypes using an extremely randomized trees classifier (*extra-trees*, a particular type of RF method; see Geurts et al., 2006). In this framework, a supervised classifier learns from a large collection of examples what are the relevant features to take into consideration when searching for the reference haplotype that best locally matches a target haplotype and how to combine them. We show that the learning framework accurately finds the best local matches by comparing it to a state-of-the-art HMM-based framework equivalent to IMPUTE2 (Howie et al., 2009). We eventually discuss the main findings of our framework in terms of the importance of features and propose improvements.

## MATERIALS AND METHODS

### Long-Range Haplotype Pre-phasing

All computations and results presented here come from genotypes (for the lower-density map) and WGS (for the higher density map) of the first bovine autosome (BTA1) of 91 dairy cattle from New Zealand (67 bulls and 24 cows; partitioned as 36 Holstein-Friesian, 24 Jersey and 31 crossbred individuals). All individuals have been genotyped with the BovineSNP50k (v1 and v2) genotyping array from Illumina. A total of 2,321 SNPs remained for BTA1 after cleaning the initial data as described in Faux and Druet (2017) and shaped a lower density map, later referred to as the "LD map." Those genotypes were phased using both linkage disequilibrium and family information.

Besides genotyping, all individuals were sequenced at high coverage (15× or more). Details about sequencing and downstream filters can be found in the study by Charlier et al. (2016). A map of 328,045 SNPs from chromosome BTA1 was obtained using stringent filtering rules (described in Faux and Druet, 2017); this map is later referred to as the *higher-density* (HD) map and includes the 2,321 SNPs from LD map. Using stringent rules allowed reducing the proportion of noise in our data set (e.g., assembly errors, false variants, incorrect genotypes, or phasing errors). These stringent filtering rules include, among others: (1) comparisons to other sets of WGS SNPs (markers are kept if they were observed in other

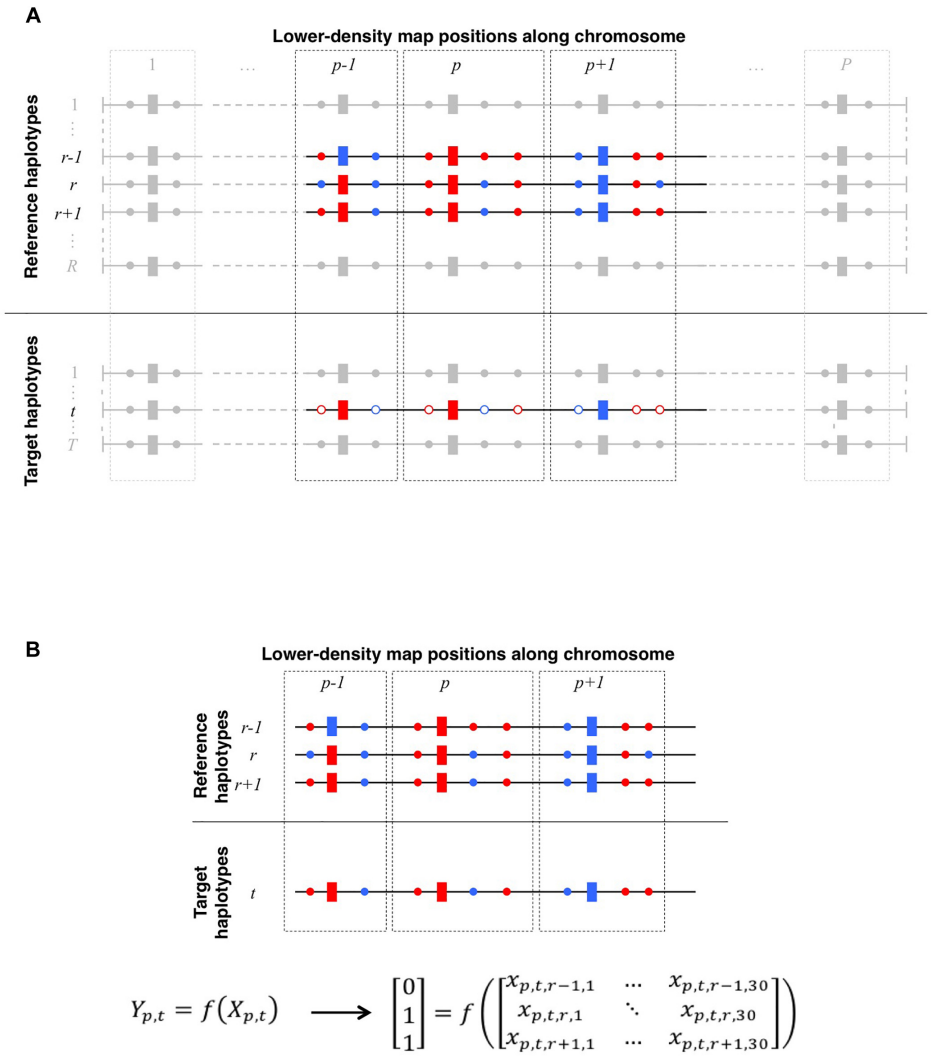


**FIGURE 1 | (A)** Schematic representation of local haplotype matching. Each horizontal line features a whole-chromosome haplotype (phased from red/blue bi-allelic genotypes), to be locally matched (target) to other haplotypes (reference). Both target and reference haplotypes have *P* positions observed on the LD map (rectangles) whereas reference haplotypes may be also observed on a HD map (circles, plain color if observed), thereby allowing imputation of the target haplotype. For a given target haplotype *t*, the question is to find which one of the *R* reference haplotypes matches the best with *t*, in the neighborhood of LD position *p* (delimited by dotted lines). Here, at positions $p - 1$, *p,* and $p + 1$, *t* perfectly matches with *r* and $r + 1$, however, *t* perfectly matches on HD positions only with $r + 1$. Therefore, locally matching haplotypes in such case comes down to match *t* to $r + 1$ rather than to *r*. **(B)** Translating local haplotype matching into machine-readable language. At a LD map position *p*, a target haplotype *t* can be matched to *R* reference haplotypes. Because target haplotypes are also observed on the HD map, we measure the success of each of the *R* local matches by computing the similarity between *t* and each reference haplotypes on HD markers that are closer to the LD position *p* than to any other LD position. Reference haplotypes returning the highest similarity with *t* earn a 1 (*success*) in the observation vector $Y_{p,t}$ whereas others earn a 0 (*fail*). Additionally, we compute a vector $X_{p,t,r}$. of observed features (see **Table 2**) for any reference haplotype *r*. The machine learns how to discriminate *successes* from *fails* in $Y_{p,t}$ according to features in $X_{p,t}$. Here, on HD markers closest to *p*, the target haplotype *t* is identical to reference haplotypes *r* and $r + 1$. This is therefore the maximum similarity observable for haplotype *t* at position *p*. Thus, both reference haplotypes *r* and $r + 1$ earn a *success* ($Y_{p,t,r} = Y_{p,t,r+1} = 1$) whereas any other reference haplotype less similar to *t* (e.g., $r - 1$) earns a *fail* ($Y_{p,t,r-1} = 0$).

available bovine WGS datasets and if they displayed correct Mendelian segregation in another WGS dataset), (2) removal of genomic regions because of a high suspicion of incorrect mapping, and (3) removal of SNPs based on additional rules for error detection.

The HD map was then phased by the two-step method outlined in Faux and Druet (2017). In a few words, this method exploits the haplotypes estimated on a genotyped population much larger (∼58,000 dairy cattle individuals from New Zealand – more details in Faux and Druet, 2017) than the 91 sequenced individuals used in the present study. Therefore, the resulting 182 haplotypes are very accurate: 99.72% of the SNPs whose phasing can be assessed using Mendelian segregation rules were proved to be assigned to their correct parental origin. Based on these results, we consider these haplotypes as the *true* haplotypes in the present study.

## Criteria for Methods Comparison

In this study, we detail a framework for automatic learning of rules to locally match haplotypes and we compare it to an HMM-based method designed for the same purpose. That comparison method is inspired from Howie et al. (2009) and fully described in the section "Hidden Markov Model for Local Haplotype Matching." In order to quantify the ability of each method to accurately achieve this purpose, we partition the full set of 182 haplotypes in reference and target panels. Haplotypes in the target panel are observed only on the LD map whereas those in the reference panel are observed on both LD and HD maps. Any given target haplotype is locally matched to all reference haplotypes on the LD map. Then based on the quality of these local matches, the target haplotype is inferred as a mosaic of the reference haplotypes (which are observed on the HD map).

The first and main criterion to compare methods is, for any target haplotype, the difference between the inferred and the true haplotypes on the HD map, measured by the metric $e_A$ as the proportion of the 328,045 SNPs whose inferred allele is different from the true allele. Such haplotype-based comparison is possible because we consider the phased haplotypes as correct enough to be the true ones. To get rid of the remaining phasing errors in method comparisons, we used a second criterion based on genotypes rather than on haplotypes: imputation reliability ($r^2$), measured, for any SNP specific to the HD map, as the squared correlation between imputed and observed genotypes of all target individuals (see section "Cross-Validation Plan," for partitioning the population in reference and target). Details are given in the next sections on how imputation is performed within the random forests framework and the HMM. We also observed the number of switches from a reference haplotype to another one. Such an observation does not reflect the ability of the methods to reach their objective but provides information on their properties (how many segments from reference haplotypes does the method use when modeling a target haplotype as a mosaic).

## Cross-Validation Plan

The cross-validation plan is outlined in **Figure 2**. In order to obtain numerous cross-validation groups (of uniform size) while keeping a training set of a reasonable size, we have chosen to partition the 91 individuals in thirteen groups of cross-validation (13-fold cross-validation scheme – as detailed in section 7.10.1 of Hastie et al., 2017). In each one of them, fourteen target haplotypes (i.e., those of seven target individuals) are inferred as mosaics of 168 reference haplotypes (i.e., those of 84 reference individuals). Then, the missing genotypes of the seven target individuals are imputed on the HD map. The seven animals forming each batch are randomly picked among the 91 animals. In each of these cross-validation groups, the fourteen target haplotypes are simultaneously imputed and modeled as a mosaic of segments from reference haplotypes. The fourteen imputed haplotypes are then summed pairwise (per individual) to obtain seven imputed genotypes per HD marker. Once cross-validation is achieved over all the 13 groups, there are 182 target haplotypes inferred as mosaic of reference haplotypes and 91 imputed genotypes per HD marker. Comparison criteria $e_A$ and $r^2$ are then measured respectively on all the inferred target haplotypes and on all HD markers for 91 imputed genotypes.

## Machine Learning Framework for Local Haplotype Matching

### General Framework

The purpose of local haplotype matching is to answer the following question (see **Figure 1A**): at a given position $p$ along the chromosome, which of the $R$ reference haplotypes would match at best with a given target haplotype $t$? Answering that question for the $P$ map positions leads to the reconstruction of haplotype $t$ as a mosaic of segments picked from the $R$ reference haplotypes. Hereafter, we detail a framework that makes this question answerable using an automatic classifier.

Let us consider a target haplotype $t$ and a panel of $R$ reference haplotypes. Both are observed on two maps of different densities (LD and HD maps). At a given position $p$, we assume that $t$ could be matched to $R$ haplotypes (see **Figure 1B**); therefore, among the $R$ possible local matches with $t$, we expect at least one to be better than others. To find this one out, we first compute a local difference, denoted $d_{p,t,r}$, for any couple of haplotypes $t$ (target) and $r$ (reference) at position $p$. Considering all the HD positions for which $p$ is the closest position on the LD map, the difference between $r$ and $t$ is computed as the number of these HD positions that carry a different allele between $r$ and $t$. This difference is basically a measure of local similarity between haplotypes. Once all the $R$ differences are obtained, a *success* score (1) award the reference haplotype(s) showing the lowest difference with $t$ whereas other reference haplotypes earn a *fail* score (0), returning thus a $r$-long scoring vector $\mathbf{y}_{p,t}$ whose elements are computed as follows:

$$y_{p,t,r} = \begin{cases} 1, & \text{if } \frac{(d_{p,t,r} - \min(d_{p,t}))}{n_{HD}} \leq 0.01 \\ 0, & \text{otherwise} \end{cases}$$

where $n_{HD}$ is the number of HD positions for which $p$ is the closest LD position. As expressed in the previous formula, more than one reference haplotype may earn a *success* score: obviously all those whose local difference with $t$ is the lowest, but also those whose local difference with $t$ is very close to the lowest local difference (arbitrarily defined as less than 1% of difference in similarity with the best matching haplotype).

The machine learning task is to build up a classifier that discriminates the best reference haplotype from others. For this purpose, we have to feed the classifier with observations on the same features for all the $R$ reference haplotypes. There are many featured observations that may prove to be helpful, e.g., the genetic relationship between haplotype $t$ and any reference haplotypes or the fact that a long identical segment is shared by $t$ and a given reference haplotype on the LD map. Those features can be specific to one map position (as the latter example) or not (as the former one). Measuring these features for all the $R$ reference haplotypes at all the $P$ LD positions shapes a $R$-by-$P$-by-$N$ collection of observations (where $N$ is the number of features). Each observation of the learning sample from which to train the classifier is therefore a vector $\mathbf{x}_{p,t,r}$ of observed features that corresponds to a specific triplet ($p$,$t$,$r$) with $p$ a LD position,

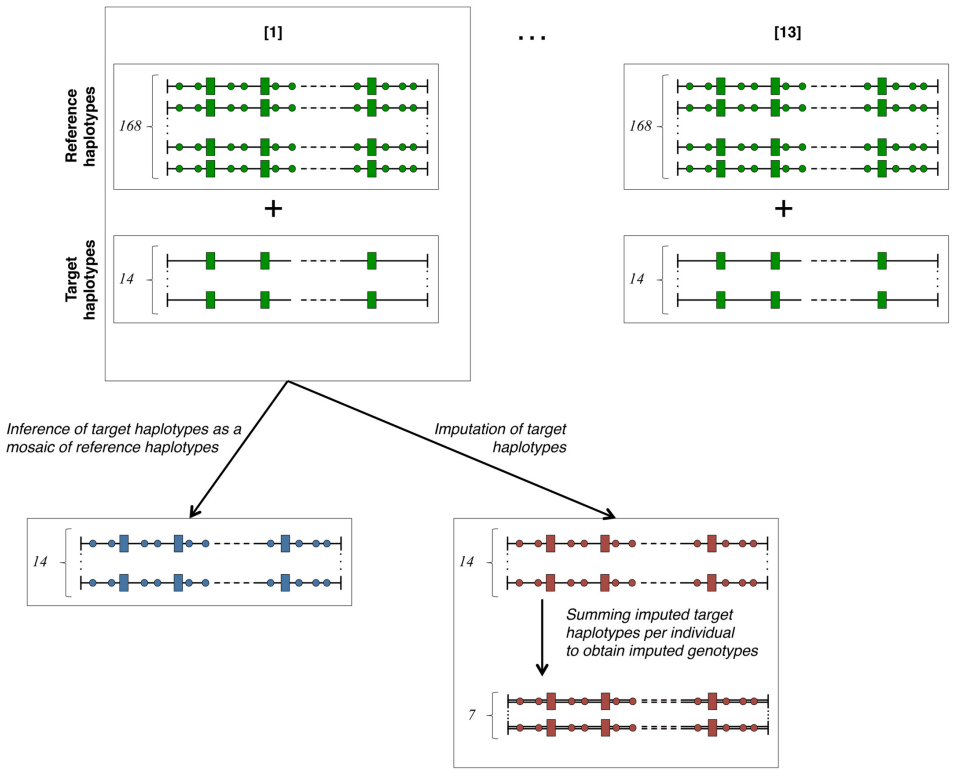


**FIGURE 2 |** Cross-validation plan. The total number (182) of haplotypes in our study is divided into 13 cross-validation groups containing 168 reference haplotypes known on both LD (rectangles) and HD (circles) maps and 14 target haplotypes only known on the LD (rectangles) map. The 14 target haplotypes are each inferred as a mosaic of segments from reference haplotypes and simultaneously imputed from reference haplotypes. Summing per individual the imputed haplotypes returns seven imputed genotypes for each cross-validation group. Green-faced shapes are known (true haplotypes), blue-faced are modeled as mosaic of references and red-faced are imputed.

*t* a target haplotype, and *r* a reference haplotype. The number of observed features defines the length of each vector $\mathbf{x}_{p,t,r}$. Following the terminology of the machine learning community the *success*|*fail* score that corresponds to each observation is hereafter referred to as the *label*. The learning sample thus contains labeled observations, whereas samples with data to predict would contain unlabeled observations (i.e., observed features for each point *p,t,r* but not their score, which remains to predict). The goal of the machine learning algorithm is now to exploit observations in the learning sample and their labels in order to build up a classifier that efficiently discriminates *successes* from *fails*.

## Specific Implementation With Extra-Trees Classifier

The following section details the implementation of the general framework specifically achieved to address the second research objective of this study, namely, to compare the efficiency of the machine learning classifier to locally match haplotypes to an HMM-based method.

Supervised classification is here achieved using the extremely randomized trees method (*extra-trees* hereafter), an ensemble method based on random forests (originally proposed by Geurts et al., 2006). Growing a decision tree works by gathering labeled observations showing identical values of features into a node and then splitting the node if a substantial proportion of these observations have distinct labels (*success* or *fail* in our specific case). The growing process can be illustrated with the theoretical example in **Table 1**: the observations listed in that table are considered as pertaining to the same node of a decision tree. In that theoretical example, we consider two features: the length of a segment shared by target and reference haplotypes (LSS) and the genomic relationship between target and reference gamete on the current chromosome (GENGc). A node split gathering all observations that have a value of LSS greater than 1,000 kb would completely discriminate *successes* from *fails*. The resulting leaves would therefore be "pure": in one leave (LSS $<$ 1,000 kb), all observations are *fails*, in the other one (LSS $>$ 1,000 kb) all observations are *successes*. Such a node split uses only one feature to classify the observations according to their labels and the cut-point value that allowed this split is 1,000 kb. Node splits are determined automatically during tree growing, by going through all features and cut points and looking for the combination that minimizes the label impurity of the leaves defined by this combination. Label impurity reduction is quantified through a score measure, with the most common ones based on Gini index or information entropy (we use the former in our experiments). A complete decision tree is obtained by repeatedly applying these splitting operations on the whole learning sample until the

**TABLE 1 |** Schematic example of a learning sample.

| Features | | | Label |
|---|---|---|---|
| **LSS (in kb)** | **GENGc** | ... | |
| 100 | 0.51 | | Fail |
| 1,500 | −0.02 | | Success |
| 350 | 0.49 | | Fail |
| 400 | 0.36 | | Fail |
| 15,000 | 0.52 | | Success |
| 5,400 | 0.55 | | Success |
| 240 | 0.04 | | Fail |
| 850 | 0.38 | | Fail |
| 350 | 0.44 | | Fail |
| 400 | 0.45 | | Fail |
| 15,000 | 0.44 | | Success |
| 1,500 | 0.56 | | Success |
| 350 | 0.32 | | Fail |

*A target haplotype is compared to a panel of reference haplotypes at any LD map position. Two features (LSS, length of a shared segment; GENGc, genomic relationship between target and reference gamete on the current chromosome) are observed. Each observation can be a success (being the best matching reference haplotype at that position) or a fail, computed using HD map information.*

resulting leaves are either pure (all examples they contain have the same label) or contain too few examples from the learning sample (this threshold is optimized by a parameter – see here below).

A single decision tree usually does not perform well in terms of predictive performance. Better results are obtained by aggregating the predictions, through a majority vote, of an ensemble of decision trees (called forests). Several ways to obtain the different decision trees that compose forests do exist. In Breiman's (2001) original RF algorithm each tree is grown from a bootstrap sample drawn from the original learning sample and node splitting is modified so that the best split (feature and cut point) is searched within a random sample of $k$ features, redrawn at each node. In contrast, in the extra-tree's method, each tree is grown from the original learning sample without bootstrapping. When splitting a node, the best split is searched for among a subset of $k$ randomly selected features like in standard RF, with the difference that the cut-point for each feature is selected randomly instead of being optimized to reduce label impurity as in standard RF. Extra-trees have been shown to be competitive with classical RF in terms of predictive performances while being more computationally efficient because of the extra-randomization (Geurts et al., 2006). For our specific case, they have also proven to yield more accurate results than classical RF (see **Supplementary Material S1**).

In this study, we used the extra-tree classifier implemented as part of the Python SciKit-Learn package (Pedregosa et al., 2011). Among the seventeen parameters of this implementation of the classifier, two were set to a value different than the default one (*n_estimators*, the number of trees, was set to 200 and *min_samples_split*, the minimum number of examples required to split a node, was set to 1) and two were set to vary as they were influencing results more than other parameters during exploratory runs (unpublished results). The first one (*max_features*, the number $k$ of features randomly selected at each node) was set to vary over the range of values [1, 2, 3, 4, 5] and the second one (*min_samples_leaf*, the minimum number of examples required at a leaf node) was set to vary over the range of values [50, 150, 250, 500, 1000, 1500, 2000, 2500].

After the learning stage, extra-trees return the importance of each feature, which is a measure of the total reduction of impurity brought by that feature within the forest. The higher the importance of a given feature in the forest, the more relevant this feature is in predicting the label. Therefore, importance values can be used afterward to rank the features from the most to the least relevant and to gain some understanding of the problem.

## Optimization of Extra-Trees Parameters

To tune these parameters, we used a second internal cross-validation loop. More precisely, each of the 13 groups of the external cross-validation loop (outlined in **Figure 2**) is further divided into 12 subgroups. Each of these 12 subgroups are divided into target and reference panels in the same way as for the 13 groups of the outer loop (see **Figure 2**). For each of the 5-by-8 combinations of the *max_features* and *min_samples_leaf* parameters and for each of the 12 subgroups, all target haplotypes are modeled as a mosaic of reference haplotypes and imputed, and the comparison criteria $e_A$ and $r^2$ are computed. For each criterion, the combination of parameters yielding the best values over all twelve subgroups is retained as the optimal one, returning therefore the two best combinations (one per criterion) used for the parent cross-validation group. Such two-level cross-validation is necessary to avoid artificial inflation of results that might arise if we would have used the target panel from the cross-validation group in the optimization of parameters.

## Building the Learning Samples

The learning sample of each of the 13 cross-validation groups is built by successively considering each one of the 84 reference individuals as a target. Therefore, two haplotypes considered as targets are matched to 166 haplotypes considered as references along the 2,321 positions of our LD map. The maximal number of labeled observations in the learning sample of the cross-validation group is thus close to 65 million ($2{,}321 \times 2 \times 166 \times 84$). Handling such a large learning sample would be tricky computationally speaking. Furthermore, we expect much of it to be redundant, which is the reason why we have downsized the number of labeled observations to two fixed sizes of 100,000 and 1,000,000, randomly picked from the 65 million possibilities and, respectively, denoted as EXT-100k and EXT-1M hereafter.

## Selection of Features

Features from which observations are made were selected during exploratory analyses (unpublished results) and are listed in **Table 2**. We have listed 30 of them and ordered them in three main types : (1) those gathering information about local similarity between haplotypes, (2) those estimating the relationships between individuals, gametes, and haplotypes, and (3) those outputted from other methods for locally matching haplotypes.

**TABLE 2 |** List of all features investigated for use in the random forests framework, with their names and ranges of variation.

| Type | Name | Description | Range | |
|---|---|---|---|---|
| | | | Min | Max |
| Features based on position along the chromosome and local haplotype sharing (16 features) | POS | Position along the SNPs of the LD panel | 1 | *P* |
| | NSS | Length (in #POS) of the shared segments | 0 | *P* |
| | R1-NSS | Ranking (standard*) of the length (in #POS) of the shared segment | 1 | *R* |
| | R2-NSS | Ranking (dense*) of the length (in #POS) of the shared segment | 1 | *R* |
| | DLN | Distance (in #POS) to the left edge of the shared segment + 1 | 0 | *P* + 1 |
| | DRN | Distance (in #POS) to the right edge of the shared segment + 1 | 0 | *P* + 1 |
| | DMN | Distance (in #POS) to the closest edge of the shared segment + 1 | 0 | *P* + 1 |
| | R1-LSS | Ranking (standard*) of the physical length of the shared segment | 1 | *R* |
| | R2-LSS | Ranking (dense*) of the physical length of the shared segment | 1 | *R* |
| | iDLN | Inverse of DLN, as 2-(DLN)$^{-1}$ when DLN > 0; 0 otherwise | 0 | 2 |
| | iDRN | Inverse of DRN, as 2-(DRN)$^{-1}$ when DRN > 0; 0 otherwise | 0 | 2 |
| | iDMN | Inverse of DMN, as 2-(DMN)$^{-1}$ when DMN > 0; 0 otherwise | 0 | 2 |
| | LSS | Physical length of the shared segments (in kb) | 0 | *L* |
| | DLL | Physical distance to the left edge of the shared segment | 0 | *L* |
| | DRL | Physical distance to the right edge of the shared segment | 0 | *L* |
| | DML | Physical distance to the closest edge of the shared segment | 0 | *L* |
| Features based on estimation of relationship (11 features) | PEDI | Pedigree relationship between reference and target individuals | 0 | 2 |
| | PEDG | Pedigree relationship between reference and target gametes | 0 | 1 |
| | GENI | Genomic relationship (as in Yang et al., 2010) between reference and target individuals on all chromosomes | (n.b.) | |
| | GENG | Genomic relationship (as in Yang et al., 2010) between reference and target gametes on all chromosomes | (n.b.) | |
| | GENIc | Genomic relationship (as in Yang et al., 2010) between reference and target individuals on the current chromosome | (n.b.) | |
| | GENGc | Genomic relationship (as in Yang et al., 2010) between reference and target gametes on the current chromosome | (n.b.) | |
| | SIMI | Genomic similarity between reference and target individuals on all chromosomes | 0 | 1 |
| | SIMG | Genomic similarity between reference and target gametes on all chromosomes | 0 | 1 |
| | SIMIc | Genomic similarity between reference and target individuals on the current chromosome | 0 | 1 |
| | SIMGc | Genomic similarity between reference and target gametes on the current chromosome | 0 | 1 |
| | MNT | Minimum number of ties to join the reference and target gametes using the pedigree (equal to 100 when MNT > 99) | 1 | 100 |
| Features outputted from other methods (3 features) | PBLM | Probability of IBD obtained by the HMM-HP-LD method | 0 | 1 |
| | R2-PBLM | Ranking (dense*) of reference haplotypes according to their PBLM | 1 | *R* |
| | MASW | Moving average of the number of switches between longest shared segments in the surrounding 5 Mb | 0 | (n.b.) |

**Standard ranking is "1134" whereas dense ranking is "1123." The dense ranking allows comparing a situation where many reference haplotypes are the local best match to a situation where only one is the local best match: in both cases the second top-ranked reference has a ranking equal to 2. nb: not bounded.*

Features of the first type contain information about local similarity between target and reference haplotypes, according to their position along the phased chromosome. The LD position itself is one of these features, as well as a group of features related to the size of the segment shared between reference and target haplotypes (expressed in number of SNPs, in kb, or ranked) and a group of features related to the position inside a shared segment, expressed as the distance to the edges of the segment. If target and reference haplotypes do not share a segment at a given position, only the LD position is non-zero; as no identity was observed, there are no shared segments and therefore their length and distance to their edges are set to zero.

Then come features related to (individual, gametic, haplotypic) relationships. Note that we understand the term "gamete" to mean the whole set of alleles inherited from each parent, as mentioned in previous studies involving gametic relationships (e.g., Schaeffer et al., 1989). Estimations are based on pedigree information and/or genomic information brought by the LD map. In the present study, haplotypes from individuals with ancestors in the sample are identified according to their

parental origins (e.g., paternal vs. maternal haplotype). This allows the use of gametic relationships (e.g., based on the genealogy, the paternal haplotype is linked with both haplotypes from its father and eventually to haplotypes from paternal grandparents, when these are present in the sample, but it is not linked to the haplotypes from its mother, assuming both parents are unrelated). Following notations in **Figure 3**, PEDI and PEDG are the additive relationships [estimated using pedigree information as defined in Wright (1922)], respectively, between individuals (e.g., $I_i$ and $I_j$) and gametes (e.g., $G_{i,p}$ and $G_{j,p}$, or $G_{i,p}$ and $G_{j,m}$). Genomic relationships (between individuals, gametes – on all autosomes – or haplotypes – only on current autosome and denoted with suffix "c") are computed using the formula by Yang et al. (2010). That formula weights the relationship according to allelic frequencies. Conversely, the genomic similarities (between the same pairs of individuals, gametes, and haplotypes as for genomic relationships) do not take into account allelic frequencies (computed using Eq. 6 in Speed and Balding, 2014). Considering the pedigree as a directed graph, we have computed the feature MNT (for the *minimum number of ties*) as the shortest path from any gamete to another one.

Lastly come features outputted from other methods for locally matching haplotypes: (1) the probability that any reference haplotype would be the best local match haplotype for a given target haplotype (PBLM), as computed in our implementation of the HMM and ranked from highest to lowest (R1-PBLM), and (2) the average number of switches in the 5 Mb surrounding the current position (MASW), using a simple (unpublished) heuristic that reconstructs the target haplotype as a mosaic of segments from reference haplotypes under constraint of a minimal number of segments. Here, the rationale is that a high value of MASW could pinpoint a chromosomal region where no large reference haplotype could be assigned to the target haplotype. Through PBLM, the classifier is fed the data used by the HMM-HP-LD modality of our HMM (see the description here below, section "Modeling Target Haplotypes As a Mosaic of Reference Haplotypes") without, however, specifying its selection rule (namely, the reference haplotype with the highest probability is chosen).

## Tests With Reduced Number of Features

In order to better understand properties of the machine learning classifier, we have applied a similar evaluation protocol to four modalities corresponding to four relevant sets of features. Each of them was obtained from the learning samples used in the EXT-100k modality by hiding some features. EXT-100k-L contains all

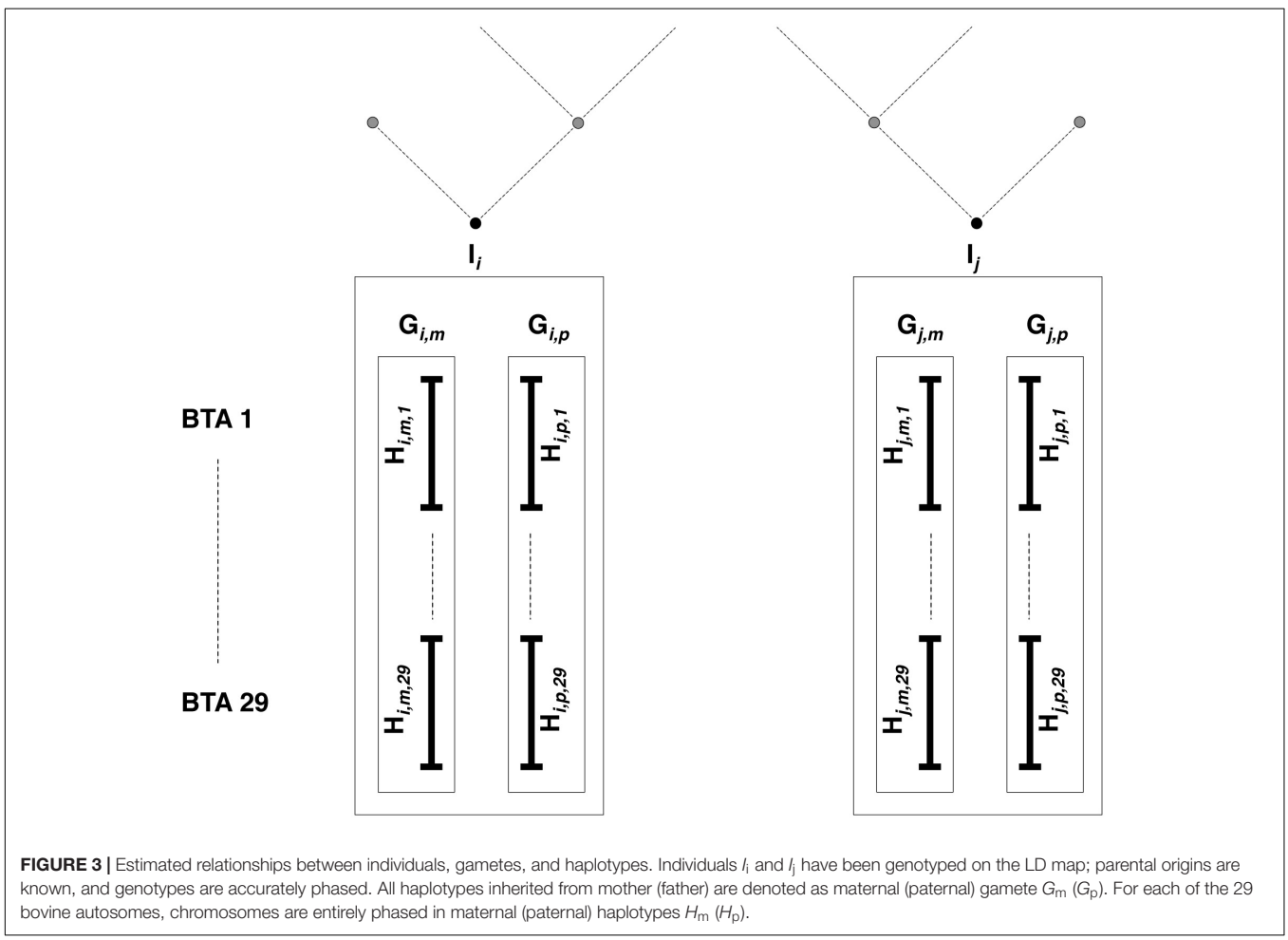


**FIGURE 3 |** Estimated relationships between individuals, gametes, and haplotypes. Individuals $I_i$ and $I_j$ have been genotyped on the LD map; parental origins are known, and genotypes are accurately phased. All haplotypes inherited from mother (father) are denoted as maternal (paternal) gamete $G_m$ ($G_p$). For each of the 29 bovine autosomes, chromosomes are entirely phased in maternal (paternal) haplotypes $H_m$ ($H_p$).

features from the first type (cf. **Table 2**), EXT-100k-LR contains all features from the first and second types, EXT-100k-H only contains the two features obtained from the HMM (PBLM and R1-PBLM) and the last one, EXT-100k-HR contains the two HMM features plus all features from the second type. In this case, the cross-validation plans, the comparison criteria and the learning samples are the same. The only difference lies in the range of tested values for optimization of the *max_features* parameter ([1, 2] instead of [1, 2, 3, 4, 5] to not exceed the number of features of the group with the lowest number of features).

### Obtaining Evaluation Criteria

Once extra-trees have learnt discrimination rules using the learning sample, the rules are applied to unlabeled observations and, for any of them, the extra-tree classifier provides the probability that the observation belongs to the two score modalities: $P_s$, the probability of *success*, complement $P_f$, the probability of *fail*. For any target haplotype at any LD position, $P_s$ are computed for each reference haplotype. The best match is the one that has obtained the highest (predicted) probability of success (in case of equality, the reference haplotype occurring at first in the vector of probability is chosen). Doing so for each LD position results in modeling the target haplotype as a mosaic of segments from the locally best matching reference haplotypes. The main criterion to assess the correctness of the mosaic target haplotype, the metric $e_A$, is obtained by summing the difference of allelic content between a true target haplotype observed on the HD map and its modeling as a mosaic of HD segments from the reference haplotypes.

A first imputation of the target haplotypes (only observed on the LD map) may be achieved by considering the inferred mosaic of reference haplotypes (observed on both maps) on the HD map. However, haplotype imputation may yield better results if we consider more reference haplotypes rather than only the best matching one, e.g., if there are more than one best matching haplotype, or if some reference haplotypes have a $P_s$ very close to the highest one. Therefore, we impute the allelic content $a_i^t (a_i \in [0, 1])$ of a target haplotype $t$ at SNP $i$ by averaging over the allelic contents of all $Q$ best-matching reference haplotypes among $R (Q \leq R)$ according to a weight $w_q$ as follows:

$$a_i^t = \sum_{q=1}^{Q} (w_q \cdot a_i^q)$$

The weight $w_q$ is computed according to the probabilities of the best local match ($P_s$) of the $Q$ best-matching reference haplotypes at the LD position closest to HD position $i$:

$$w_q = \frac{P_{s(q)}}{\sum_{q=1}^{Q} P_{s(q)}}$$

The $Q$ best-matching reference haplotypes are selected as those having a $P_s$ greater or equal to a fraction $c (c \in [0, 1])$ of the highest $P_s$. For instance, setting $c$ to 0 leads to a weighted average of all the $R$ reference haplotypes. Nonetheless, such an option is not optimal: the best imputation results were obtained during exploratory runs with $c$ close to 1.

For a given individual, the imputed HD dosages are obtained by summing the allelic contents of the two imputed haplotypes. Once genotype imputation is achieved for all animals, the imputation reliability ($r^2$) can be computed at every HD map position. Note that the optimization of extra-tree parameters *max_features* and *min_samples_leaf* are independently achieved for each criterion chosen for comparison; optimized parameters, and thus optimized extra-trees, are different, whether the purpose was to optimize $e_A$ or the imputation of $r^2$. For imputation purposes, the value of $c$ is optimized along with *max_features* and *min_samples_leaf* by setting it to vary in the range [0.75, 0.80, 0.85, 0.90, 0.95, 1.00].

## Hidden Markov Model for Local Haplotype Matching

### Modeling Target Haplotypes as a Mosaic of Reference Haplotypes

IMPUTE2 (Howie et al., 2009) returns imputed genotypes without providing information on the best matching reference haplotypes. To obtain the mosaic structure, we have implemented an HMM equivalent to IMPUTE2 and similar to models underlying other HMM-based methods, e.g., MaCH (phasing and imputation, Li et al., 2006) or ChromoPainter (local ancestry inference, Lawson et al., 2012). Our model corresponds to settings where genotypes are pre-phased, thus it does not include a phasing step, nor does it integrate phasing uncertainties. Working straight from phased haplotypes rather than genotypes makes the method comparable to the random forests framework.

In this HMM, we model each target haplotype as an unobserved mosaic of the $R$ reference haplotypes (hidden states). Emission probabilities $P_e$ correspond to the probability to observe allele $k$ ($k = 0|1$) at a position $p$ when the underlying hidden state is a reference haplotype $r$ and accounts for genotyping errors. Denoting the probability of error as $P_{error}$, $P_e$ is equal to $1 - P_{error}$ if alleles are identical and to $P_{error}$ if alleles are not identical. Between positions $p$ and $p + 1$, separated by a distance $d_{p,p+1}$ (in cM), the probability of transition $P_{t;p,p+1}$ from hidden state $r$ to hidden state $s (r, s \in [1, R])$ is estimated as:

$$P_{t;p,p+1} =$$

$$\begin{cases} 1/R \cdot \left(1 - exp\left(-N_g d_{p,p+1}\right)\right) & \text{if } r \neq s \\ exp\left(-N_g d_{p,p+1}\right) + 1/R \cdot \left(1 - exp\left(-N_g d_{p,p+1}\right)\right) & \text{if } r = s \end{cases}$$

In the formula above, $N_g$ is a parameter corresponding to the expected number of generations from the target haplotype to the reference haplotype. Since the maximum number of reference haplotypes is low in our case ($R = 168$ at maximum, see **Figure 2**), we do not restrict the space of hidden states.

At each position, we compute the probability that the reference haplotype $r$ contributes to the unobserved mosaic structure of target haplotype $t$ according to the HMM. That probability is later referred to as the "best local match probability"

(for consistency with definition used for the random forests framework) and is computed with the forward–backward algorithm (described in Rabiner, 1989). This algorithm efficiently computes the probabilities over all possible sequences of unobserved states and conditionally on all observations and on the parameters of the model.

Inferring a discrete mosaic sequence is achieved in two ways: (1) HMM-VI, selecting the most likely mosaic sequence using the Viterbi algorithm (also described in Rabiner, 1989), or (2) HMM-HP, selecting the hidden state (reference haplotype) with highest probability at each map position. The HMM is trained on the two genetic maps, LD and HD, leading therefore to four mosaic sequences (HMM-VI-LD, HMM-VI-HD, HMM-HP-LD, HMM-HP-HD).

The parameters $P_{error}$ and $N_g$ of the so-defined HMM have been chosen to mimic at best the behavior of IMPUTE2 with option *allow_large_regions* and default parameters except for *k_hap* (set to 168) and *Ne* (set to 200). The selected values are $P_{error} = 0.0005$ and $N_g = 4.7619$. The model was then applied to all 14 target haplotypes of each of the 13 cross-validation groups (see **Figure 2**).

### Imputation of Target Haplotypes and Genotypes Using the HMM

For any map position, haplotype imputation of a given target haplotype is obtained by averaging the allelic content of all reference haplotypes according to their respective best local match probability (computed using forward–backward algorithm). When the HMM is trained on the LD map, HD positions that are unobserved on that map are imputed using probabilities computed at the closest LD positions. Imputed haplotypes are eventually paired per individual to yield imputed dosages. With the aforementioned values for parameters $P_{error}$ and $N_g$ and trained on the HD map, our implementation of the model behaves similarly enough to IMPUTE2 (using option *allow_large_regions* and the fore-mentioned values for parameters *k_hap* and *Ne*) to consider them as identical imputation methods (see correlations between imputation methods in **Supplementary Material S2**). Hereafter, genotype imputation results using the HD map are obtained by running IMPUTE2 (with fore-mentioned parameters) and results using the LD map are obtained by running our implementation of the HMM (denoted HMM-LD and written in Fortran 90).

## RESULTS

### Importance of Features

After supervised learning on the learning samples of the 13 cross-validation groups (see **Figure 2**), the importance of each of the 30 features was computed and averaged over the 13 cross-validation groups. The features are ranked by importance in **Figure 4**, for each case of size of learning sample and each purpose (inference of a target haplotype as a mosaic of reference haplotypes and genotype imputation from LD to HD map). The ranking is quite conserved between the four cases: from 96.9 to 99.7% of Spearman's correlation, less correlated between purposes than between sizes of LS. The three top-ranked features are always iDMN, iDRN, and iDLN, three features expressing the distance to the edge of a shared segment (respectively the minimal, right and left distances) on an inverse scale. These three features mostly form a top group, well delimited from other features. It may be worth noting that those three

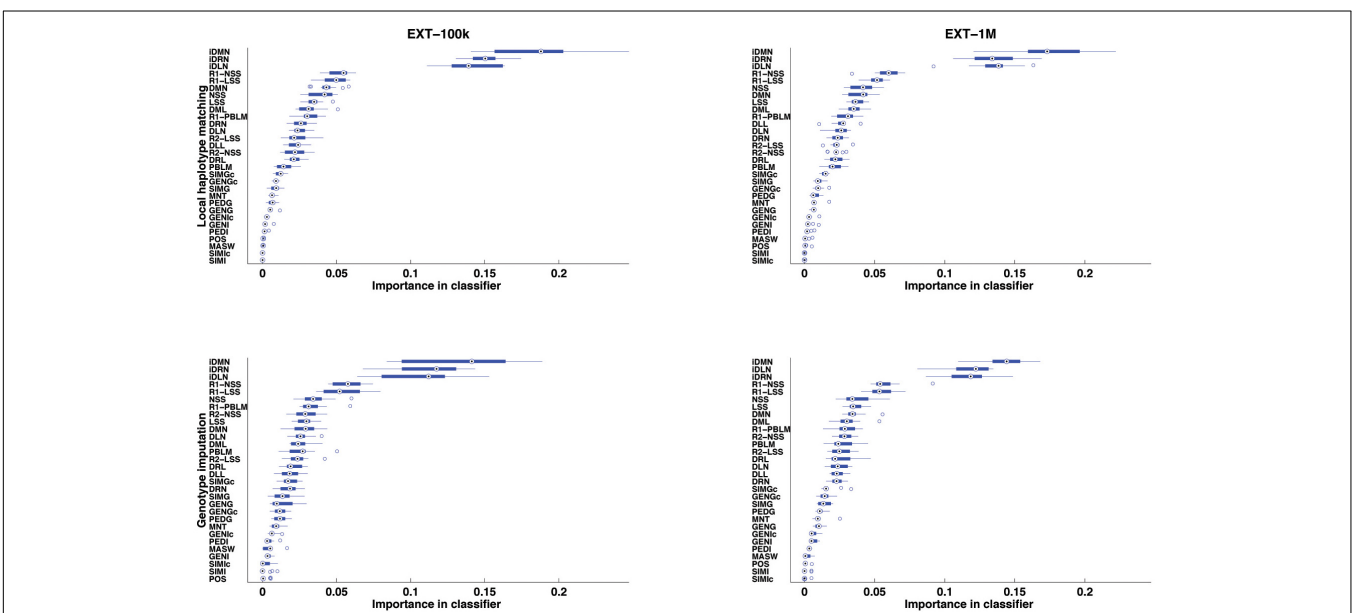


**FIGURE 4 |** Features ranked by their importance in extra-trees (averaged over 13 cross-validations), for the purpose of locally matching haplotypes (top) or genotype imputation (bottom) and for two sizes of learning sample (100,000 and 1,000,000 labeled observations).

features are always preferred to their corresponding ones on the regular scale (DMN, DRN, and DLN). Those are ranked in a second group of importance, alongside features related to the size of shared segments (NSS, LSS and their rankings). Features related to estimation of relationships (between gametes or individuals) are always low in rankings: SIMGc earns the highest ranking (17th) for a feature of this kind, ∼22 times less important than iDMN in that ranking. About features related to other assignation methods, the ranking of the best local match probability (R1-PBLM) is always more important than the probability itself (PBLM). The estimated number of switches in the neighboring 5 Mb (MASW) is consistently the least important feature, in the bottom group along with similarity between individuals.

The distribution of four selected features (iDMN, DMN, NSS, and GENGc) are given in **Figure 5** (the detailed information is given in **Supplementary Material S3**). In that figure, the range of each of these features is divided in 20 equally spaced bins. The relative size of each bin is then computed as the proportion of observations falling into this bin. Among those observations, some are labeled with *success* (in blue), others with *fail* (in red). The purity of the bin is measured by the proportion of objects in this bin and labeled with *success*. This figure therefore shows how each of these four features is linked to the label. For each of them, the lower the value of the feature, the lower the purity and the larger the bins. However, feature iDMN reaches a better compromise between purity and size than feature GENGc does, for instance: less than 1% of the observations fall in the last bin of GENGc, in which 99.9% of the observations are successes, whereas 5.5% of the observations fall in the last bin of iDMN, in which purity is reasonably high (94.5% of the observations are successes). This may explain why iDMN is a good feature for classification.

## Differences Between True Haplotypes and Haplotypes Predicted Using Extra-Trees or the HMM

The 182 target haplotypes were modeled (per group of 14, see the cross-validation plan in **Figure 2**) as mosaics of HD segments from the best matching reference haplotypes. The metric $e_A$ was then measured by comparing the modeled haplotypes to their known phase, for the four modalities of the HMM and the two modalities of the random forests framework. Results are averaged over the 182 haplotypes in **Table 3**. On these results, we see that the extra-trees classifier performs better than the other methods, whether the learning sample contains 1E5 or 1E6 objects. When a target haplotype is inferred as a mosaic of HD segments from the reference haplotypes that are locally classified as the best match, 98.75–98.77% of the HD positions have allelic content identical to the known target haplotype on the HD map. The HMM-HP-xx returns a lower median value than the extra-trees classifier; that median value difference is, however, much lower than the average difference.

Among the four HMM mosaic sequences, the method for selection of the local reference haplotype has more impact than that of the map on which the HMM was trained. Building the mosaic by selecting the hidden states (reference haplotypes) with the highest best local match probability (HMM-HP-xx) performs better on both maps

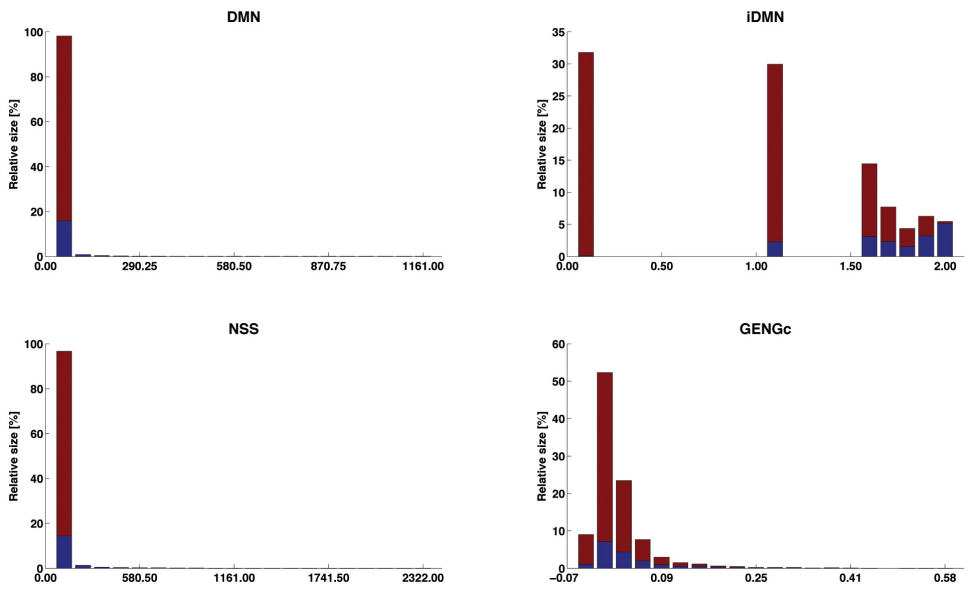


**FIGURE 5 |** Distribution of the *success* labels along the ranges of four selected features. The range of each feature is divided into 20 equally spaced bins; the relative size of each bin (in %) is given by its height and its proportion of observations labeled with success is blue faced. The four features are DMN [distance (in #POS) to the closest edge of the shared segment +1], iDMN [inverse of DMN, as 2-(DMN)$^{-1}$ when DMN > 0; 0 otherwise], NSS [length (in #POS) of the shared segments], and GENGc (genomic relationship between reference and target gametes, on the current chromosome).

**TABLE 3 |** Inference of target haplotype as a mosaic of reference haplotypes.

| | $e_A$ [%] | | | | Number of switches in inferred mosaic | | | |
|---|---|---|---|---|---|---|---|---|
| | **Min** | **Avg** | **Med** | **Max** | **Min** | **Avg** | **Med** | **Max** |
| HMM-VI-LD | **0.004** | 1.441 | 0.430 | 11.936 | **0** | 15.7 | **9.0** | 73 |
| HMM-HP-LD | 0.005 | 1.304 | 0.413 | 7.401 | **0** | 19.5 | **9.0** | 91 |
| HMM-VI-HD | 0.005 | 1.413 | 0.409 | 8.327 | **0** | **14.9** | **9.0** | **67** |
| HMM-HP-HD | 0.005 | 1.310 | **0.394** | 7.403 | **0** | 27.6 | **9.0** | 671 |
| EXT-100k | 0.005 | **1.226** | 0.410 | **6.941** | 4 | 70.5 | 47.0 | 285 |
| EXT-1M | 0.006 | 1.231 | 0.414 | 7.026 | 4 | 95.8 | 71.0 | 367 |

*Distribution of the difference between predicted and true haplotypes ($e_A$) and of the number of switches in the mosaic, on 182 haplotypes and 328,045 HD SNPs. Best results are boldfaced.*

than by selecting the best mosaic sequence with the Viterbi algorithm (HMM-VI-xx).

Methods are ranked almost reversely when looking at the number of switches in the mosaic in **Table 3**: the best mosaic sequences on $e_A$ tend to model the target haplotype with more segments. For instance, when using the HMM, the mosaic obtained by the Viterbi algorithm (HMM-VI-xx) is less prone to switches than the mosaic obtained by selecting the reference haplotype with highest best local match probability (HMM-HP-xx), whatever the map (VI does 19 and 46% less switches than HP, respectively, for LD and HD maps). Conversely, the HP mosaic sequences have a lower proportion of error than the VI mosaic sequences (e.g., the average $e_A$ is equal to 1.41% for HMM-VI-HD and 1.31% for HMM-HP-HD).

## Comparisons of Imputation Reliability Between Extra-Trees and HMM

In **Table 4**, results of imputation from LD to HD maps are detailed for the four methods of imputation: HMM using LD and HD maps (respectively HMM-LD and IMPUTE2) and extra-trees with 100,000 and 1,000,000 observations in the learning samples (respectively EXT-100k and EXT-1M). The imputation $r^2$ are categorized by minor allele frequency (MAF) and position along the BTA1 chromosome. These results show that the extra-trees classifier performs as good as HMM: extra-trees classifiers are better on average imputation $r^2$ whilst IMPUTE2 has a greater number of variants that are better imputed (higher median). Although slightly better on rare variants (MAF $< 0.05$) and between first and last Mb of the chromosome, the machine learning model is distinctly better than the HMM on chromosome edges: SNPs located on the last Mb of BTA1 have an average imputation $r^2$ 2.23% higher for the best extra-trees (EXT-100k) than for the best HMM (IMPUTE2).

The statistics in **Table 4** relate to the SNPs that do not pertain to the LD map and for which imputation reliability was always computable (for that reason, SNPs imputed as monomorphic by one of the four methods were excluded). The numbers of SNP excluded for being imputed as monomorphic are proportionally very low (0.14% of the total number of only HD SNPs) but the random forests framework has imputed SNPs as monomorphic ∼3 to ∼4 times more than the HMM.

Another way of categorizing SNPs to highlight imputation differences between methods is given in **Figure 6**. That figure shows the average imputation $r^2$ in regard to the distance between the imputed HD SNP and the closest observed LD SNP. Ten classes of distance (from 0–2.9 to 66–389 kb) were designed so that they all include the same number (∼33k) of HD SNPs. For the HMM-based imputations, the figure shows that both maps return an equal average reliability up to ∼13 kb and then the HD map (IMPUTE2) overtakes the LD map (HMM-LD). Besides, whatever the size of the learning sample (EXT-100k or EXT-1M), the random forests framework always imputes better than the HMM which uses the same map (HMM-LD). As a result of these two trends, the random forests framework always yields better results than the HMM, except for the most distant class (>66 kb), where IMPUTE2 overtakes it. However, in that last distance class, the average imputation $r^2$ drops for all methods.

### Machine Learning With Reduced Number of Features

The results (**Table 5**) obtained when considering only the features of the first type (i.e., those based on the position along the chromosome) are quite close to the results obtained with all features, much more for inferring the target haplotype as a mosaic of segments than for genotype imputation. Adding the eleven relationship features further enhances these results. Note that the differences between **Tables 3**, **4** on average imputation $r^2$ for a given method are due to the exclusion of more SNPs in **Table 5**, for being imputed as monomorphic in at least one of the tests.

Though lower, the results achieved by an automatic classifier only fed with two features – the features returned by the HMM (the probability of best local match and its ranking) – are still close to the "full" automatic classifier and actually slightly better than HMM-HP-HD for the purpose of inferring the target haplotype as a mosaic of segments. For that purpose, using the two HMM features with machine learning returns the same results as the HMM using the LD map (HMM-HP-LD). Surprisingly however, adding the relationship features yields worse results. The fact that the

**TABLE 4 |** Genotype imputation of target haplotypes.

| | | Overall | NMA[1] = 2 | MAF < 0.05 | MAF > = 0.05 | First Mb | Last Mb | Between first and last Mb | Number of SNP imputed as monomorphic |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | 325,358 | 4,020 | 41,931 | 283,427 | 2,587 | 2,370 | 320,401 | |
| HMM-LD | Avg | 91.86 | 71.89 | 80.96 | 93.47 | 87.89 | 87.74 | 91.92 | 125 |
| | Med | 94.93 | 99.15 | 90.22 | 95.04 | 92.61 | 90.30 | 95.00 | |
| IMPUTE2 | Avg | 91.93 | 71.85 | 81.00 | 93.55 | 87.91 | 87.76 | 92.00 | 157 |
| | Med | 94.97 | 99.14 | 90.20 | 95.10 | 92.21 | 90.39 | 95.03 | |
| EXT-100k | Avg | 92.01 | 72.31 | **81.52** | 93.56 | 88.74 | **89.99** | 92.05 | 455 |
| | Med | 94.89 | 99.43 | 90.65 | 95.00 | 92.51 | 93.34 | 94.94 | |
| EXT-1M | Avg | **92.08** | **72.33** | 81.50 | **93.65** | **89.28** | 89.60 | **92.12** | 444 |
| | Med | 94.94 | 99.43 | 91.16 | 95.08 | 92.48 | 92.89 | 95.00 | |

*Average and median imputation $r^2$ (as percentages) of four different imputation methods, partitioned by allele frequency and by position on BTA1, after exclusion of LD SNPs as well as any SNP imputed as monomorphic by at least one of the four methods. For each partition, the best average result is boldfaced. [1]NMA, number of occurrences of Minor allele.*

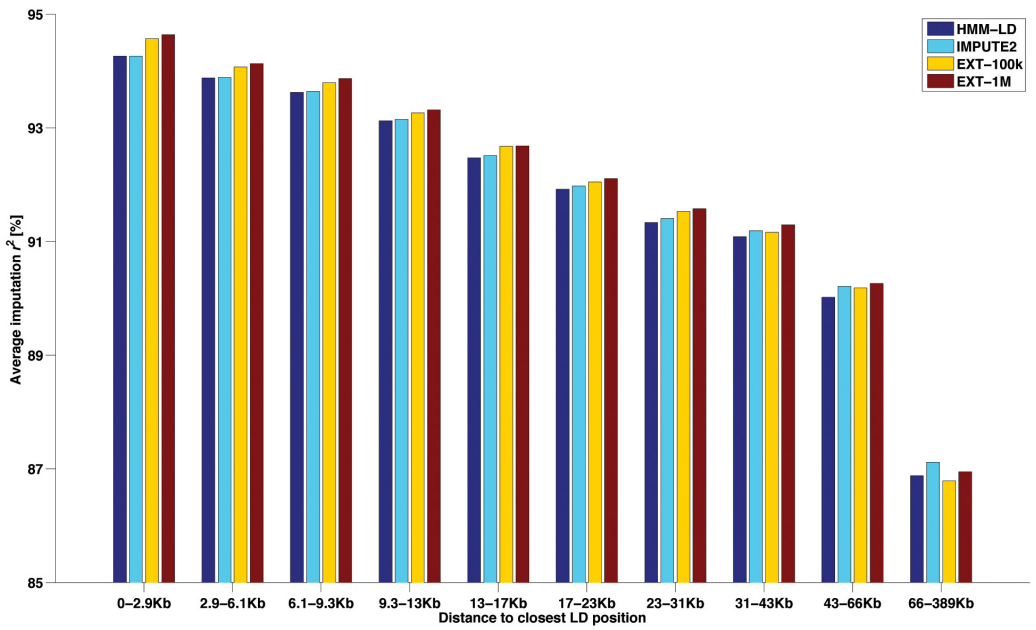


**FIGURE 6 |** Average imputation $r^2$ by four methods with regard to the distance between the imputed SNP (from the HD map) and the closest observed SNP (from the LD map), for different classes of distance containing the same number of imputed SNP.

*max_features* parameter was set to vary between few and low values (1 or 2) could explain this unexpected result. For the purpose of imputation, considering only some features never reach average imputation reliabilities higher than those of the HMM.

# DISCUSSION

## Genotype Imputation Illustrates the Effectiveness of the Random Forests Framework

When imputing WGS genotypes from 50k dense genotypes, the implemented random forests framework reaches average reliabilities similar to those achieved by IMPUTE2. We consider therefore these reliabilities as fair evidence of the ability of our framework to efficiently learn how to locally match haplotypes from examples (the labeled observations) for two main reasons. First, such a measure is independent of phasing, thus it does not embed potential phasing errors (even though those remain scarce). Second, using the imputation criterion makes it comparable to a state-of-the-art method, here IMPUTE2. Imputation results of the two types of methods are very similar, although we observed two main differences between HMM and the random forests framework. The first is that the random forests framework performs better on both edges of chromosomes: a difference of ∼2% of average imputation $r^2$ is observed. The second difference is that IMPUTE2 imputes genotypes at distant positions from known genotypes with higher accuracy; this is due

**TABLE 5 |** Effect of considering only some features and not others, on average difference $e_A$ between predicted and true target haplotypes and on average imputation $r^2$.

| | $e_A$ | $r^2$ | Number of SNP imputed as monomorphic |
|---|---|---|---|
| ***N*** | **182** | **324,738** | |
| HMM-HP-LD\| HMM-LD | 1.304 | 92.00 | **125** |
| HMM-HP-HD\| IMPUTE2 | 1.310 | 92.07 | 157 |
| EXT-100k | **1.236** | **92.15** | 455 |
| EXT-100k-L | 1.240 | 91.73 | 577 |
| EXT-100k-LR | 1.238 | 91.83 | 692 |
| EXT-100k-H | 1.304 | 91.47 | 613 |
| EXT-100k-HR | 1.345 | 91.03 | 914 |

*Both comparison criteria are given as percentages and best results are boldfaced.*

to its use of the HD map, as shown by comparison with HMM-LD in **Figure 6**.

## Conceptual Differences Between the HMM and the Random Forests Framework

The differences in imputation results could be explained by the views behind the two types of methods, which also are quite distinct. The very basic conceptual difference between them lies in their modeling objectives: the HMM seeks to find the sequence of reference haplotypes that most likely reproduces an observed target haplotype (hence, essentially minimizing the number of segments) while our proposed framework searches for the best match locally (independently of the whole sequence). In some particular designs, the reference haplotypes correspond to the true ancestors of the target haplotype (e.g., Mott et al., 2000; Druet and Farnir, 2011; Zheng et al., 2015); then the HMM models the biological process of chromosomes transmission over a few generations. In contrast, the sequence returned by the random forests framework has no pretention to model that biological process but aims at imputing the target haplotype as well as possible, chunk after chunk. When the reference haplotypes are not the true ancestors of the target haplotype (e.g., when the target haplotype is not a true mosaic of reference haplotypes), the HMM framework no longer aims at finding the reference haplotype that is the most likely to be identical-by-descent (IBD) with the target haplotype at a given position but essentially minimizes the number of segments in the mosaic. Conversely, the random forests framework searches for the best match haplotype similarly to methods estimating IBD probability, considering the number of identical-by-state SNPs on both sides of the position (e.g., Meuwissen and Goddard, 2001). The natural consequence of these two different modeling purposes is a much higher level of "mosaicism" for the random forests framework (given in **Table 3**).

Beyond that first conceptual difference, another two are of interest. First, our framework does not allow for small differences between shared segments: a mismatch between target and reference haplotypes terminates a shared segment. For some methods (e.g., Beagle – Browning and Browning, 2009), more efficient imputation results have been observed without allowing differences. Not allowing differences also partially explains why the extra-trees makes more switches than the HMM. Note that the same constraint could be imposed in the HMM framework by setting $P_{error}$ to 0. Second, the two types of methods use different map information: the random forests framework only obtains information from the LD map whereas the HMM may additionally obtain information from the HD map. That difference matters since the HMM achieves better imputation with the HD map than with the LD map (particularly for HD SNPs distant from a LD position, see **Figure 6**). When it uses the entire map, the HMM better accounts for distances between SNP positions and for the structure of linkage disequilibrium between SNPs. It subsequently produces a better estimation of the haplotype blocks: a block is defined by SNPs in perfect linkage disequilibrium, not by those closest to a LD position. Integrating the information from the HD map into the random forests framework would therefore be profitable.

## Main Lessons of the Extra-Trees Classifier

Beyond its use, the random forests framework also reveals some useful lessons for the development of methods for local haplotype matching. The most informative lesson comes from the importance ranking of the features: top-ranked features are those expressing the distance to an edge of a shared segment (e.g., DMN, minimal distance to the left or right edge of the shared segment, or iDMN, its expression on an inverse scale). When such a feature is not equal to zero, it contains a double information: (1) that both haplotypes are, at this position, in a shared segment and (2) the value of the distance to the edges of the segment. A high value of DMN (or a value of iDMN close to 2) reveals that both haplotypes share a long identity segment (at least twice the length of the value of DMN) and that the current position is quite distant from the closest edge of this identity segment. The distance to the edge of a shared segment is thus more important than the length of this shared segment. As discussed above, the distance to the closest edge might better reflect relative local IBD probabilities than the length of the shared segment. Accordingly, minimizing the number of segments in the mosaic as done in the HMM does not guarantee the identification of the reference haplotype with the highest local IBD probability.

Before going further, note that the precedence of iDMN over DMN (and similarly for iDRN, iDLN) can be explained by the nature of extra-trees itself: for any node split when growing a decision tree, the extra-trees algorithm randomly picks up the value of the cut-point for a feature uniformly between the min and max value of this feature in the node to split. However, the sizes of classes of iDMN are more uniformly distributed over its (bounded) range than the sizes of classes of DMN (see **Figure 5**: for DMN, >98% of the observations fall into the first bin of range). Therefore, when picking at random a cut-point for node splitting, there is a higher chance of having an informative

discrimination with iDMN compared with DMN. With classical random forests (where cut-points are optimized over the full range of values), iDMN and DMN have similar importance (see **Supplementary Material S4**).

Features rankings (**Figure 4**) also show that features of the first group (i.e., 16 features related to the position along the chromosome) unambiguously take the precedence over the ones of the second group (relationships). Such hierarchy was then confirmed by the tests with a reduced number of features (**Table 5**). This result was expected in the sense that the relationship features express identity between haplotypes at maximum at the chromosome level (feature GENGc, which actually is the most important of these features) whereas features from the first group express identity between haplotypes at a segment level (e.g., a high value of feature LSS reveals an identity spanning on several Mb). A second lesson is thus that relationship features have a small but not null impact: removing them from the random forests framework leads to average imputation reliabilities lower than those of the HMM (**Table 5**). Our explanation is that these relationships are still useful to discriminate between reference haplotypes bearing a shared segment of the same length, although for most of the cases the length of the shared segment already captures the familial information (long segments indicating close relationships). Consequently, using relationship to pre-select the subset of reference haplotypes, as done by SHAPEIT2 (Delaneau et al., 2011) or by LDMIP (Meuwissen and Goddard, 2010), is probably already a good way to use this information. Similarly, we observed that adding the relationship information to the HMM information (in the random forests framework) did not improve our accuracy.

The rankings of features (**Figure 4**) bring other minor lessons about features expressing the same aspect, but in a different way. First, feature NSS is always preferred to feature LSS, whereas both express the length of a shared segment between target and reference haplotypes (respectively in number of LD map positions and in kb). Second, the dense rankings are of little help: standard rankings ("R1-") always take precedence over them ("R2-"). The rationale behind the use of the dense rankings was to make comparable cases where many reference haplotypes were the best match to cases where only one reference haplotype was the best match. In both situations, with dense ranking ("1123"), the second-best reference haplotype is ranked second whereas, with standard ranking ("1134"), the second-best reference haplotype is ranked $n + 1$, where $n$ is the number of best matching haplotypes.

## Perspectives and Improvements for Routine Use of the Random Forests Framework

As implemented in our study, the random forests framework is not computationally competitive compared to the existing HMM approaches. Hence, prior to a routine application, two entangled aspects have to be considered: how does one achieve routine predictions with higher accuracy, and with lower computational demand than the random forests framework as implemented so far? Both aspects can be circumscribed to the constitution of the learning samples, summarizing the previous question to reducing the dimensions of these learning samples (number of labeled observations per number of features) along with improving accuracy.

On the aspect of the number of features, the tests conducted in this study have shown that discarding features could lead to very limited losses of precision but should not be done in a group-wise manner. Now that the hierarchy of features have been established inside each group, some features could be trimmed off to avoid redundancy, i.e., giving preference to iDMN over DMN, to NSS over LSS, or to R1- over R2. For instance, an optimized set of features may also be obtained through recursive feature elimination (Guyon et al., 2002). Besides removing less important features, new ones could also be investigated. Note that preliminary investigations are, however, always necessary for new features; for instance, we had considered the gametic linkage (as estimated in Wang et al., 1995) but too few relationships were non-zero so that it was helpless to identify best local matches between haplotypes. The IBD probabilities, as estimated by Beagle (Browning and Browning, 2009) or LDMIP (Meuwissen and Goddard, 2010), could also be considered although the usefulness of such features might be hampered by the time requested for computing them. Other features to consider are the allele (as in Maples et al., 2013), the MAF and the position of HD SNPs. These features would extend the learning sample to all HD positions, which would undoubtedly be profitable for accuracy. Conversely, this would directly impact the computational aspect. For that reason, an intermediate solution would be to consider blocks of linkage disequilibrium of HD SNPs (and their allele, MAF and position) instead of operating on these HD SNPs. All lengths and distances could also be expressed on a different scale to account for the average number of generations between target and reference haplotypes as in the HMM framework (e.g., using genetic distances and the number of generations to estimate recombination probabilities).

The number of labeled observations is the second aspect to consider and should be optimized alongside the number of features. Our results show a limited improvement when using a learning sample 10-times larger (EXT-1M vs. EXT-100k). The number of labeled observations could therefore be reduced. In addition, their selection could be achieved in a wiser manner, e.g., selecting them in order to contain the most different examples rather than randomly. The problem of the selection of the best training examples is known as active learning in machine learning literature (Settles, 2012).

# CONCLUSION

We herein outlined a new framework for automatically matching haplotypes along the chromosome and have

illustrated that extremely randomized trees can effectively combine multiple sources of information to identify the best matching reference haplotypes. As an example, our implementation of the extremely randomized trees achieved slightly better imputation results than IMPUTE2. The random forests framework also allows identifying which features are the most important for a specific prediction. In the present case, distance to the edges of the shared segment appeared as the most important variable and adding genomic relationships only marginally improved results. To conclude, this approach might be further enhanced, for instance by including additional features, or could also be applied to other related applications such as identification of carriers of genetic defects or imputation of structural variants (by including features as distance with known carriers, genotyping intensity, etc.).

## AUTHOR CONTRIBUTIONS

PF, PG, and TD conceived the study, interpreted the results, and wrote the manuscript. PF and TD developed the tools and software. PF carried out the experiments. All authors read and approved the final manuscript.

## FUNDING

This research was supported by the Fonds National de la Recherche Scientifique (F.R.S.-FNRS) (TechILA project – Grant T.1086.14) and the University of Liège (BluePOOL project – Fonds Spéciaux de la Recherche). The supercomputing facilities of the Consortium d'Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles (CECI) was funded by the F.R.S.-FNRS.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge Livestock Improvement Corporation (Hamilton, New Zealand) for providing the material used in this study. TD is a Senior Research Associate from the F.R.S.-FNRS.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00562/full#supplementary-material

## REFERENCES

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367. doi: 10.1093/bioinformatics/bts144

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

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

Burdick, J. T., Chen, W.-M., Abecasis, G. R., and Cheung, V. G. (2006). In silico method for inferring genotypes in pedigrees. *Nat. Genet.* 38, 1002–1004. doi: 10.1038/ng1863

Charlier, C., Li, W., Harland, C., Littlejohn, M., Coppieters, W., Creagh, F., et al. (2016). NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res.* 26, 1333–1341. doi: 10.1101/gr.207076.116

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., and Goddard, M. E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics* 189, 317–327. doi: 10.1534/genetics.111.128082

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

Druet, T., and Farnir, F. P. (2011). Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. *Genetics* 188, 409–419. doi: 10.1534/genetics.111.127720

Druet, T., and Georges, M. (2010). A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* 184, 789–798. doi: 10.1534/genetics.109.108431

Faux, P., and Druet, T. (2017). A strategy to improve phasing of whole-genome sequenced individuals through integration of familial information from dense genotype panels. *Genet. Sel. Evol.* 49:46. doi: 10.1186/s12711-017-0321-6

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition, Corrected at 12th Printing*. New York, NY: Springer.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* 40, 1068–1075. doi: 10.1038/ng.216

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* 8:e1002453. doi: 10.1371/journal.pgen.1002453

Li, Y., Ding, J., and Abecasis, G. R. (2006). Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* 79:S2290.

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi: 10.1016/j.ajhg.2013.06.020

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

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

Meuwissen, T. M. H., and Goddard, M. E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* 33, 605–634. doi: 10.1051/gse:2001134

Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12649–12654. doi: 10.1073/pnas.230304397

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., et al. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5:e1000519. doi: 10.1371/journal.pgen.1000519

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

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

Schaeffer, L. R., Kennedy, B. W., and Gibson, J. P. (1989). The inverse of the gametic relationship matrix. *J. Dairy Sci.* 72, 1266–1272. doi: 10.3168/jds.s0022-0302(89)79231-6

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802

Settles, B. (2012). Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 6, 1–114.

Speed, D., and Balding, D. J. (2014). Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* 16, 33–44. doi: 10.1038/nrg3821

Su, Z., Cardin, N., The Wellcome Trust Case Control Consortium, Donnelly, P., and Marchini, J. (2009). A bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.* 24, 430–450. doi: 10.1214/09-STS311

Wang, T., Fernando, R. L., van der Beek, S., Grossman, M., and von Arendonk, J. (1995). Covariance between relatives for a marked quantitative trait locus. *Genet. Sel. Evol.* 27, 251–274. doi: 10.1186/1297-9686-27-3-251

Wright, S. (1922). Coefficients of Inbreeding and relationship. *Am. Nat.* 56, 330–338. doi: 10.2307/2456273

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Zheng, C., Boer, M. P., and van Eeuwijk, F. A. (2015). Reconstruction of genome ancestry blocks in multiparental populations. *Genetics* 200, 1073–1087. doi: 10.1534/genetics.115.177873

**Conflict of Interest Statement:** 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: 23 September 2019
doi: 10.3389/fgene.2019.00890

# Rediscover and Refine QTLs for Pig Scrotal Hernia by Increasing a Specially Designed $F_3$ Population and Using Whole-Genome Sequence Imputation Technology

*Wenwu Xu, Dong Chen, Guorong Yan, Shijun Xiao, Tao Huang, Zhiyan Zhang* and Lusheng Huang**

*State Key Laboratory for Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University, Nanchang, China*

OPEN ACCESS

**Edited by:**
*Gábor Mészáros, University of Natural Resources and Life Sciences Vienna, Austria*

**Reviewed by:**
*Fabyano Fonseca Silva, Universidade Federal de Viçosa, Brazil*
*Juan José Arranz, Universidad de León, Spain*

***Correspondence:**
*Zhiyan Zhang*
*bioducklily@hotmail.com*
*Lusheng Huang*
*lushenghuang@hotmail.com*

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

**Received:** *18 June 2019*
**Accepted:** *23 August 2019*
**Published:** *23 September 2019*

**Citation:**
*Xu W, Chen D, Yan G, Xiao S, Huang T, Zhang Z and Huang L (2019) Rediscover and Refine QTLs for Pig Scrotal Hernia by Increasing a Specially Designed $F_3$ Population and Using Whole-Genome Sequence Imputation Technology. Front. Genet. 10:890. doi: 10.3389/fgene.2019.00890*

Pig scrotal hernia is one of the most common congenital defects triggered by both genetic and environmental factors, leading to severe economic loss as well as poor animal welfare in the pig industry. Identification and implementation of genomic regions controlling scrotal hernia in breeding is of great appeal to reduce incidences of hernia in pig production. The aim of this study was to identify such regions or molecular markers affecting scrotal hernia in pigs. First of all, we summarized and analyzed the results of some international teams on scrotal hernia and designed a specially population which contains 246 male individuals. We then performed genome-wide association study (GWAS) in this specially designed population using two scenarios, i.e., the target panel data before and after imputation, which contain 42,365 SNPs and 18,756,672 SNPs, respectively. In addition, a series of methods including genetic differentiation analysis, linkage disequilibrium and linkage analysis (LDLA), and haplotype sharing analysis were appropriate to provide for further analysis to identify the potential gene underlying the QTL. The GWAS in this report detected a highly significant region affecting scrotal hernia within a 24.8Mb region (114.1–138.9Mb) on SSC8. And the result of genetic differentiation analysis also showed a strong genetic differentiation signal between 116.1 and 132.7Mb on SSC8. In addition, the QTL interval was refined to 2.99Mb by combining LDLA and genetic differentiation analysis. Finally, two susceptibility haplotypes were identified through haplotype sharing analysis, with one potential causal gene in it. Our study provided deeper insights into the genetic architecture of pig scrotal hernia and contributed to further fine-mapping and characterize haplotype and gene that influence scrotal hernia in pigs.

**Keywords: GWAS, imputation, haplotype, specially designed population, scrotal hernia, pigs**

## INTRODUCTION

Pig hernias are of the most common congenital defects which cause severe economic losses as well as poor animal welfare in the pig industry. The most common types of hernias in pig are scrotal and umbilical hernia. Scrotal hernia is the phenomenon of abdominal contents falling into scrotum from the unilateral or bilateral inguinal rupture, causing local expansion bulge (Grindflek et al.,

2006; Du et al., 2009; Zhao et al., 2009). As a complex congenital defect, the reason of scrotal hernia formation is unclear; some abnormal phenomena and problems occurred at the stage of the development and obliteration of processus vaginalis in descent of testis, which have been considered to be the main reason for the development of scrotal hernia (Clarnette and Hutson, 1997; Clarnette et al., 1998). The genetic mechanical of scrotal hernia is also poorly clarified, only with the knowledge of cause by both multiple genetic and environmental factors. In the pig breeding industry, the occurrence of scrotal hernia is varied from 1.7 to 6.7% across from pig breeds and populations, and the heritability estimation varied from 0.2 to 0.6 in disparate studies (Mikami and Fredeen, 1979; Thaller et al., 1996). Environmental factors, as a potential factor in the occurrence of complex genetic diseases, have a great influence on the occurrence of scrotal hernia. Research reports showed that the incidence of scrotal hernia in Dutch Landrace and large white pig was 1.36 and 1.31%, respectively, while the corresponding incidence rate of Dutch Landrace and large white pig of Hypor was 0.54 and 0.22% (PK, 2006). In 2010, the European Breeding Corporation reported that the incidence of scrotal hernia in Dutch Landrace and large white pig was 0.383% (Walters, 2010). Obviously, the difference of environment will make the incidence of scrotal hernia different.

In breeding practice, it is not effective to decrease the incidence of pig scrotal hernia by conventional phenotypic selection. One of the methods of hernia resistance breeding is to isolate and identify susceptibility loci and major causative genes and then implement marker assisted selection. Currently, several research groups have identified the susceptible loci and potential positional candidate genes for scrotal hernia. Grindflek et al. reported several susceptibility QTLs for pig scrotal hernias on eight chromosomes (Grindflek et al., 2006). Ding et al. have revealed seven regions on SSC2, 4, 8, 10, 13, 16, and 18 for scrotal hernia in a White Duroc and Erhualian $F_2$ intercross using nonparametric genome-wide linkage (NPL) analysis and transmission disequilibrium test (TDT) (Ding et al., 2009). Du et al. found that four regions surrounding *ELF5, KIF18A, COL23A1* on chromosome 2, and *NPTX1* on chromosome 12 may contain the genetic variants important for the development of the scrotal hernia development using a family-based analysis (Du et al., 2009). Sevillano et al. reported a susceptibility region on SSC13 between 34 and 37 Mb for scrotal hernia (Sevillano et al., 2015). However, these susceptibility areas are rarely further confirmed in other research groups; even using bigger population sizes, the genetic control of scrotal hernia has still not been clarified.

In the 10 years, we performed two statistical methods (TDT and NPL) in the $F_2$ population using 194 microsatellites and identified one chromosomal region distributed on SSC8 for the scrotal hernia. Generally speaking, nonparametric linkage analysis (NPL) evaluates allele sharing among affected individuals and comes to a result without particular model assumptions, and the TDT was proposed as a family-based association test for the presence of genetic linkage between a genetic marker and a trait; more computational details with this 2 statistical methods were showed by Ding et al. (2009). Using the same population, we perform GWAS study in 60K genotypes, the result manifested that none of SNPs achieved the genome-wide significance threshold (Su et al., 2014). The feasible reasons for the "missing QTLs" in GWAS study probably are the low linkage disequilibrium between markers and low incidence rate in the subject population, or due to the intricacy genetic basis of this congenital defect. To overcoming these problems and exploring this congenital defect, we designed a specially $F_3$ population which was mated with full-sibs or half-sib of the affected individuals and imputed the chip SNPs to whole-genome sequences (**Supplementary Figure S1**) then implemented several classical genetic methods to rediscover and refine QTLs for pig scrotal hernia. Our aim in this study was to identify susceptibility loci of pig scrotal hernia and provided a novel insight for further analysis of the genetic basis of this congenital defect.

## MATERIAL AND METHOD

All procedures including experimental animals established and tissue collection were performed in accordance with the guidelines approved by the Ministry of Agriculture of China. This study was approved by the ethics committee of Jiangxi Agricultural University.

## ANIMALS OF THE TARGET POPULATION

A four-generation resource population was developed from the intercross of 2 White Duroc boars (PIC 1075) and 17 Chinese Erhualian sows between 2,000 to 2,006. In briefly, two White Duroc boars were crossed to 17 Erhualian sows, then 9 $F_1$ boars, and 59 $F_1$ sows were randomly selected to produce a total of 1,912 $F_2$ pigs in 6 batches avoiding full-sib mating (Guo et al., 2009). Last, 62 $F_2$ boars and 149 $F_2$ sows were selected to produce two types of $F_3$ population. The ordinary experiment population contains 661 $F_3$ offspring from an intercross of randomly chosen $F_2$ avoiding full-sib mating; the particular hernia population in this study contains 851 $F_3$ offspring, which were designed to mate the health full-sibs or half-sibs of affected individuals. Affected pigs were diagnosed and recorded carefully by veterinarians at three age stages: 46, 90, and 240 days. In summary, 23 affected pigs from $F_2$ population were confirmed, 5 affected pigs from ordinary $F_3$ population, and 23 affected pigs from $F_3$ hernia study population were diagnosed, respectively. A total of 1,020 individuals (19 $F_0$, 68 $F_1$, and 933 $F_2$) and 500 $F_3$ were genotyped. For this study, 246 male $F_3$ pigs were chosen for GWAS analysis, which contain 18 available DNA samples for affected individuals. Furthermore, 19 $F_0$, 68 $F_1$, and 516 $F_2$ male pigs, and 246 $F_3$ male pigs were used in haplotype sharing analysis.

Genomic DNA was isolated from ear tissue with a standard phenol/chloroform extraction method. All DNA samples were qualified and diluted to a final concentration of 50 ng/μl in 96-well plates. A total of 1,020 $F_2$ and 500 $F_3$ were genotyped with the Illumina PorcineSNP60 BeadChip and GeneSeek GGP Porcine 50K BeadChip on an iScan System (Illumina, USA) following the manufacturer's protocol, respectively (Ramos et al.,

2009). Physical positions of SNPs on chromosomes referred to the swine reference genome sequence assembly (Sus_scrofa11.1) (http://asia.ensembl.org/Sus_scrofa/Info/Index). Quality control procedures were implemented by PLINK (version 1.07). Briefly, SNPs were removed if their positions on the genome build 11.1 were unspecific, call rate <90%, and minor allele frequency (MAF) <1%. Animals more than 10% missing genotypes were removed. To keep the alleles consistency with the sequencing data, we firstly aligned the primer sequences of each SNP to the reference porcine genome assembly Sus scrofa 11.1 by BLAST. Then, the genotypes of reversed SNP strands in target panel were flipped using PLINK (v1.9) software (Chang et al., 2015); SNPs without positions were excluded for further analysis.

## HAPLOTYPE CONSTRUCTION OF REFERENCE PANEL

In this study, a wide collection of 109 whole-genome sequence individuals from 14 difference populations were used as a reference; each breed contained 2 to 22 individuals. More details on the origins, breeds, and sample size are shown in **Table 1**. We firstly trimmed the raw reads according to a quality score threshold greater than 15; then, BWA (Burrows–Wheeler Aligner) was used to align the raw reads which passed chastity filtering to the reference porcine genome assembly Sus scrofa11.1 (Li and Durbin, 2009). Variants were identified using the GATK (Genome Analysis Toolkit) (McKenna et al., 2010); PCR duplications were firstly marked by Picard MarkDuplicates (http://broadinstitute.github.io/picard/), and GATK IndelRealigner option was carried out for local realignments. Then, variants were filtered with GATK VariantFiltration option. VCFtools was used to remove the structural variants. Subsequently, the haplotypes of 109 individuals with cleaned SNP data were constructed by Beagle (v4.1) (Browning and Browning, 2007). Specifically, the number of markers to include in each sliding window was set to 100,000, and the overlap between windows was set to 3,000 markers. Then, the number of phasing iterations was set to 50. Finally, the other options involving in the imputation follow the default setting.

**TABLE 1** | The components of the reference panel.

| Breeds | Sample Size | Depth | Location |
|---|---|---|---|
| Bamei | 6 | 24.9 | Shanxi, China |
| Hetao | 6 | 24.4 | Inner Mongolia, China |
| Laiwu | 6 | 27.5 | Shandong, China |
| Min | 6 | 25.8 | Heilongjiang, China |
| Bamaxiang | 6 | 28.1 | Guangxi, China |
| Luchuan | 6 | 26.4 | Guangxi, China |
| Wuzhishan | 6 | 26.1 | Hainan, China |
| Jinhua | 6 | 26.2 | Zhejiang, China |
| Erhualian | 19 | 28.1 | Jiangsu, China |
| Tibet | 22 | 26.9 | Southwest China |
| Baoshan | 6 | 26.5 | Yunnan, China |
| Neijiang | 6 | 26.2 | Sichuan, China |
| White Duroc | 2 | 31.1 | USA |
| Wild boar | 6 | 28.9 | South China; North China; Sumatra, Indonesia |

## IMPUTATION

Whole-genome sequence imputation between target and reference panel was conducted by Beagle (v4.1) using the default parameter settings (Browning and Browning, 2016). Specifically, the size of imputed region was set to 50,000 markers per window, and the overlap between windows was set to 3,000 SNPs. This software first constructed local haplotypes using the hidden Markov chain Monte Carlo (MCMC) algorithm and then resampled new estimated haplotypes for each individual based on a hidden Markov model (HMM).

Imputation accuracy should be further investigated in whole-genome sequence data because of the low density and common variants in 50k. Browning et al. and Williams et al. have fully exhibited the number of individuals present in a population is a crucial factor in determining how well the phase can be estimated for haplotype construction (Browning and Browning, 2011; Williams et al., 2012). Therefore, 109 whole-genome sequence pigs including 19 $F_0$ who were the progenitor of the 500 $F_3$ populations were also regarded as reference panel in order to obtain more accurate phase information. Then, the genotypic concordance rate and the squared correlation (R2) between best-guess imputed and the original variants as imputation accuracy. The genotypic concordance rate used a cross-validation strategy described in previous studies (Brondum et al., 2014; van Binsbergen et al., 2014; Pausch et al., 2017). More specifically, two thousand loci in the target sample were deleted randomly then imputed in the same strategy. The number of 2,000 alleles imputed correctly divided by total 2,000 loci (the allelic correct rate) was taken to calculate the accuracy of imputation. Finally, in order to balance the imputation accuracy and missing proportion in the next analysis process, we excluded the variants with call rate <90% and MAF <0.03.

## GWAS

GEMMA was utilized to perform the association analyses underlining the standard linear mixed model (Zhou and Stephens, 2012). Sex and batch were included as fixed effects. Heritability was estimated by using –lmm procedure implemented in GEMMA using genomic relationship matrix. Population stratification and were adjust by including genomic relationship matrix. Briefly, this model is denoted as:

$$y = W\alpha + X\beta + u + \epsilon; u \sim MVN_n(0, \lambda^{-1}\mathrm{K}), \epsilon \sim MVN_n(0, \lambda\tau^{-1}\mathrm{I}_n)$$

where y is a n element vector of phenotypic values (or case/control labels), α is a c-vector of fixed effects, $\beta$ is the effect size of SNPs, W is a design matrix of covariates, x is a vector of genotypes at each locus, and u is the vector of random effects following the multivariate normal distribution $MVN_n(0, \lambda\tau^{-1}K)$, where $\tau^{-1}$ is the variance of the residual errors, and λ is the ratio between $\tau^{-1}$ and the variance of the residual errors; K is a known kinship matrix, ∈ is an vector of errors following the multivariate distribution $MVN_n(0, \lambda\tau^{-1}I_n)$, and $I_n$ is an n × n identity matrix. Normally, significance threshold of multiple

test in chip array-based GWAS was adjusted by naïve Bonferroni corrections, which is 0.05 divided number of examined SNPs. However, this approach would lead to over correction and decreasing the detection power in GWAS as these tests are non-independent for the linkage disequilibrium between markers. We herein used 5E−08 as a genome-wide suggestive significance threshold following Pe'er et al. and Johnson et al. (Pe'er et al., 2008; Johnson et al., 2010). The population stratification is one of the factors that affects the validity of genome-wide association study (Pearson and Manolio, 2008). To check if stratification exists in our result, quantile–quantile plots (Q–Q plots) were implemented to evaluate population stratification effects. The Q–Q plots were constructed with R software. Measures of linkage disequilibrium (r and $r^2$) between SNPs were estimated by plink 1.07 (Clarnette and Hutson, 1997), the default settings for minimum linkage between SNPs at threshold $r^2 = 0.8$.

## GENETIC DIFFERENTIATION ANALYSIS

To elucidate whether there is genetic differentiation exist in scrotal hernia pigs and health pigs, we divided the affected pigs and the unaffected pigs into two groups, as the method did by Zhang et al. (2019) then assessed allele frequency differentiation using the unbiased genetic differentiation estimated of the fixation index (Fst). Akey et al. have fully described estimation of unbiased Fst fixation index in his paper using SNP dataset (Akey et al., 2002). Briefly, Fst was estimated as follows:

$$Fst = \frac{MSP - MSG}{MSP + (n_c - 1)MSG}$$

where MSG represents the observed mean square errors for loci within populations, MSP denotes the observed mean square errors for loci between populations, and $n_c$ is the average sample size across samples, which incorporates and corrects for the variance in the sample size over population

$$MSG = \frac{1}{\sum_{i=1}^{s} n_i - 1} \sum_{i}^{s} n_i p_{Ai}(1 - p_{Ai})$$

$$MSP = \frac{1}{s-1} \sum_{i}^{s} n_i (p_{Ai} - \bar{p}_A)^2$$

$$n_c = \frac{1}{s-1} \sum_{i=1}^{s} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$$

In the above formulae, $n_i$ and $p_{Ai}$ denote the sample size and the frequency of SNP allele A in the *i*th population, respectively, and $\bar{p}_A$ is a weighted average of $p_A$ across populations. The negative Fst didn't have any biological interpretation and were set to 0 to fit the definition of Fst ranging from between 0 and 1 (Wright, 1951). The top 1% of loci according to genetic differentiation values was served as candidate regions to host resistance or susceptibility to pig scrotal hernia (Zhang et al., 2019).

## LINKAGE DISEQUILIBRIUM AND LINKAGE ANALYSIS (LDLA)

The haplotypes of F3 on SSC8 were reconstructed using a hidden Markov model by beagle (Zhang et al., 2012) and then the graphical model for the haplotype clusters with beagle was directly generated, which is a directed acyclic graph (DAG). The parameters for both processes are set to scale equals 2 and shift equals 0.1. Haplotypes within a cluster are likely to descend from the same ancestral haplotype and to carry the same DSV (DNA sequence variants) and combination of alleles, which is actually the principle used in linkage analysis. The linkage disequilibrium or association mapping information is generated by ancestral recombinations and detected by population level associations between individuals. Then, the clustered haplotypes were converted into diallelic markers by pseudomarker program, which can be imported into a program like R for statistical analysis. Thus, haplotype data contains both linkage and linkage disequilibrium information and can be imported into a mixed model framework:

$$Y = Xb + Zu + e$$

where Y is the vector of phenotypes, and b is fixed effects including sex and batch. The haplotypes could be treated as random here, as there are likely to be many of them, and some haplotypes will occur only a small number of times. Therefore, the random additive genetic effect following the distribution $u \sim N(0, \sigma_u^2 G)$, in which G is the individual–individual similarity matrix, and $\sigma_u^2$ is the polygenetic additive variance, and X and Z are incidence matrices for b and u, respectively. The residual random effect "e" following the distribution $e \sim N(0, \sigma_e^2 I)$. The LDLA analysis was carried out using a homemade R scripts (**Supplementary Data Sheet 2**). The most likely position of the QTL was obtained by the 2-LOD drop method (Karim et al., 2011).

## HAPLOTYPE SHARING ANALYSIS

The haplotypes in the target QTL region were constructed by fastPHASE. Firstly, we tried to find the sharing susceptibility haplotype by thoroughly scanning the haplotypes of affected individuals in $F_3$ population. Then, we tried to identify whether the same sharing susceptibility haplotype existed in $F_2$ affected individuals and tried to trace it to the $F_1$ and $F_0$ generations. It should be noted that we take the intersection of SNPs of $F_3$ and $F_2$ due to the different density of 50 and 60k chip.

## CONDITIONAL ASSOCIATION TEST

To elucidate whether there are additional QTLs for scrotal hernia in the identified QTL region, we extracted genotypes of the top SNP and included as a covariate to the univariate linear mixed

model, which was performed in the single-marker GWAS as we described above then performed a conditional test to retest the association between SNPs and phenotypes. If additional signal was detected, then there were multiple QTLs that cooperated to control scrotal hernia. Otherwise, there was only one QTL that affected scrotal hernia.

## BOOTSTRAP TEST

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement (Efron and Tibshirani, 1993). It can be used to estimate summary statistics such as the mean, standard deviation, confidence interval, or correlation coefficient, which is done by repeatedly taking small samples, calculating the statistic, and taking the average of the calculated statistics. We herein carried out bootstrap test to verify the reliability of GWAS in this study. First, we randomly resampled for 1,000 times with replacement, in which some affected individuals can be sampled for multiple times, while some may be sampled for 0 times, the total number of affected individuals that may either increase or decrease, and the same resample results were acquired in unaffected individuals. Then, we conducted GWAS for 1,000 times to see if there were still significant signals in the susceptibility region which was identified in our study.

## RESULTS

### Phenotype Statistics and SNP Characteristics After Quality Control

Incidences of scrotal hernia were estimated to be 0.7 and 2.7% in the ordinary $F_3$ population and in the specially designed $F_3$ population, respectively. It is obvious that the incidence of scrotal hernia in the specially designed $F_3$ population was significantly higher than in the ordinary $F_3$ population. Heritability for scrotal hernia was estimated at 0.39 using the standard linear mixed model, which implies that there is a genetic contribution to scrotal hernia.

After quality control, a total of 42,365 SNPs and 246 pigs had retained for further analyses. Imputation was produced using Beagle software. The summarization of imputation results is presented in **Table 2**. After imputation, a total of 46,483,626 SNPs for 246 individuals were obtained, and 18,756,672 SNPs were retained after filtering with MAF > 0.03. The average genotypic concordance rate was 84.8%, and the average correlation between best-guess and true variants reached with an average of 71% after we delete sites where $R^2$ is equal to 0 and MAF is less than 0.03 (**Supplementary Figure S2**).

**TABLE 2 |** The distribution of SNPs in different chromosomes.

| Chr | Before QC | After QC |
|---|---|---|
| Chr1 | 4,735,710 | 1,871,922 |
| Chr2 | 3,000,496 | 1,128,593 |
| Chr3 | 2,841,406 | 1,161,477 |
| Chr4 | 2,711,334 | 1,140,103 |
| Chr5 | 2,259,813 | 971,685 |
| Chr6 | 3,399,128 | 1,410,997 |
| Chr7 | 2,647,087 | 1,094,716 |
| Chr8 | 2,786,786 | 1,193,047 |
| Chr9 | 2,942,680 | 1,159,718 |
| Chr10 | 1,840,363 | 856,736 |
| Chr11 | 1,894,832 | 815,114 |
| Chr12 | 1,502,345 | 649,471 |
| Chr13 | 3,648,208 | 1,351,511 |
| Chr14 | 2,828,077 | 1,122,063 |
| Chr15 | 2,743,392 | 1,095,268 |
| Chr16 | 1,788,429 | 728,729 |
| Chr17 | 1,544,662 | 562,261 |
| Chr18 | 1,368,878 | 562,261 |
| Whole genome | 46,483,626 | 18,875,672 |

*Chr, chromosome number; QC, quality control. The QC condition was MAF > 0.03.*

## SUMMARY OF GWAS

We conducted a GWAS on the $F_3$ population in two scenarios, i.e., the target data before and after imputations. In the scenario with experimental 50k chips data, we identified a total of 18 SNPs that surpassed the genome-wide significance level (**Figure 1A**). The most significantly associated SNP rs320409365 (*P-value* = 2.64 × $e^{-14}$) locates at 124.1Mb within a 10.6Mb region (116.1–126.7Mb) on SSC8 (**Table 3**). In the scenario with imputed sequence data, 3,236 significant SNPs were located on SSC1, 3, 6, 7, 8, 9, 10, 12, 13, and 16 (**Figure 1B**), and the most significantly SNP rs319603861 (*P-value* = 1.52 × $e^{-18}$) locates at 122.2Mb within a 21Mb region (115.5–136.5Mb) on SSC8 (**Table 4**). In addition, to validate the possibility of spurious SNPs caused by population stratification, the Q–Q plots for these GWAS were explored (**Supplementary Figure S3**). The average inflation factors (λ) of the GWAS were 1.17 and 1.2 in the two scenarios, respectively. Indicating that population structures were properly corrected.

## GENETIC DIFFERENTIATION SCORES

Fst were estimated to determine the extent of population differentiation between the affected and unaffected pigs. We identified a total of 26 SNPs beyond the empirical threshold on SSC8 (**Figure 2B**); the strongest genetic differentiation loci rs320409365 (Fst = 0.535) locates at 124.1Mb within a 16.6Mb region (116.1–132.7Mb), indicating the affected pigs and the unaffected pigs had a large genetic differentiation in this interval. All the SNPs beyond the empirical threshold in this interval are shown in **Table 5**.

## FINE MAPPING ON SSC8 USING LDLA AND GENETIC DIFFERENTIATION ANALYSES IN THE $F_3$ POPULATION

To further narrow down the confidence interval of SNPs SSC8 for scrotal hernia, we perform linkage and linkage disequilibrium (LDLA) for scrotal hernia on SSC8. The LDLA results showed the strongest association SNP was rs330263452 (*P*-value = 1.58 × $10^{-17}$); the most likely confidence interval of

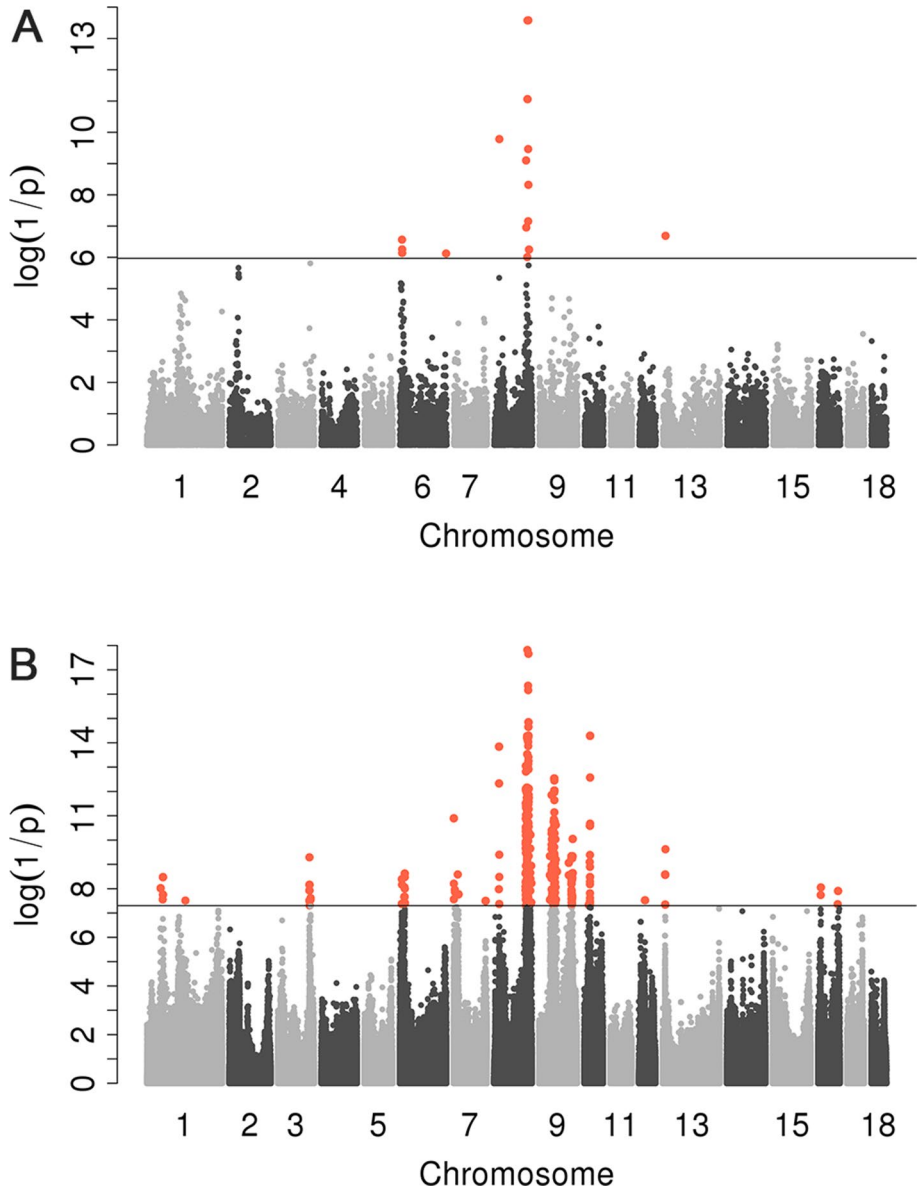


**FIGURE 1 |** Manhattan plots for scrotal hernia with data before imputation and after imputation. log10 (1/P) are shown for all qualified SNPs, which were plotted against genomic position. In Manhattan plot **(A)**, black solid line indicates the 5% genome-wide significant threshold. In Manhattan **(B)**, the black line indicated the significance threshold [–log10(5E–08)]. All SNPs surpassing the genome-wide threshold are highlighted in pink.

the QTL was approximately 3Mb (121–123.99Mb), based on the LOD drop off 2 (**Figure 2A**). We herein concluded a common QTL region located on SSC8 between 121.02 and 123.99Mb mapped by LDLA and genetic differentiation analysis.

## HAPLOTYPE SHARING ANALYSIS WITHIN THE CONFIDENCE INTERVAL

The result of haplotype sharing analysis on $F_3$ population was showed on **Figure 3B**. To put the result in detail, 15 of 18th affected pigs shared two types of haplotype in this 2.97Mb region flanked by markers rs318390967 and rs81404172. Those two shared haplotypes were associated with pig scrotal hernia and presumably $Q_1$-bearing and $Q_2$-bearing haplotypes, respectively. Further investigation revealed 27 of 228 unaffected pigs also carried the $Q_1$ or $Q_2$ haplotype. To test the risk ration and significance of individual carried Q haplotype, we summarized the number of affected pigs and unaffected pigs who carried and uncarried Q haplotype and conducted chi-square test with them (chi-square test *P*-value = $8.46 \times E^{-15}$). This result is indicative of that the hypothesized Q haplotype was involved in the occurrence of scrotal hernia in pigs. Next, we tried to identified whether there is the same sharing susceptibility haplotype existed in $F_2$ affected

**TABLE 3 |** Description of the most significant 13 SNPs associated with scrotal hernia on chromosome 8 in $F_3$ population with the 50k data.

| Chr | ps | rs | P_wald |
|---|---|---|---|
| 8 | 124,136,332 | rs320409365 | 2.62E−14 |
| 8 | 121,414,739 | rs333147082 | 2.64E−14 |
| 8 | 121,443,468 | rs81404013 | 8.72E−12 |
| 8 | 121,025,652 | rs318390967 | 8.72E−12 |
| 8 | 123,546,433 | rs334430596 | 3.45E−10 |
| 8 | 116,106,612 | rs329921419 | 7.94E−10 |
| 8 | 124,435,610 | rs81306859 | 4.79E−09 |
| 8 | 123,575,503 | rs327837715 | 7.01E−08 |
| 8 | 116,743,649 | rs81284684 | 1.11E−07 |
| 8 | 126,744,562 | rs81404481 | 5.65E−07 |
| 8 | 126,706,775 | rs339470982 | 5.65E−07 |
| 8 | 120,387,726 | rs81403944 | 9.81E−07 |
| 8 | 124,688,011 | rs345674547 | 1.80E−06 |

*Chr, chromosome number; rs, SNP IDs and SNPs that do not possess ID were named after Chr_ps, by the author; ps, base positions on the chromosome; P_wald, P-value from the Wald test.*

**TABLE 4 |** Description of the most significant 20 SNPs associated with scrotal hernia on chromosome 8 in $F_3$ population with the data after imputation.

| Chr | ps | rs | p_wald |
|---|---|---|---|
| 8 | 122,211,833 | rs319603861 | 1.53E−18 |
| 8 | 125,809,848 | rs337122565 | 2.18E−18 |
| 8 | 125,809,964 | rs339744702 | 2.18E−18 |
| 8 | 125,809,992 | rs318592275 | 2.18E−18 |
| 8 | 124,541,337 | rs695816095 | 4.51E−17 |
| 8 | 125,085,997 | rs344335641 | 6.84E−17 |
| 8 | 125,845,805 | rs321787225 | 1.42E−15 |
| 8 | 125,845,868 | rs332303403 | 1.42E−15 |
| 8 | 126,802,357 | 8_126802357 | 1.44E−15 |
| 8 | 125,810,544 | rs340831415 | 2.22E−15 |
| 8 | 125,818,781 | rs324505118 | 2.22E−15 |
| 8 | 125,818,811 | rs324505118 | 2.22E−15 |
| 8 | 125,810,459 | rs327695191 | 5.07E−15 |
| 8 | 125,811,139 | rs336507639 | 5.07E−15 |
| 8 | 125,812,250 | rs81404378 | 5.07E−15 |
| 8 | 125,813,294 | rs337489662 | 5.07E−15 |
| 8 | 125,817,665 | rs321431992 | 5.07E−15 |
| 8 | 125,817,706 | rs341020016 | 5.07E−15 |
| 8 | 120,993,480 | rs790867883 | 5.29E−15 |

*Chr, chromosome number; rs, SNP IDs and SNPs that do not possess ID were named after Chr_ps, by the author; ps, base positions on the chromosome; P_wald, P-value from the Wald test.*

individuals and trace this susceptibility haplotype back to the $F_1$ and $F_0$ generations. The result showed that 13 of the 19 affected pigs in the $F_2$ population also carried $Q_1$ or $Q_2$ haplotype flanked by markers rs81275702 and rs81404172 (**Figure 4**), and another carried other types of haplotypes. According to the pedigree (**Table 6**), we also found that the parents of those 13 affected pigs also carried $Q_1$ or $Q_2$ haplotype in the same region, while the other 5 parents with other types of haplotypes individuals did not. Most of all, we found $Q_1$ and $Q_2$ haplotypes were come from of one White Duroc boars ($F_0$-73) and three Chinese Erhualian sows ($F_0$-74, $F_0$-94, $F_0$-124) when we traced those two haplotypes to the $F_0$ generation, respectively. Therefore, it is concluded that two susceptibility haplotypes underlying the SSC8 were identified for pig scrotal hernia, and there should be some important pathogenic mutations. In addition, it was worth mentioning that the significantly associated SNP rs81404013 ($P$-value = $8.72\times E^{-12}$), rs318390967 ($P$-value = $8.72\times E^{-12}$), and rs333147082 ($P$-value = $2.64\times E^{-14}$) that contained in this confidence interval have strong linkage disequilibrium extents ($r^2 > 0.9$) to each other (**Figure 3A**). However, the most significantly associated SNP rs320409365 ($P$-value = $2.62\times E^{-14}$) has a low linkage disequilibrium extents ($r^2 < 0.5$) with those three loci. We take a region flanked by markers rs341392224 and rs326688253, which contain rs320409365, as well as it's left and right two loci. Then, we count the types of haplotype in this interval and take a chi-square test with them (**Supplementary Table 1**); the result showed that haplotype CACGT ($P$-value = $1.02\times E^{-12}$) was significantly associated with scrotal hernia.

## CANDIDATE GENE *EIF4E* FOR GENOME-WIDE SIGNIFICANT QTL

The 2.95Mb region on SSC 8 in pig (Ensembl 2018) encompasses eight annotated genes (*ADH6, ADH4, ADH5, METAP1, EIF4E, TSPAN5, RAP1GDS1, STPG2*), which indicated that few genes are the most likely candidate genes that caused scrotal hernia in pigs (Zerbino et al., 2018). Of the eight genes, *EIF4E* stood out as a potential candidate based on its biochemical and physiological functions. *EIF4E* is a protein-coding gene, which regulates the expression of the Eukaryotic translation initiation factor 4E protein, and translation initiation factor 4E is regulating the expression of *MID1* gene (Pelletier et al., 1991; Jones et al., 1997). Winter et al. demonstrated that loss-of-function mutations in the *MID1* gene may cause the malformations of the ventral midline, which always lead to a series of urogenital abnormalities, such as cryptorchidism, ambiguous genitalia, hypoplastic scrotum, and umbilical and inguinal hernias (Winter et al., 2016). In addition, both top SNPs rs333147082 ($P$-value = $2.64\times e^{-14}$) and rs81404013 ($P$-value = $8.72\times e^{-12}$) located in the intron of *EIF4E* gene when we condition the strongest significantly associated SNP rs333147082; no additional association signals appeared in this loci (QTL) was detected (**Supplementary Figure S4**), which showed the additional evidence for the causality of *EIF4E* incorporating functional and conditional association studies. These results were more evidence that the *EIF4E* is the susceptibility gene for pig scrotal hernias.

## DISCUSSION

In the current study, we obtained 18,756,672 variants with 84.8% genotypic concordance rate. In the study on imputation, few researches reported the imputation accuracy from 60K to whole-genome sequence in pig, compared to most studies focused on imputation from low-density genotypes to 60k variants with correlations ranging from 0.938 to 0.992 for imputation from 3 to 60K (Cleveland and Hickey, 2013). Yan et al. showed an average genotypic concordance of 89% with imputing 60K to whole-genome sequence variants in a

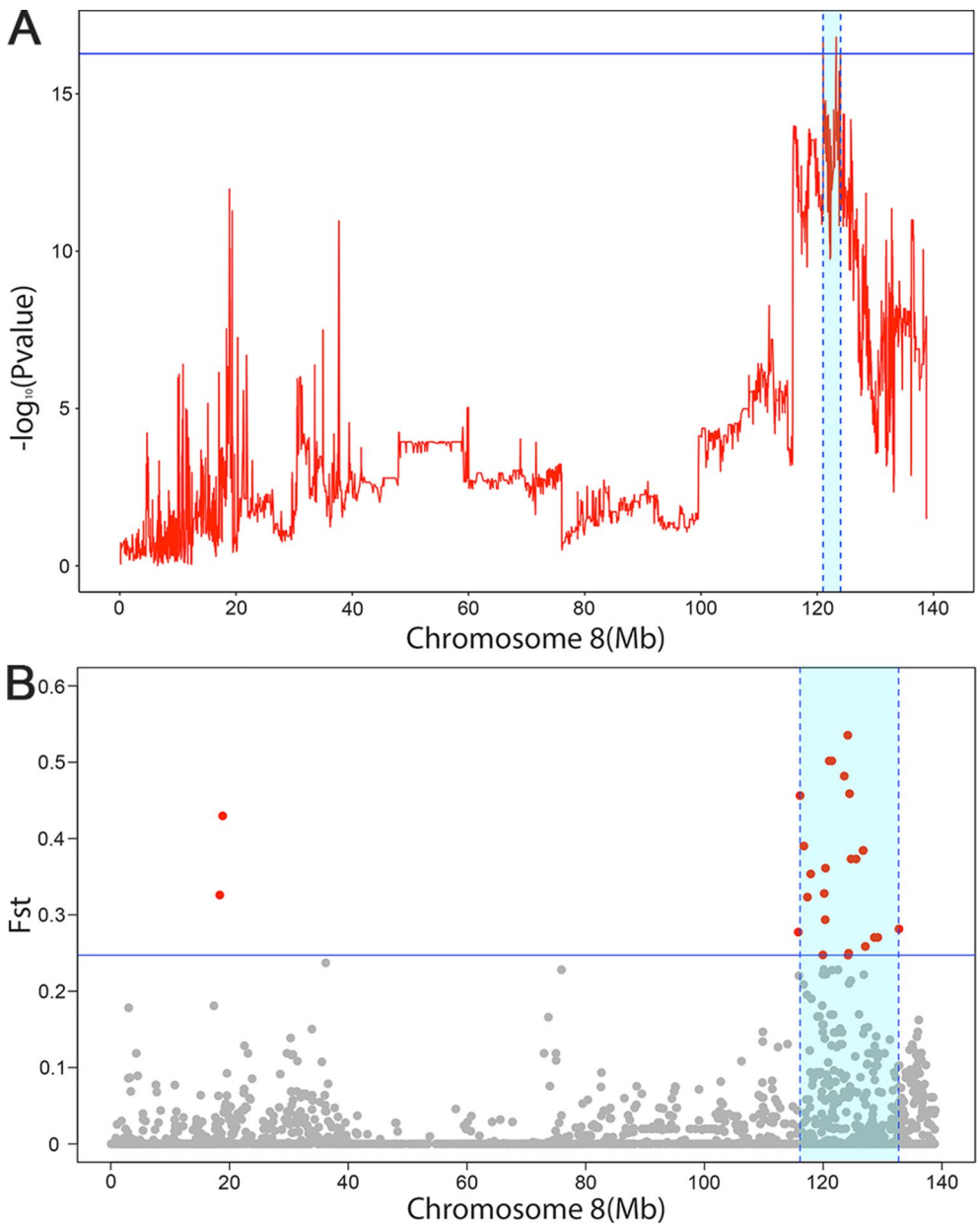


**FIGURE 2 |** The significant associated region on SSC8 in LDLA analysis **(A)** and genetic differentiation analysis **(B)**. **(A)** The y-axis shows negative $\log_{10}$ (*P*-values) from haplotype-based association study, and the x-axis indicates the SNP positions on SSC8. The red lines represent the haplotype. The horizontal line indicated the 95% of confidence interval by LOD drop off two from the most significant haplotype. **(B)** The significant associated region on SSC8 were represented as light blue. The x-axis indicates the SNP positions on SSC8, and y-axis shows Fst. The horizontal line indicated the top 1 of confidence interval. All SNPs surpassing the threshold are highlighted in pink. Region with a large genetic differentiation were represented as light blue.

large-scale swine F2 resource population (Yan et al., 2018), and Zhang et at. reported the genotypic concordance was 85.6% from 650K to whole-genome sequence variants using a stepwise imputation strategy in 1,363 Duroc pigs (Zhang et al., 2018); the genotypic concordance rate (84.8%) in our study is almost to their level. Moreover, we adopted $R^2$ to estimate imputation accuracy on account of genotypic concordance rate that is highly sensitive to MAF and is not appropriate for comparing genotypes with different MAF (Yan et al., 2018).

In the present study, $R^2$ decreased from 58 to 8% when MAF decreased from 0.1 to 0. The same trend was found in other studies (Daetwyler et al., 2014; Yan et al., 2018). And the average correlation between best-guess and true variants reached with an average of 71% after we delete sites where $R^2$ is equal to 0 and MAF is less than 0.03. Besides, Yan et al. showed that the average correlation is lower than the genotypic concordance rate, which was consistent with our result in this study (Yan et al., 2017). In addition, there are many other factors that affect the accuracy of

**TABLE 5 |** Genome-wide loci beyond the empirical threshold on chromosome 8 for pig inguinal/scrotal hernias identified by genetic differentiation analysis.

| Chr | ps | rs | Fst |
|---|---|---|---|
| 8 | 124,136,332 | rs320409365 | 0.53536008 |
| 8 | 121,025,652 | rs318390967 | 0.501778292 |
| 8 | 121,443,468 | rs81404013 | 0.501778292 |
| 8 | 123,546,433 | rs334430596 | 0.481862641 |
| 8 | 124,435,610 | rs81306859 | 0.458791183 |
| 8 | 116,106,612 | rs329921419 | 0.456265642 |
| 8 | 116,743,649 | rs81284684 | 0.390162465 |
| 8 | 126,706,775 | rs339470982 | 0.384398437 |
| 8 | 126,744,562 | rs81404481 | 0.384398437 |
| 8 | 125,530,778 | rs334269805 | 0.37324505 |
| 8 | 124,688,011 | rs345674547 | 0.37324505 |
| 8 | 120,387,726 | rs81403944 | 0.361238378 |
| 8 | 117,897,490 | rs336417589 | 0.353602403 |
| 8 | 120,167,202 | rs81403910 | 0.328123759 |
| 8 | 117,335,274 | rs81324515 | 0.323287852 |
| 8 | 120,335,533 | rs81403964 | 0.293693666 |
| 8 | 132,760,090 | rs81323639 | 0.281592392 |
| 8 | 115,799,722 | rs81307505 | 0.277478219 |
| 8 | 129,198,702 | rs329385027 | 0.270487326 |
| 8 | 128,613,004 | rs336466493 | 0.270487326 |
| 8 | 127,090,631 | rs81317149 | 0.258692328 |
| 8 | 124,278,247 | rs332687320 | 0.249736644 |
| 8 | 119,943,759 | rs81330386 | 0.247590093 |
| 8 | 124,189,324 | rs81340120 | 0.247278921 |

*Chr, chromosome number; rs, SNP IDs and SNPs that do not possess ID were named after Chr_ps, by the author; ps, base positions on the chromosome; Fst, the genetic differentiation scores.*

imputation, such as the relationships between target panel and reference (van Binsbergen et al., 2014) and LD and reference size (van Binsbergen et al., 2014). Here we sequenced 19 ancestors of F3 to ensure our imputation reliability. Overall, imputation accuracy can be affected by different aspects, and high accuracy of imputation will lead to a reliable GWAS.

GWAS has become an exceedingly effective and widely used approach in identification of genetic variants associated with common diseases or complex traits since the first application of GWAS research was performed successfully in 2005 by Klein et al. (2005). Previously, by performing haplotype-based GWAS in $F_2$ population for scrotal hernia using Porcine SNP60 BeadChip, 108 chromosome-wise significance SNPs were identified to be associated with scrotal hernia; however, there was no marker surpassed the genome-wide significance level. The feasible reasons for the low detection power in this study was probably the low incidence and penetrance rate in $F_2$ population. But the most possible reason is the intricacy molecular genetic mechanism of scrotal hernia. So far, many international teams have identified the susceptibility loci of scrotal hernia on almost all chromosomes. Complex interactions between environmental factors and susceptibility alleles of multiple genes are the most normal process resulting in such a complex genetic background diseases. As a complex genetic defect, the polygene model may be the main pathogenesis, under polygene model that lots of susceptibility genes cause a change for disease. Therefore, whether a certain mutation is not directly related to scrotal hernia, but does have a role in the occurrence of it, this is why single-marker GWAS can't detect any significant signal in the $F_2$ population. Therefore, we generate a particular hernia population which was mated with full-sibs or half-sib of the affected individuals. The incidences of scrotal hernia will increase significantly in this population. Next, we will systematically describe the feasibility of our idea.

In the current study, we first designed a specially $F_3$ population to increase the incidence and penetrance rate by crossing full-sibs or half-sib of the affected individuals in the $F_2$ population. Statistics manifested that prevalence of scrotal hernia in the specially designed $F_3$ population was 3.6 times and 2.4 times higher than the ordinary $F_3$ population and $F_2$ population, respectively, indicating the $F_3$ specially designed population is completely successful in increasing the incidence rate of scrotal hernia. Most importantly, in the $F_2$ population, the frequency of a mutation associated with scrotal hernia will be greatly increased in $F_3$ specially designed population, as the health full-sibs or half-sib of the affected individuals in the $F_2$ population also have the mutant sites, which will pass on to the $F_3$ population.

As we predicted, 13 SNPs were located on SSC8 between 116.1 and 126.7Mb surpassed the genome-wide significance level after we conducted a GWAS on the specially designed $F_3$ population with experimental 50-k chip data, and this QTL must have come from $F_2$, which overlaps with a region previously identified by Sevillano et al. (2015). The basic principle of single-marker GWAS was to test association between phenotypes and genotypes. Normally, this association was indirect correlation as the causative mutation was not included in the study locus. Potentially, significant signals could be missed in a GWAS analysis if there were low LDs among paired markers. To improve the LD between markers, we performed imputation analysis by increasing the marker density in the study population using 109 sequenced data as reference panel. Consequently, we obtained 18,756,672 variants with relatively high imputation accuracy (average CR = 84.8%). After performing the whole-genome association study with sequence data, 3,252 significant SNPs reached the significant level. Three regions located on SSC3, SSC8, and SSC10 were similar to corresponding interval previously identified by Sevillano et al. (2015), especially the region on SSC8 between 115.6 and 136.5Mb overlaps a region they previously identified. To our knowledge, it is the first time that the other eight QTL regions identified on SSC1, 6, 7, 9, 13, and 16 are found to be associated with scrotal hernia, although some studies have reported that different regions on these chromosomes harbor QTL for scrotal hernia.

According to our original intention, we identified 13 SNP loci significantly associated with scrotal hernia on chromosome 8 through GWAS analysis with the specially designed $F_3$ population. In the subsequent analysis, we divided the affected individuals and unaffected individuals into two independent groups and calculated the genetic differentiation index to verify that there is genetic differentiation on SSC8. The result showed that a strong genetic differentiation signal located on rs320409365 (Fst = 0.535) within a 16.6Mb region (116.1–132.7Mb) on SSC 8 was detected. This result indicated that the affected pigs and the unaffected pigs had a greater genetic differentiation in this confidence interval. Moreover, there is a high coincidence of the top SNPs detected through genetic differentiation analyses and GWAS.

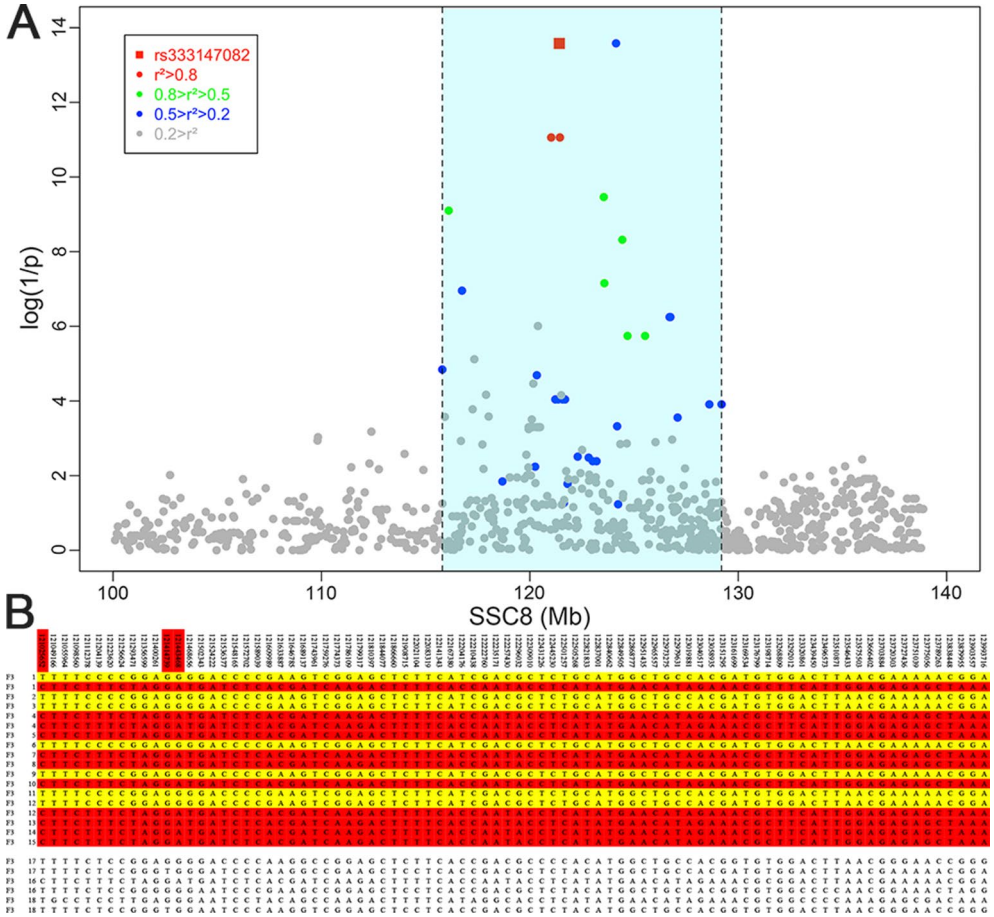


**FIGURE 3 |** Fine mapping of the target region by the haplotype sharing analysis in the $F_3$ population. **(A)** Regional association plot of SNPs in linkage disequilibrium with rs333147082. The colored diamonds indicate different linkage disequilibrium (LD) levels between rs333147082 and other SNPs. The light blue region indicates the interval which SNPs and rs333147082 with LD greater than 0.2. **(B)** Haplotypes of the target region between 121 ~ 123.99 Mb on chromosome 8 are shown. Golden diamonds and red diamonds represent the $Q_1$ and $Q_2$ haplotypes with affected pigs, respectively. The last six lines indicate that three affected pigs who carried other types of haplotypes.

Additionally, in consideration of single-marker GWAS, it was hard to properly estimate the confidence interval of the detected QTL, as LD varied severely among nearby SNPs while haplotypes have stable LD than SNPs. Thus, we conducted haplotype-based LDLA analysis, by simultaneously taking advantage of recent and ancestral recombination events to increase the efficiency and detect confidence interval. The LDLA results showed that the SNP with the strongest association at the locus was rs330263452 ($P$-value = $1.58 \times 10^{-17}$), and the most likely confidence intervals around the 121–123.99Mb region on SSC8. Furthermore, we found out that the confidence intervals mapped by LDLA contained within the region mapped by genetic differentiation analysis. We narrow the confidence interval to 2.99Mb by picking up the intersection of those two intervals for further analysis.

Lastly, we identified two susceptibility haplotypes underlying the SSC8 associated with scrotal hernia after performed a haplotype sharing analysis, and those two haplotypes were from one White Duroc boar ($F_0$-73) and three Chinese Erhualian sows ($F_0$-74, $F_0$-94, $F_0$-124), respectively. It is incomprehensible that the White Duroc boar ($F_0$-73) carried the susceptibility haplotype, but it was unaffected. Actually, whether a certain mutation is not directly related to scrotal hernia as we explained earlier. Similarly, 163 of 497 unaffected pigs in the $F_2$ population also carried the susceptibility haplotypes, echoing the result that there was no significant signal when we performed GWAS in the $F_2$ population. When we merge the $F_2$ and $F_3$ populations and then conducted chi-square test with them (chi-square test $P$-value = $8.32\times E^{-11}$), this result is also indicative of that the hypothesized Q haplotype was involved in the occurrence of scrotal hernia in pigs. In addition to discovering two susceptibility haplotypes, we further found that there are nine annotated genes in this 2.95Mb interval in total, and the *EIF4E* was selected as potential candidate gene based on its biochemical and physiological functions.

Although there are some crucial discoveries revealed by these studies, there are a slice of limitations to our study, such as the relatively small number of samples in the F3 population. Therefore, we herein carried out bootstrap test to verify the reliability of GWAS in this study. The result showed that there are 957 of the 1,000 GWAS

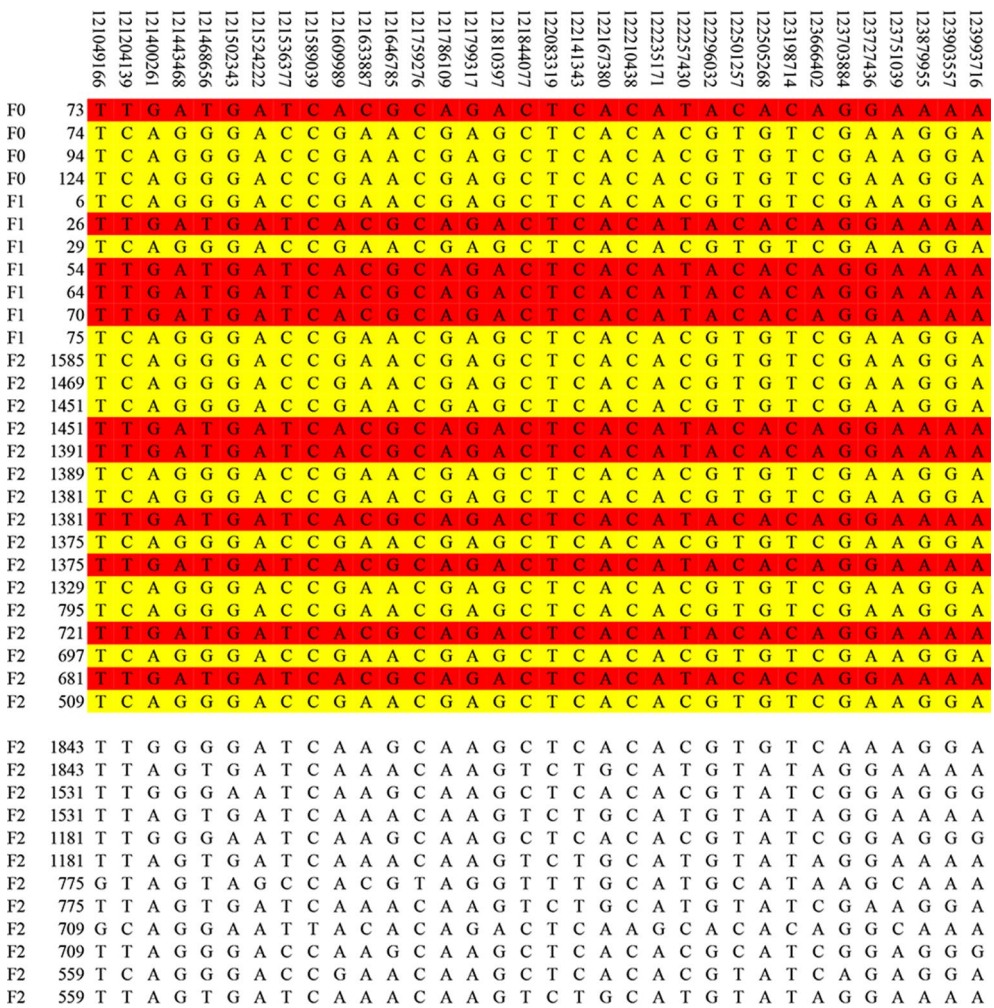


**FIGURE 4 |** The haplotype sharing analysis in the $F_2$ population. The figure showed that $Q_1$ and $Q_2$ haplotypes contained in $F_2$ affected individuals and traced back to the $F_1$ and $F_0$ generations. The last 12 lines indicate six affected pigs that carried other types of haplotypes.

**TABLE 6 |** The pedigree of $F_2$ affected individuals.

| | Parent's generation | | Grandparent's generation | | | |
|---|---|---|---|---|---|---|
| | **Male** | **Female** | **Male** | **Female** | **Male** | **Female** |
| $F_2$-1,585 | $F_1$-29 | $F_1$-46 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-58 |
| $F_2$-1,469 | $F_1$-29 | $F_1$-52 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-58 |
| $F_2$-1,451 | $F_1$-75 | $F_1$-26 | $F_0$-75 | $F_0$-94 | $F_0$-73 | $F_0$-90 |
| $F_2$-1,391 | $F_1$-29 | $F_1$-64 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-202 |
| $F_2$-1,389 | $F_1$-29 | $F_1$-64 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-202 |
| $F_2$-1,381 | $F_1$-29 | $F_1$-64 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-202 |
| $F_2$-1,375 | $F_1$-29 | $F_1$-64 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-58 |
| $F_2$-1,329 | $F_1$-35 | $F_1$-6 | $F_0$-73 | $F_0$-58 | $F_0$-73 | $F_0$-124 |
| $F_2$-795 | $F_1$-75 | $F_1$-32 | $F_0$-75 | $F_0$-94 | $F_0$-73 | $F_0$-90 |
| $F_2$-721 | $F_1$-49 | $F_1$-70 | $F_0$-75 | $F_0$-94 | $F_0$-73 | $F_0$-202 |
| $F_2$-697 | $F_1$-29 | $F_1$-46 | $F_0$-75 | $F_0$-74 | $F_0$-73 | $F_0$-58 |
| $F_2$-681 | $F_1$-49 | $F_1$-54 | $F_0$-75 | $F_0$-94 | $F_0$-73 | $F_0$-58 |
| $F_2$-559 | $F_1$-3 | $F_1$-36 | $F_0$-73 | $F_0$-124 | $F_0$-75 | $F_0$-74 |
| $F_2$-509 | $F_1$-35 | $F_1$-6 | $F_0$-73 | $F_0$-58 | $F_0$-73 | $F_0$-124 |

that were detected significant signals in the 116–126Mb interval on chromosome 8, which indicated that the fluctuation in the number of affected and unaffected individuals has no effect on GWAS (FDR < 0.05). Therefore, the significant signals obtained in our GWA study were not accidental but were caused by differences in the genomes of affected and unaffected individuals, which were reliable.

## CONCLUSION

In summary, in the first place, we discovered a major quantitative trait loci (QTL) for pig scrotal hernia on chromosome 8 in an $F_3$ specially designed population using GWAS. There is one more point: two susceptibility haplotypes ($Q_1$ and $Q_2$) flanked by markers rs81275702 and rs81404172 and one potential causal gene underlying the SSC8 were identified through a series of methods including genetic differentiation analysis, LDLA, and haplotype sharing analysis. Last but not the least, we explain why many international research teams do not have a high repeatability of the results of scrotal hernia research, and some research studies haven't even found any associated locus with scrotal hernia. Further studies will be devoted to confirming the detected haplotype and gene in outbred populations.

## DATA AVAILABILITY

The genotypic data of 246 F3 individuals as well as the phenotype of scrotal hernia for this study can be found in the figshare Digital Repository (https://figshare.com/s/d661962aa6c0740caeab), and the genotypic data of 516 F2 individuals analyzed for this study can be found in the Dryad Digital Repository (https://doi.org/10.5061/dryad.7kn7r) (Ma et al., 2013). The raw reads of the whole-genome sequence can be found from the NCBI sequence read archive (SRA) under the accession codes SRA065461 and SRP159212 (Ai et al., 2015; Yan et al., 2018).

## ETHICS STATEMENT

This study was approved by the ethics committee of Jiangxi Agricultural University. All procedures including experimental animals established and tissue collection were performed in accordance with the guidelines approved by the Ministry of Agriculture of China.

## AUTHOR CONTRIBUTIONS

LH and ZZ conceived and designed the experiments. WX, GY, and TH analyzed the data. DC and SX contributed materials and analysis tools. WX, ZZ, and LH wrote the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was supported by National Natural Science Foundation of China (31760656) and Guangdong Sail Plan Introduction of Innovative and Entrepreneurship Research Team Program (No. 2016YT03H062).

## ACKNOWLEDGMENTS

We are grateful to all members who participated in this study from the State Key Laboratory for Pig Genetic Improvement and Production Technology.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00890/full#supplementary-material

## REFERENCES

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* 47 (3), 217–225. doi: 10.1038/ng.3199

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12 (12), 1805–1814. doi: 10.1101/gr.631202

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

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

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81 (5), 1084–1097. doi: 10.1086/521987

Browning, S. R., and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12 (10), 703–714. doi: 10.1038/nrg3054

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

Clarnette, T. D., and Hutson, J. M. (1997). Is the ascending testis actually 'stationary'? Normal elongation of the spermatic cord is prevented by a fibrous remnant of the processus vaginalis. *Pediatr. Surg. Int.* 12 (2/3), 155–157. doi: 10.1007/BF01349987

Clarnette, T. D., Lam, S. K., and Hutson, J. M. (1998). Ventriculo-peritoneal shunts in children reveal the natural history of closure of the processus vaginalis. *J. Pediatr. Surg.* 33 (3), 413–416. doi: 10.1016/S0022-3468(98)90080-X

Cleveland, M. A., and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. *J. Anim. Sci.* 91 (8), 3583–3592. doi: 10.2527/jas.2013-6270

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brondum, R. F., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46 (8), 858–865. doi: 10.1038/ng.3034

Ding, N. S., Mao, H. R., Guo, Y. M., Ren, J., Xiao, S. J., Wu, G. Z., et al. (2009). A genome-wide scan reveals candidate susceptibility loci for pig hernias in an intercross between White Duroc and Erhualian. *J. Anim. Sci.* 87 (8), 2469–2474. doi: 10.2527/jas.2008-1601

Du, Z. Q., Zhao, X., Vukasinovic, N., Rodriguez, F., Clutter, A. C., and Rothschild, M. F. (2009). Association and haplotype analyses of positional candidate genes in five genomic regions linked to scrotal hernia in commercial pig lines. *PLoS One* 4 (3), e4837. doi: 10.1371/journal.pone.0004837

Efron., B., and Tibshirani., R. J. (1993). *An introduction to the bootstrap*. Washington, DC: Chapman & Hall/CRC.

Grindflek, E., Moe, M., Taubert, H., Simianer, H., Lien, S., and Moen, T. (2006). Genome-wide linkage analysis of inguinal hernia in pigs using affected sib pairs. *BMC Genet.* 7, 25. doi: 10.1186/1471-2156-7-25

Guo, Y., Mao, H., Ren, J., Yan, X., Duan, Y., Yang, G., et al. (2009). A linkage map of the porcine genome from a large-scale White Duroc x Erhualian resource population and evaluation of factors affecting recombination rates. *Anim. Genet.* 40 (1), 47–52. doi: 10.1111/j.1365-2052.2008.01802.x

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., et al. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11, 724. doi: 10.1186/1471-2164-11-724

Jones, R. M., MacDonald, M. E., Branda, J., Altherr, M. R., Louis, D. N., and Schmidt, E. V. (1997). Assignment of the human gene encoding eukaryotic initiation factor 4E (EIF4E) to the region q21-25 on chromosome 4. *Somat. Cell Mol. Genet.* 23 (3), 221–223. doi: 10.1007/BF02721373

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. A., Baurain, D., et al. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat. Genet.* 43 (5), 405–413. doi: 10.1038/ng.814

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720), 385–389. doi: 10.1126/science.1109557

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Ma, J., Yang, J., Zhou, L., Zhang, Z., Ma, H., Xie, X., et al. (2013). Genome-wide association study of meat quality traits in a White DurocxErhualian F2 intercross and Chinese Sutai pigs. *PLoS One* 8 (5), e64047. doi: 10.1371/journal.pone.0064047

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

Mikami, H., and Fredeen, H. T. (1979). A genetic study of cryptorchidism and scrotal hernia in pigs. *Can. J. Genet. Cytol.* 21 (1), 9–19. doi: 10.1139/g79-002

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

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* 32 (4), 381–385. doi: 10.1002/gepi.20303

Pearson, T. A., and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA* 299 (11), 1335–1344. doi: 10.1001/jama.299.11.1335

Pelletier, J., Brook, J. D., and Housman, D. E. (1991). Assignment of two of the translation initiation factor-4E (EIF4EL1 and EIF4EL2) genes to human chromosomes 4 and 20. *Genomics* 10 (4), 1079–1082. doi: 10.1016/0888-7543(91)90203-Q

PK, C. (2006). Congenital defects in pigs: 1. hernias and ridglings.

Ramos, A. M., Crooijmans, R. P., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 4 (8), e6524. doi: 10.1371/journal.pone.0006524

Sevillano, C. A., Lopes, M. S., Harlizius, B., Hanenberg, E. H., Knol, E. F., and Bastiaansen, J. W. (2015). Genome-wide association study using deregressed breeding values for cryptorchidism and scrotal/inguinal hernia in two pig lines. *Genet. Sel. Evol.* 47, 18. doi: 10.1186/s12711-015-0096-6

Su, Y., Ruan, G. R., Long, Y., Yang, B., Zhang, Z. Y., Deng, W. Y., et al. (2014). Genome-wide association study reveals candidate susceptibility loci for pig scrotal hernia using both F2 intercross and outbred populations. *Sci. Agric. Sin.* 47 (14), 2872–2880. doi: 10.3864/j.issn.0578-1752.2014.14.017

Thaller, G., Dempfle, L., and Hoeschele, I. (1996). Maximum likelihood analysis of rare binary traits under different modes of inheritance. *Genetics* 143 (4), 1819–1829.

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

Walters, J. R. (2010). Have we forgotten about inherited disease? AGBU Pig Genetics Workshop –October 2010.

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* 91 (2), 238–251. doi: 10.1016/j.ajhg.2012.06.013

Winter, J., Basilicata, M. F., Stemmler, M. P., and Krauss, S. (2016). The MID1 protein is a central player during development and in disease. *Front. Biosci. (Landmark Ed.)* 21, 664–682. doi: 10.2741/4413

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15 (4), 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

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

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

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46 (D1), D754–D761. doi: 10.1093/nar/gkx1098

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

Zhang, M., Huang, T., Huang, X., Tong, X., Chen, J., Yang, B., et al. (2019). New insights into host adaptation to swine respiratory disease revealed by genetic differentiation and RNA sequencing analyses. *Evol. Appl.* 12 (3), 535–548. doi: 10.1111/eva.12737

Zhang, Z., Guillaume, F., Sartelet, A., Charlier, C., Georges, M., Farnir, F., et al. (2012). Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. *Bioinformatics* 28 (19), 2467–2473. doi: 10.1093/bioinformatics/bts348

Zhao, X., Du, Z. Q., Vukasinovic, N., Rodriguez, F., Clutter, A. C., and Rothschild, M. F. (2009). Association of HOXA10, ZFPM2, and MMP2 genes with scrotal hernias evaluated via biological candidate gene analyses in pigs. *Am. J. Vet. Res.* 70 (8), 1006–1012. doi: 10.2460/ajvr.70.8.1006

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–824. doi: 10.1038/ng.2310

**Conflict of Interest Statement:** 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: 09 October 2019
doi: 10.3389/fgene.2019.00968

# Novel lncRNA lncFAM200B: Molecular Characteristics and Effects of Genetic Variants on Promoter Activity and Cattle Body Measurement Traits

*Sihuan Zhang[1], Zihong Kang[1], Xiaomei Sun[1,2], Xiukai Cao[1], Chuanying Pan[1], Ruihua Dang[1], Chuzhao Lei[1], Hong Chen[1] and Xianyong Lan[1]**

*[1] College of Animal Science and Technology, Northwest A&F University, Yangling, China, [2] College of Animal Science and Technology, Yangzhou University, Yangzhou, China*

Skeletal muscle is one of the three major muscle types in an organism and has key roles in the motor system, metabolism, and homeostasis. RNA-Seq analysis showed that novel lncRNA, *lncFAM200B*, was differentially expressed in embryonic, neonatal, and adult cattle skeletal muscles. The main aim of this study was to investigate the molecular and expression characteristics of *lncFAM200B* along with its crucial genetic variations. Our results showed that bovine *lncFAM200B* was a 472 nucleotide (nt) non-coding RNA containing two exons. The transcription factor binding site prediction analysis found that *lncFAM200B* promoter region was enriched with SP1 transcription factor, which promotes the binding of myogenic regulatory factor MyoD and DNA sequence. The mRNA expression analysis showed that *lncFAM200B* was differentially expressed in embryonic, neonatal, adult bovine muscle tissues, and the *lncFAM200B* expression trend positively correlated with that of *MyoG* and *Myf5* in myoblast proliferation and differential stages. To identify the promoter active region of *lncFAM200B*, we constructed promoter luciferase reporter gene vector pGL3-Basic plasmids containing *lncFAM200B* promoter sequences and transfected them into 293T, C2C12, and 3T3-L1 cells. Our results suggested that *lncFAM200B* promoter active region was from −403 to −139 (264 nt) of its transcription start site, covering 6 SP1 potential binding sites. Furthermore, we found a novel C-T variation, named as SNP2 (ERZ990081 in European Variation Archive) in the promoter active region, which was linked to the nearby SNP1 (rs456951291 in Ensembl database). The genotypes of SNP1 and combined genotypes of SNP1 and SNP2 were significantly associated with Jinnan cattle hip height. The luciferase activity analysis found that the SNP1-SNP2 haplotype CC had the highest luciferase activity, which was consistent with the association analysis result that the combined genotype CC-CC carriers had the highest hip height in Jinnan cattle. In conclusion, our data showed that *lncFAM200B* is a positive regulator of muscle development and that SNP1 and SNP2 could be used as genetic markers for marker-assisted selection (MAS) breeding of beef cattle.

**Keywords: bovine, lncFAM200B, muscle development, promoter, body measurement traits**

OPEN ACCESS

***Edited by:***
*Marco Milanesi,*
*São Paulo State University, Brazil*

***Reviewed by:***
*Xiaozhu Wang,*
*Auburn University, United States*
*Lisui Bao,*
*University of Chicago, United States*

***Correspondence:**
*Xianyong Lan*
*lanxianyong79@126.com*

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

***Received:*** *15 May 2019*
***Accepted:*** *10 September 2019*
***Published:*** *09 October 2019*

***Citation:***
*Zhang S, Kang Z, Sun X, Cao X, Pan C, Dang R, Lei C, Chen H and Lan X (2019) Novel lncRNA lncFAM200B: Molecular Characteristics and Effects of Genetic Variants on Promoter Activity and Cattle Body Measurement Traits. Front. Genet. 10:968. doi: 10.3389/fgene.2019.00968*

## INTRODUCTION

Long non-coding RNA (lncRNA) is an important class of non-coding RNAs (ncRNAs), which are involved in a variety of biological processes. LncRNAs are usually greater than 200 nucleotide (nt) in length, mostly were transcribed by RNA polymerase II, and some were transcribed by RNA polymerase III. Similar to mRNAs, the expression of lncRNAs have obviously temporal (the same tissue on different development stages) as well as the spatial (different tissues) specificity. LncRNA gene has its own promoter, which can be recognized by specific transcription factors. In the last decade, lncRNAs have been showed to have multiple functions in many developmental processes, such as regulating gene expression by transcriptional, post-transcriptional, or epigenetic regulation (Yan et al., 2017; Fernandes et al., 2019). Besides, lncRNAs can serve as the sponges for miRNAs to relieve the repression of miRNAs on their target genes (Sun et al., 2016). Although the biological functions of lncRNAs are very important, their sequence conservation is low among species. Thus, it is important to understand the role of novel lncRNAs in various biological processes in different species.

Skeletal muscles account for about 40% of human body weight, which are not only the dynamic part of the motor system but also play a key role in organism metabolism and homeostasis (Li et al., 2018). Skeletal muscles are composed primarily of multinucleated myotubes, which were originally derived from myogenic progenitor cells (MPCs). MPCs are destined to become myoblasts, which subsequently turn into myotubes after proliferation, differentiation, and fusion (Li et al., 2018). This process is regulated by a variety of transcription factors and epigenetic regulators such as the myogenic regulatory factors myogenic differentiation 1 (MyoD), myogenin (MyoG), myogenic factor 5 (Myf5), and myosin heavy chain 3 (MYH3) (Bharathy et al., 2013). Recently, with the rapid development of sequencing technology, an increasing number of studies found that lncRNA played a crucial role in the development of muscle (Yu et al., 2017; Zhu et al., 2017; Li et al., 2018). In cattle, the lncRNA sequencing showed that lncRNAs were crucial in muscle development (Billerey et al., 2014; Sun et al., 2016; Liu et al., 2017). Although the functions of some lncRNAs such as *lncMD*, *lncYYW*, and *lnc133b* in bovine muscle development have been identified, the roles of numerous lncRNAs are still mysteries to be explored (Sun et al., 2016; Jin et al., 2017; Yue et al., 2017).

Muscle development is one of the main factors that affect cattle growth, and thus, ultimately influences the production economic benefits. Thus, this issue has attracted huge attention in the beef cattle breeding industry. Nowadays, marker-assisted selection (MAS) is a rapid and efficient breeding method, which is based on crucial genetic variation markers (Cui et al., 2018; Chen et al., 2019). Thus, finding muscle development associated genetic variation markers is very important for beef cattle MAS breeding. Given the important role of lncRNA, we think that it would be feasible to screen genetic variations in the muscle development associated lncRNAs region.

Sun et al. (2016) using Ribo-Zero RNA-Seq identified the lncRNA landscape of bovine embryonic, neonatal, and adult skeletal muscles. Within these three developmental stages, 401 differentially expressed lncRNAs were revealed, which included *lncMD* and some new lncRNAs (Sun et al., 2016). In these newly identified lncRNAs, NONBTAT022788 was mapped to the first intron and the second exon (sequence identity is 100%) of *Bos taurus FAM200B* gene (NCBI Reference Sequence: AC_000163.1), thus we aptly renamed it as *lncFAM200B*. In this study, we focused on *lncFAM200B* as it was differentially expressed in bovine embryonic, neonatal, and adult skeletal muscle [the fragments per kilobase of exon per million fragments mapped (FPKM) of *lncFAM200B* were 15.72, 0.41, and 5.73, respectively]. Based on the RNA-Seq results, we speculated that *lncFAM200B* probably plays an important role in the development of bovine skeletal muscle.

Therefore, in this study, we investigated the sequence and expression characteristics of bovine *lncFAM200B* and further, we identified the functional genetic variations in *lncFAM200B* gene. These results would lay the foundation for the function research of *lncFAM200B* and provide scientific data for beef cattle breeding.

## MATERIALS AND METHODS

All experiments in this study were approved by the Faculty Animal Policy and Welfare Committee of Northwest A&F University (no.NWAFAC1008). The care and use of experimental animals is in full compliance with local animal welfare laws, guidelines, and policies.

### Animal Tissue Samples Collection

To explore the expression profile of *lncFAM200B*, multiple tissue samples from Qinchuan steers at three different developmental stages: embryos of about 3 months old, newborns within 1 week, and adults of about 24 months old were collected from Shaanxi Kingbull Livestock Co., Ltd. (Baoji, China). For sampling at each of the developmental stages, three individuals were used. For each neonatal and adult individual, seven types of tissue samples were collected (heart, liver, spleen, lung, kidney, skeletal muscle, and fat tissue). For embryonic stage, only six kinds of tissue samples were collected (without fat). All samples were frozen immediately in liquid nitrogen and stored at −80°C.

### Total RNA Isolation, cDNA Synthesis, and RACE Experiments

Total RNA was isolated from samples using TRIzol reagent (TaKaRa, Dalian, China). The quality of total RNA was evaluated by 1% agarose gel electrophoresis and NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Then PrimeScript™ RT reagent Kit with gDNA Eraser (TaKaRa, Dalian, China) was used to synthesize complementary DNA (cDNA), which was used as template for quantitative reverse-transcription PCR (qRT-PCR) or full-length amplification of *lncFAM200B*.

Rapid amplification of cDNA ends (RACE) experiments were carried out to identify the full-length of bovine *lncFAM200B* using bovine fetus skeletal muscle cDNA as template.

The 3′ RACE was done using PrimeScript™ RT reagent Kit (TaKaRa, Dalian, China) and 3′ RACE universal primers $Q_T$, $Q_O$, and $Q_I$ as described in Scotto-Lavino et al. (2006). The 5′ RACE was done using SMARTer® RACE 5′/3′ Kit (Clontech, Palo Alto, CA, USA) according to the user manual and the previous study (Sun et al., 2016). The 3′ RACE and 5′ RACE specific primers for *lncFAM200B* were designed based on the sequence obtained from RNA-Seq (**Table 1**). Then the full-length of bovine *lncFAM200B* was obtained through sequences assembly based on the results of 3′ and 5′ RACE.

## The Sequence Features Analyses and Functional Prediction of Bovine *lncFAM200B*

The coding potential was predicted on Coding Potential Calculator (CPC) website (Kong et al., 2007). The known protein-coding genes CCAAT enhancer binding protein alpha (*C/EBPα*) and lncRNA H19 imprinted maternally expressed transcript (*H19*) were also calculated as control. NCBI-Open Reading Frame Finder (ORF Finder) was used to analyze the open reading frame (ORF) of *lncFAM200B*. The prokaryotic expression system was used to detect the protein coding ability of *lncFAM200B*. The full length of bovine *lncFAM200B* and enhanced green fluorescent protein (*EGFP*) were cloned into vitro prokaryotic expression system pET-28a vector using *Xho*I and *Hin*dIII restriction enzymes and In-Fusion® HD Cloning Kit (TaKaRa, Dalian, China) (Li et al., 2016). The miRDB (http://www.mirdb.org/) was used to predict the interacting miRNAs, and AliBaba2.1 (http://gene-regulation.com/pub/programs/alibaba2/index.html) was used to predict the transcription factors that may bind to the promoter region of *lncFAM200B*.

## Quantitative Reverse-Transcription PCR

The qRT-PCR was performed to detect the expression of *lncFAM200B* in tissues. The housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) was used as internal control. The primers for qRT-PCR were listed in **Table 1**. The qRT-PCR was performed in a Bio-Rad CFX Manager 3.1

**TABLE 1 |** Primers in this study.

| Primers | Primer sequences (5'→3') | Sizes (bp) | Purpose |
|---|---|---|---|
| qlncFAM200B-F | CCACTTCAAGGAAGTTCCA | 93 | qRT-PCR |
| qlncFAM200B-R | TTGTGTTGGTAGCTTGACTA | | |
| GAPDH-F | AAAGTGGACATCGTCGCCAT | 116 | qRT-PCR |
| GAPDH-R | CCGTTCTCTGCCTTGACTGT | | |
| MYOG-F | CCAGTACATAGAGCGCCTGC | 183 | qRT-PCR |
| MYOG-R | AGATGATCCCCTGGGTTGGG | | |
| MYOD-F | GAACACTACAGCGGCGACTC | 126 | qRT-PCR |
| MYOD-R | GCTGTAGTAAGTGCGGTCGT | | |
| MYH3-F | TGCTCATCTCACCAAGTTCC | 150 | qRT-PCR (Sun et al., 2016) |
| MYH3-R | CACTCTTCACTCTCATGGACC | | |
| MYF5-F | ACTACTATAGCCTGCCGGGG | 238 | qRT-PCR |
| MYF5-R | GGCAATCCAGGTTGCTCTGA | | |
| 3'RACE-F | GCTTCCCATCAGAAAGTATCAGGA | 141 | 3′ RACE |
| 5'RACE-R1 | TGCTAAACTGCTGGCTGACACTGGA | 295 | 5′ RACE |
| 5'RACE-R2 | TTCCTTGAAGTGGTGGATTC | 268 | 5′ RACE |
| Full length-F | GGTGTTGAGTAGGGAATGG | 472 | Full-length cloning |
| Full length-R | TTGTGTTGGTAGCTTGACTACG | | |
| pET-28a-F | CTCCGTCGACAAGCTTGGTGTTGAGTAGGGAATGG | 504 | Prokaryotic expression |
| pET-28a-R | GGTGGTGGTGCTCGAGTTGTGTTGGTAGCTTGACTACG | | |
| pGL3-1F | TATCGATAGGTACCGACAACATAGCAGATAATTCGAGTGT | 2787 | Luciferase reporter system construction for promoter active region identification |
| pGL3-2F | TATCGATAGGTACCGGCCAACTTTGGAGACCACTT | 1994 | |
| pGL3-3F | TATCGATAGGTACCGAATCGGTGGACTGCTAACCT | 1143 | |
| pGL3-4F | TATCGATAGGTACCGTCAGCATCACCAGTCACCAAC | 744 | |
| pGL3-5F | TATCGATAGGTACCGGCGAGAAAAGGAAACACCGC | 480 | |
| pGL3-6F | TATCGATAGGTACCGGGTTAGGCGGGAGGCTTGA | 296 | |
| pGL3-R1 | GCAGATCTCGAGCCCTCCCCCAGATCTCAAGGGAG | | |
| SNP-F | GTCTCCTCCTGCCTTCAATCT | 626 | SNP screening |
| SNP-R | CGAGCGCCAGTGTACCTC | | |
| pGL3-SNP-F | TAGCCCGGGACTCGAGTCTCCTCCTGCCTTCAATCT | 594 | Construction of luciferase reporter system of SNP1-SNP2 haplotypes |
| pGL3-SNP-R | CCGGAATGCCAAGCTTCGAGCGCCAGTGTACCTC | | |
| pGL3-SNP1-A-F | TCGCGTGTGGCCGAGAGGGGCGGCCCGGCCA | | |
| pGL3-SNP1-A-R | TGGCCGGGCCGCCCCTCTCGGCCACACGCGA | | |
| pGL3-SNP2-T-F | CTGCTTGATTGGTACTAGCCTCTTCTCCGCT | | |
| pGL3-SNP2-T-R | AGCGGAGAAGAGGCTAGTACCAATCAAGCAG | | |

(Bio-Rad Laboratories, Hercules, CA, USA) using SYBR® Premix Ex Taq™ II (Tli RNaseH Plus) (TaKaRa, Dalian, China) (Kang et al., 2019a). All samples were detected in triplicate. The relative expression levels of mRNA in tissue samples were calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). The correlations between genes were calculated using Pearson correlation analysis, and the differences between samples were calculated using Student *t*-test (Chen et al., 2018).

## Cell Culture, Plasmids Construction, and Transfection

The procedure for separating bovine myoblast from skeletal muscle was the same as the previous study of our lab (Sun et al., 2016). Then cells were cultured in incubator at 37°C with 5% $CO_2$. The proliferation medium for myoblast contains 80% Dulbecco's Modified Eagle Medium (DMEM), 20% fetal bovine serum (FBS), penicillin (10 U/ml), and streptomycin (10 mg/ml). When myoblast start to fuse, the proliferation medium was replaced by differential medium, which contains 2% horse serum, penicillin (10 U/ml), streptomycin (10 mg/ml), and DMEM. The RNA of the myoblast was collected using TRIzol reagent (TaKaRa, Dalian, China) at proliferation and differential stages. Mouse C2C12 myoblast cells, mouse 3T3-L1 embryo fibroblast, and human embryonic kidney 293T cells were used to uncover the active region of *lncFAM200B* promoter or single nucleotide polymorphisms (SNPs) effects on promoter activity. They were grown in 10% FBS, 90% DMEM, penicillin (10 U/ml), and streptomycin (10 mg/ml) medium.

To investigate the active region of *lncFAM200B* gene promoter, six fragments of the *lncFAM200B* promoter region were amplified and cloned into the pGL3-Basic vector (Promega, Madison, WI, USA) using *Sac*I and *Sma*I restriction enzymes (**Table 1**). These constructed plasmids were named as pGL3-pro1 (2,787 base pairs [bp]), pGL3-pro2 (1,994 bp), pGL3-pro3 (1,143 bp), pGL3-pro4 (744 bp), pGL3-pro5 (480 bp), and pGL3-pro6 (296 bp) according to their sequence length. The largest fragment (2,787 bp) spans from −2,446 nt to +310 nt of the *lncFAM200B* transcription start site. Additionally, four plasmids termed as pGL3-CC (SNP1-C and SNP2-C), pGL3-CT (SNP1-C and SNP2-T), pGL3-AC (SNP1-A and SNP2-C), and pGL3-AT (SNP1-A and SNP2-T) were constructed using overlap PCR to detect the effects of haplotype on promoter activity (**Table 1**). The vector pRL-TK was used as internal reference in the luciferase reporter system. The pGL3-Control and empty pGL3-Basic were used as positive and negative control, respectively (Xu et al., 2018; Kang et al., 2019b).

The plasmids were transfected into cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA). Before transfection, cells were seeded into 96-well plate. When cells covered 80% of the culture plate bottom, the plasmids were transient transfected according to the manufacturer's protocol. To normalize the transfection efficiency, the pRL-TK was transfected with constructed plasmids, and the transfection ratio of constructed plasmids and pRL-TK was 50:1 (Kang et al., 2019b). All transfections were carried out in triplicate. After 36 h, the cells were lysed, and the luciferase activity was measured using BHP9504 microporous-plate luminescence analyzer (Hamamatsu Photons Technology, Beijing, China). The relative luciferase activity of different promoter fragments were normalized by renilla luciferase activity (Xu et al., 2018; Kang et al., 2019b). The relative luciferase activity was represented by mean ± standard deviation. The one-way ANOVA and Bonferroni multiple comparisons were used to analyze the difference between groups (Yang et al., 2019).

## Genetic Variation Analyses of Bovine *lncFAM200B* Promoter Region

A total of 352 female cattle from four breeds were used in this study to identify the novel genetic variations in bovine *lncFAM200B* promoter region. The samples of Qinchuan cattle ($n$ = 139), Jinnan cattle ($n$ = 121), Nanyang cattle ($n$ = 67), and Ji'an cattle ($n$ = 25) were randomly collected from Shaanxi, Shanxi, He'nan, and Jiangxi provinces, respectively. The detailed information and records of body measurement traits for the cattle were the same as the published papers (Zhang et al., 2015; Jin et al., 2018). The blood DNA samples were isolated using high salt-extraction method (Aljanabi and Martinez, 1997). The primers (SNP-F and SNP-R) used to identify the genetic variations were designed based on the DNA sequence of bovine *lncFAM200B* gene. All the variations were identified by agarose gel electrophoresis and DNA sequencing (Sangon Biotech, Shanghai, China). After genotyping, the genotypic and allelic frequencies, population genetic diversity indexes [Hardy-Weinberg equilibrium (HWE), heterozygosity (He), effective population size (Ne), polymorphism information content (PIC)] were calculated according to the methods described as Nei (1973) using MSR website (http://www.msrcall.com/) (Wang et al., 2017; Yang et al., 2017). Then the association analyses between genotypes and records of body measurement traits were performed based on the reduced linear model below: $Y_i = u + G_i + e$, where $Y_i$ was the trait measured data for each animal; u was the over mean for each trait; $G_i$ was the effect of genotype; and e was the random error. Different breeds were analyzed separately. Due to all the cattle were 2–3 years old female and the individuals of the same breed were bred in the same farm, so this model excluded the farm, breed, years old, and sex factors. The linkage disequilibrium (LD) and haplotypes analyses were performed using SHEsis online platform (http://analysis.biox.cn[1]; Cui et al., 2018). The association analyses between genotypes or haplotypes and body measurement traits were performed by one-way ANOVA followed by Bonferroni multiple comparison (three groups) or independent-sample *t*-test (two groups) (Wang et al., 2019).

# RESULTS

## Characterization of Bovine *lncFAM200B*

Due to only partial sequence (369 nt) was obtained by RNA-Seq (Sun et al., 2016), the 5′ and 3′ RACE were carried out to obtain the full length of *lncFAM200B*. The 3′ and 5′ RACE obtained 174 bp and 323 bp sequences, respectively (**Figure 1**). The full-length of bovine *lncFAM200B* was 472 nt and had two exons

[1] Accessed: Sep 26, 2019.

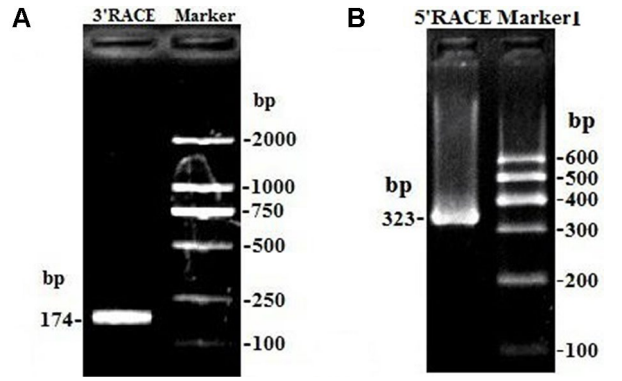


**FIGURE 1 |** The amplification products of *lncFAM200B* 3′ RACE and 5′ RACE. **(A)** the amplification product of *lncFAM200B* 3′ RACE, 174 bp = 141 bp (*lncFAM200B* sequence) + 15 bp (poly A) + 18 bp ($Q_i$). **(B)** the amplification product of *lncFAM200B* 5′ RACE, 323 bp = 268 bp (*lncFAM200B* sequence) + 33 bp (5′ RACE adapter) + 22 bp (primer).

(**Figures 2A, B**). The protein-coding potential prediction score of bovine *lncFAM200B* in CPC was −1.22524, which was far less than the scores of the known protein-coding genes *C/EBPα* and lncRNA *H19* (**Figure 2C**). Meantime, all the ORFs in *lncFAM200B* were smaller than 100 amino acids, illustrated that the coding ability of *lncFAM200B* was very low (Sun et al., 2016). To ensure the coding ability of *lncFAM200B*, the prokaryotic expression system was implemented and it showed that no protein was being encoded by *lncFAM200B* (**Figure 2D**).

The miRNA prediction analysis uncovered that 8 miRNAs might interact with *lncFAM200B*. Among these miRNAs, 5 miRNA scores were above 60, so we further predicted the target genes of these 5 miRNAs. As a result, some cell proliferation associated genes were uncovered, such as insulin like growth factor 2 mRNA binding protein 2 (*IGF2BP2*) (**Figure 2E**). Furthermore, as it is known that few lncRNAs could interact with their nearby genes, we searched the adjacent genes of *lncFAM200B*. Interestingly, we found that fibroblast growth

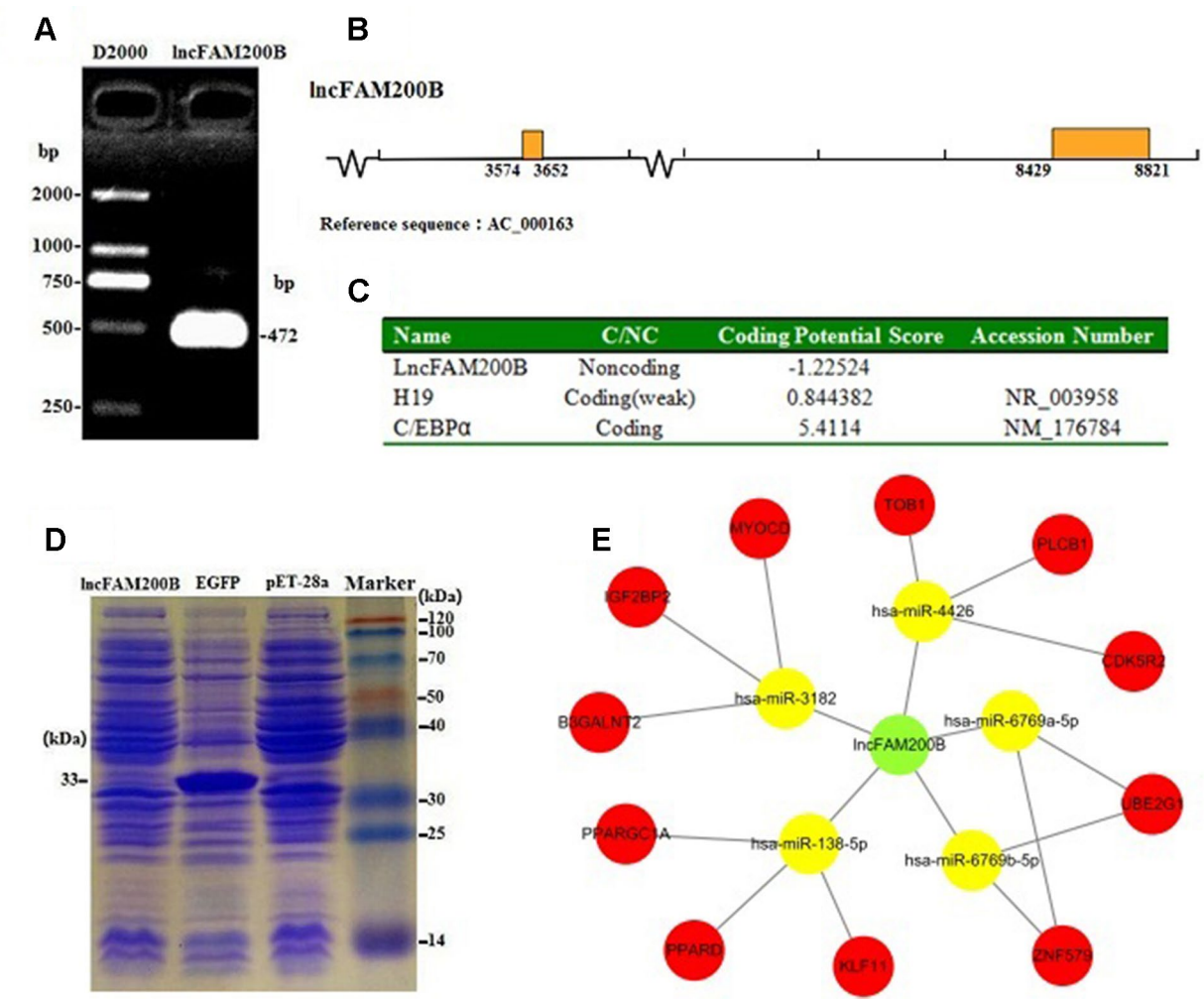


**FIGURE 2 |** Characterization of bovine *lncFAM200B*. **(A)** The full-length of *lncFAM200B*; **(B)** The distribution mode chart of *lncFAM200B* exons. The box and the line represented the exon and intron, respectively; **(C)** The coding ability prediction of *lncFAM200B* using CPC website; **(D)** The *in vitro* translation system of protein product from *lncFAM200B*; **(E)** The potential interacting miRNAs of *lncFAM200B* and miRNAs target genes.

factor binding protein 1 (*FGFBP1*) was close to *lncFAM200B*. Thus, *lncFAM200B* might interact with *FGFBP1* and affect cell proliferation and differentiation (Xie et al., 2006). The transcription factors binding sites prediction analysis found that within the 3000 bp sequence region upstream of *lncFAM200B*, there were 30 C/EBPα, 7 CCAAT/enhancer binding protein beta (C/EBPβ), and 43 SP1 transcription factor binding sites. Hayashi et al. (2016) found that the area enriched with SP1 was highly prone to promote the binding of MyoD and DNA sequence. Since the MyoD was a crucial transcription factor during muscle cell differentiation, we think that the identified region must be important for the transcription of bovine *lncFAM200B*.

## Expression Profiles of *lncFAM200B* in Bovine Tissues and Myoblasts

To reveal the function of *lncFAM200B*, we investigated the expression profiles in bovine embryonic, neonatal, and adult tissues. In various bovine tissues, *lncFAM200B* was widely expressed in three developmental stages (**Figures 3A–C**). In skeletal muscle, the expression level of *lncFAM200B* was low at each state, but was significantly different among the three developmental stages (**Figure 3D**), which was consistent with the RNA-Seq data. At the cellular level, we detected the expression level of *lncFAM200B*, *MyoD*, *MyoG*, *Myf5*, and *MYH3* genes in myoblast proliferation and differential stages, which were important in the regulation of myoblast development (**Figure 4**). The expression characteristic of *lncFAM200B* showed a significant positive correlation with the expression of *MyoG* (Pearson correlation coefficient = 0.922, $P = 0.003$) and *Myf5* (Pearson correlation coefficient = 0.741, $P = 0.035$) (**Table 2**). These results suggested that *lncFAM200B* might be involved in the development of bovine myoblasts.

## Identification of Bovine *lncFAM200B* Promoter Active Region

Considering the characteristic of *lncFAM200B* promoter region, this study further confirmed the promoter active region of bovine *lncFAM200B*. Six truncated fragments of the promoter region were constructed into pGL3-Basic plasmid and transfected into 293T, C2C12, and 3T3-L1 cells. By restriction enzyme identification and plasmids sequencing analyses, we confirmed that the recombinant plasmids were successfully constructed (**Figure 5**). The detection of double luciferase activity showed that the luciferase activity of different truncated fragments showed the same trend in these three different cell lines (**Figure 6D**). In each cell line, the luciferase activity of positive control

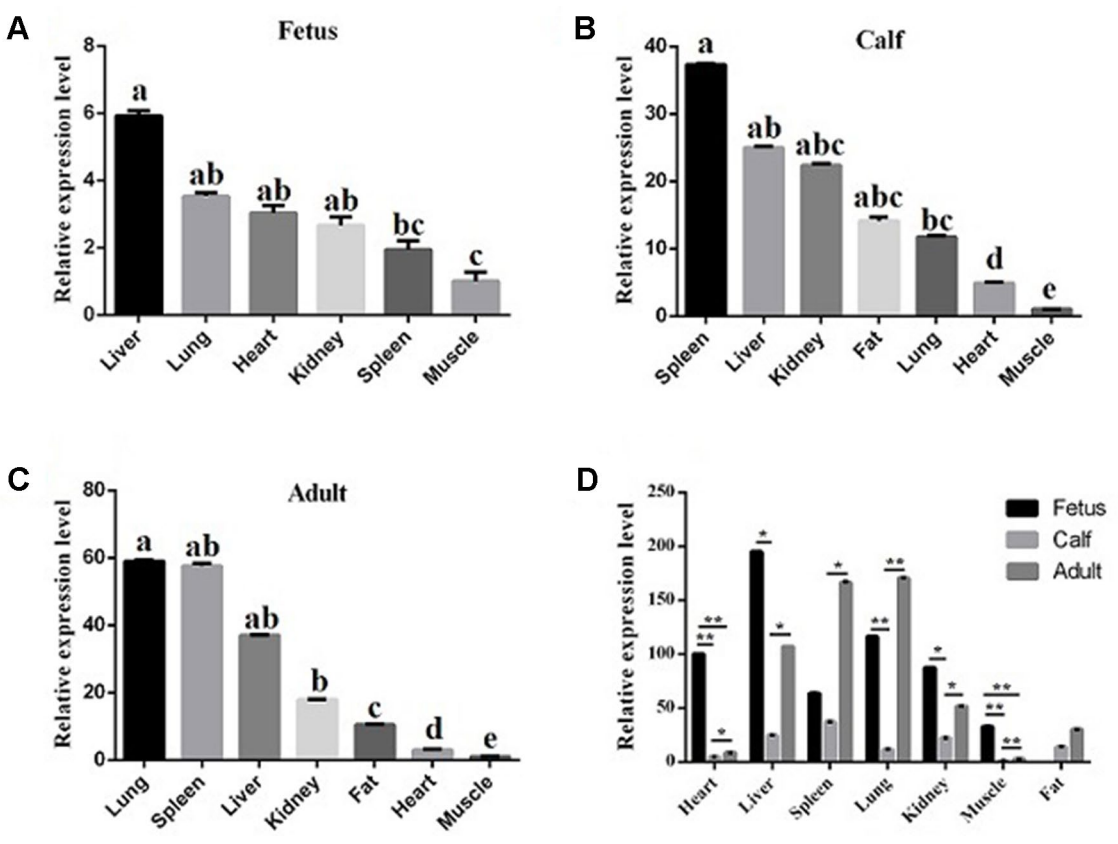


**FIGURE 3 |** The relative expression levels of lncFAM200B in tissues of Qinchuan cattle. Expression level of lncFAM200B in fetus **(A)**, calf **(B)**, adult, **(C)** tissues **(D)**. **(A**, **B**, **C)** The columns with different superscripts (a, b, c, d, e) within each figure differ significantly at $P < 0.05$ level. **(D)** $^{*}P < 0.05$; $^{**}P < 0.01$.

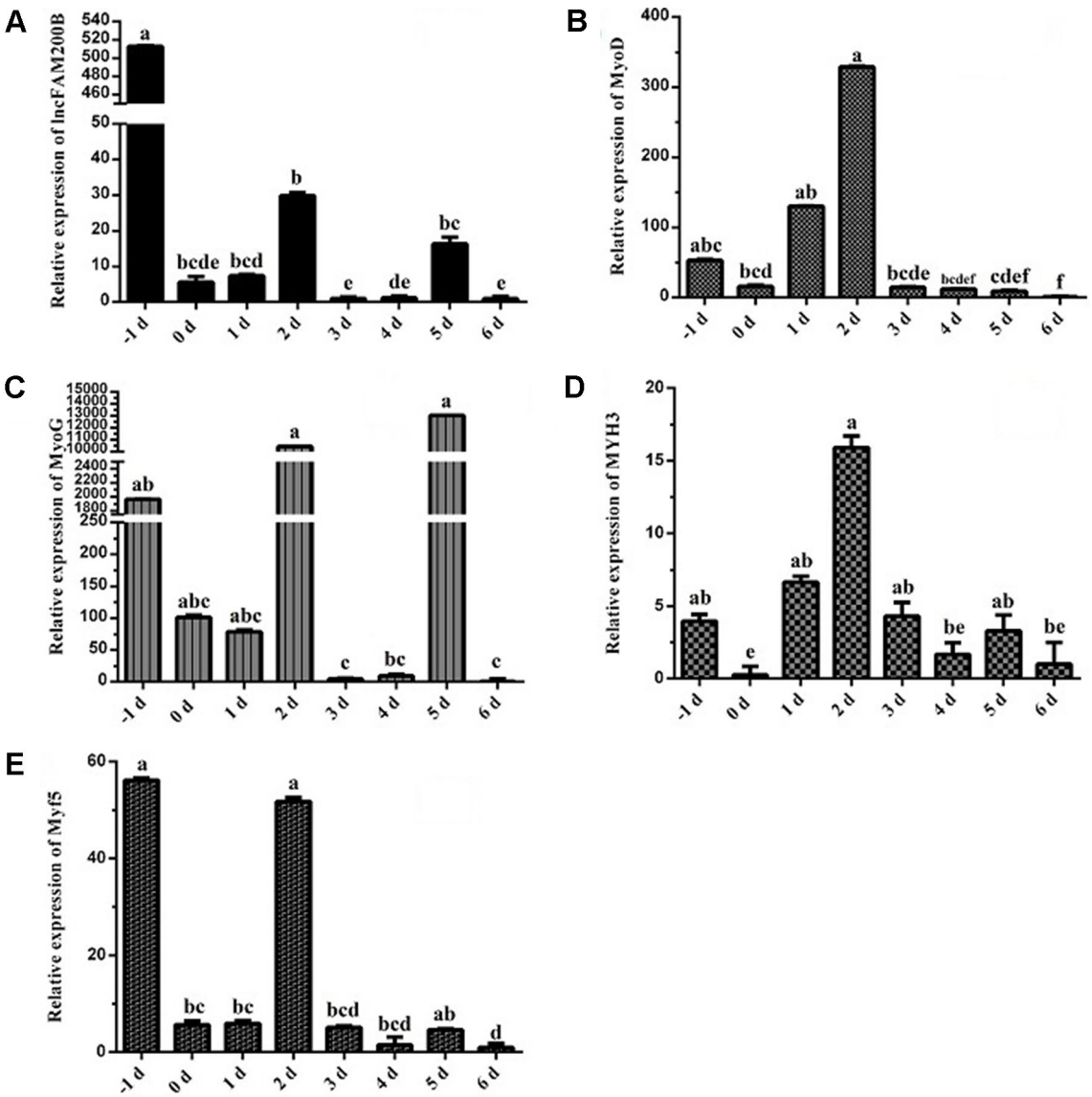


**FIGURE 4 |** Expression characteristics of *lncFAM200B* and myoblast development associated genes in bovine myoblast. Expression trend of *lncFAM200B* **(A)**, *MyoD* **(B)**, *MyoG* **(C)**, *MYH3* **(D)**, and *Myf5* **(E)** in bovine myoblast cultured in proliferation medium (–1 day) and differentiation medium (0, 1, 2, 3, 4, 5, and 6 days).

(pGL3-Control) was high, but the negative control (empty pGL3-Basic) was low (**Figures 6A–D**), providing the basis for our observations and correct experimental design. The pGL3-pro2, pGL3-pro3, and pGL3-pro4 yielded a significantly stronger luciferase activity compared to the other vectors ($P < 0.01$; **Figures 6A–C**), which suggested that these fragments contained promoter active region. The luciferase activity of the longest fragment pGL3-pro1 was lower than that of pGL3-pro2, pGL3-pro3, and pGL3-pro4 (**Figures 6A–C**), suggesting that there might be inhibitor binding sites in the region (−2,446 to −1,653) of the *lncFAM200B*. Particularly, from pGL3-pro4 to pGL3-pro5, the luciferase activity dramatically decreased ($P < 0.01$; **Figures 6A–C**), which meant that the active region was truncated in pGL3-pro5 and the active region was from −403 to −139 (264 nt) of the *lncFAM200B* transcription start site (**Figure 6D**). Besides, upon the transcription factor binding site prediction, we found 6 SP1 and 2 C/EBPα potential binding sites in the active region (−403 to −139) (**Figure 6E**). Above

**TABLE 2 |** Pearson correlation analyses between the expression of *lncFAM200B* and myoblast development associated genes in proliferation and differentiation states muscle cell.

| Gene | *MyoD* | *MyoG* | *MYH3* | *Myf5* |
|---|---|---|---|---|
| Pearson correlation coefficient | 0.527 | 0.922** | 0.442 | 0.741* |
| Sig.(2-tailed) | 0.179 | 0.003 | 0.273 | 0.035 |

*$P < 0.05$; **$P < 0.01$.

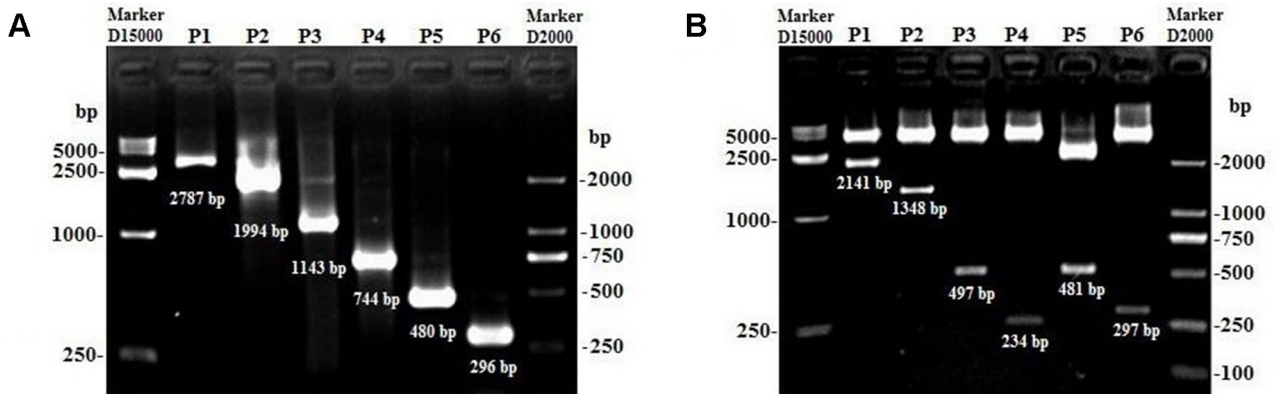


**FIGURE 5 |** The products of *lncFAM200B* promoter fragments and the identification of the plasmids using different restriction enzymes. **(A)** P1 to P6 represented the PCR products of pGL3-pro1 to pGL3-pro6. **(B)** P1 to P6 represented the recombined plasmids of pGL3-pro1 to pGL3-pro6 digested by different restriction enzymes.

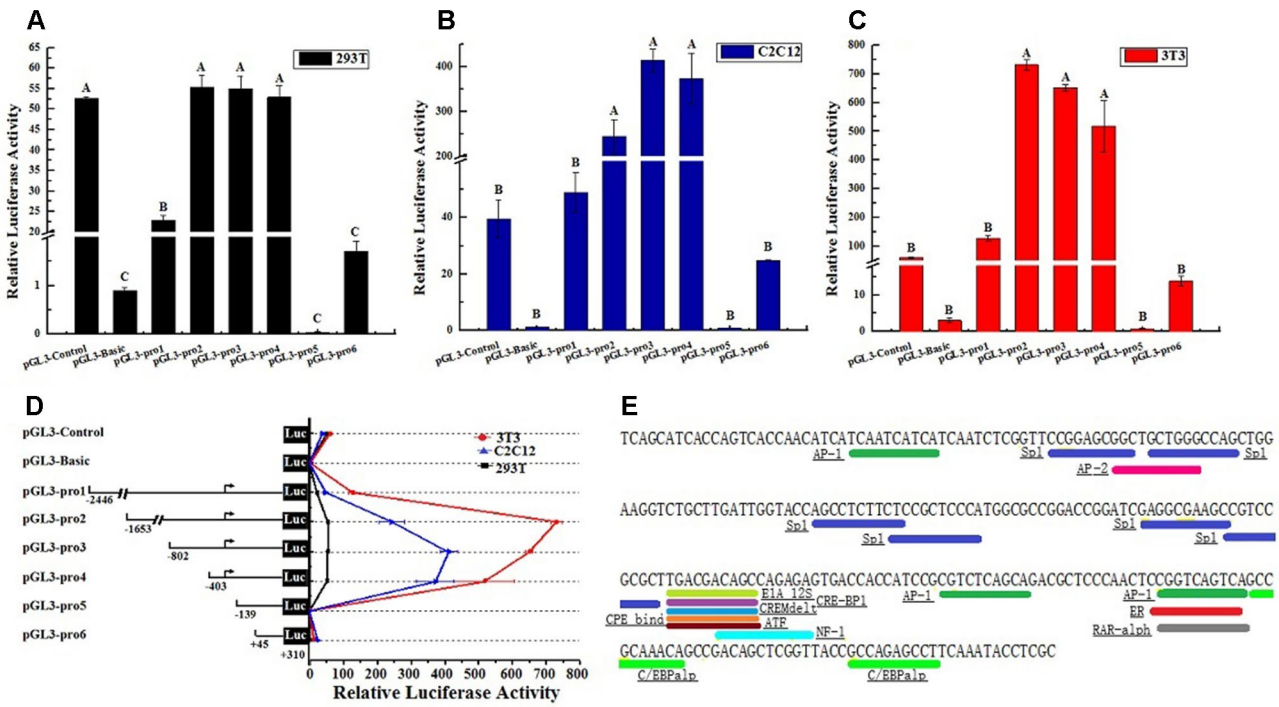


**FIGURE 6 |** The bovine *lncFAM200B* promoter active region. Relative luciferase activity of different promoter fragments in **(A)** 293T, **(B)** C2C12, **(C)** 3T3-L1 cell lines; **(D)** Relative luciferase activity changed trend; **(E)** Potential transcription factor binding sites in promoter activity region.

results suggested that the 264 nt active region was crucial for the expression of bovine *lncFAM200B*.

## Novel Genetic Variations in Bovine *lncFAM200B* Promoter Region

Promoter active region is very important for gene expression, hence we wanted to know whether there are crucial genetic variations in this region. Based on the DNA sequencing results, two SNPs were revealed in the promoter region of bovine *lncFAM200B*, SNP1 (NC_037333.1:g.110851632 C-A, rs456951291 in Ensembl database) and a novel genetic variant SNP2 (NC_037333.1:g.110851751 C-T, ERZ990081 in European Variation Archive) (**Figure 7**). Interestingly, SNP2 was in the promoter active region of bovine *lncFAM200B*.

At SNP1 locus, CC and CA genotypes were identified in cattle (three genotypes were identified in Jinnan cattle). At SNP2 locus, only CC and CT were identified in the four detected cattle breeds (**Table 3**;

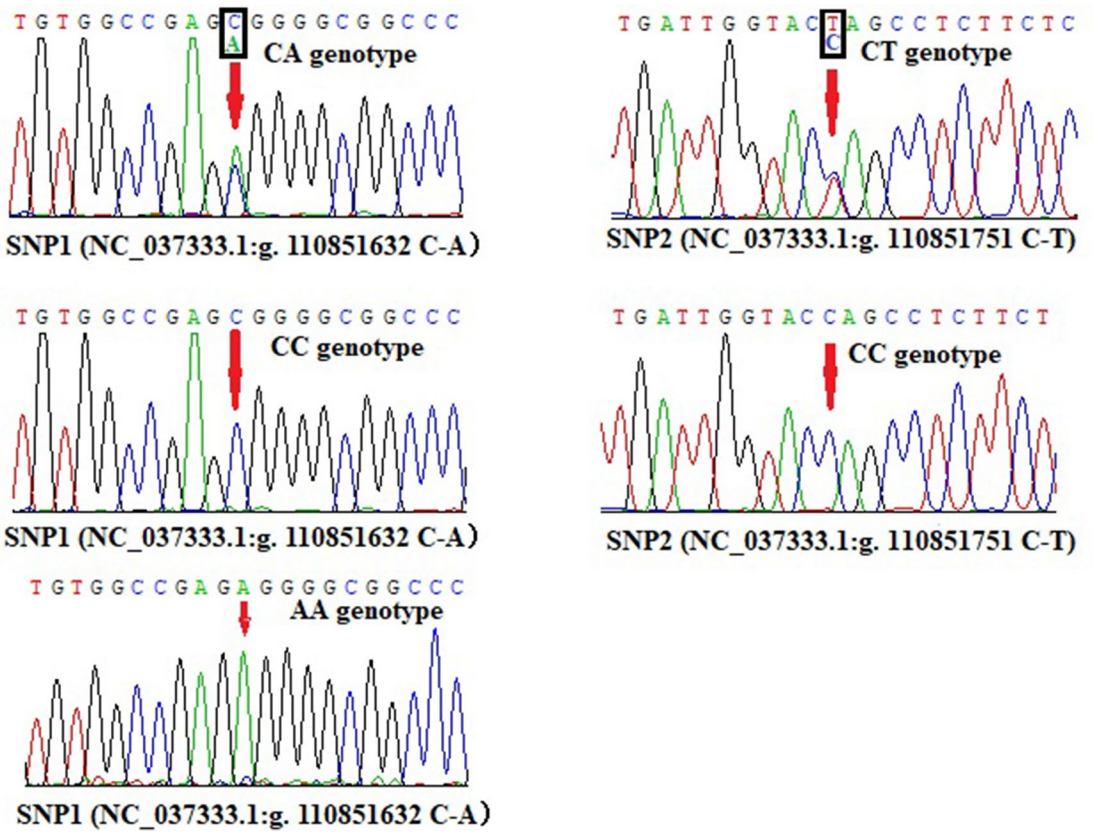


**FIGURE 7 |** Sequencing maps of SNP1 and SNP2 different genotypes in *lncFAM200B* gene promoter region.

**Figure 7**). At these two loci, C was the main allele in all the detected cattle breeds. The Chi-squared test showed that these loci were at Hardy-Weinberg equilibrium ($P > 0.05$) in the four populations (**Table 3**). Further, population genetic parameters indicated that the loci were polymorphic but belonged to low (PIC < 0.25) or moderate (0.25 < PIC < 0.50) polymorphisms categories (**Table 3**). Then LD analyses between SNP1 and SNP2 were analyzed in Qinchuan, Jinnan, and Ji'an populations ['in Nanyang cattle the individual numbers of CA (SNP1 locus) and CT (SNP2 locus) were found to be smaller than 3, so we did not perform the LD analysis and the follow association analysis]. The $D'$ and $r^2$ values in Qinchuan ($D'$ = 1.000, $r^2 = 0.735$), Jinnan ($D' = 0.611$, $r^2 = 0.049$), and Ji'an ($D' = 0.857$, $r^2 = 0.532$) cattle populations showed these two loci were linked in cattle. The $r^2$ reflects the extent of the linkage disequilibrium and $r^2 >$ 0.33 indicated that there was a sufficiently strong linkage between the two loci. When different genotypes are evenly distributed in the population, the $D' > 0.33$ can also be used to judge that there was a linkage disequilibrium (Zhao et al., 2007).

**TABLE 3 |** Calculation of the parameters of the genetic variations in bovine *lncFAM200B* promoter region.

| Loci/Breeds | Genotype numbers (frequencies) | | | Allele frequencies | | HWE | Population parameters | | |
|---|---|---|---|---|---|---|---|---|---|
| **SNP1** | **CC** | **CA** | **AA** | **C** | **A** | ***P* values** | **He** | **Ne** | **PIC** |
| Nanyang | 65 (0.97) | 2 (0.03) | / | 0.99 | 0.01 | 0.901 | 0.029 | 1.030 | 0.029 |
| Qinchuan | 119 (0.86) | 20 (0.14) | / | 0.93 | 0.07 | 0.361 | 0.134 | 1.154 | 0.125 |
| Jinnan | 46 (0.38) | 58 (0.48) | 17 (0.14) | 0.62 | 0.38 | 0.851 | 0.459 | 1.848 | 0.354 |
| Ji'an | 14 (0.56) | 11 (0.44) | / | 0.72 | 0.28 | 0.158 | 0.343 | 1.523 | 0.284 |
| **SNP2** | **CC** | **CT** | **TT** | **C** | **T** | ***P* values** | **He** | **Ne** | **PIC** |
| Nanyang | 65 (0.97) | 2 (0.03) | / | 0.99 | 0.01 | 0.901 | 0.029 | 1.030 | 0.029 |
| Qinchuan | 124 (0.89) | 15 (0.11) | / | 0.95 | 0.05 | 0.501 | 0.102 | 1.114 | 0.097 |
| Jinnan | 103 (0.85) | 18 (0.15) | / | 0.93 | 0.07 | 0.377 | 0.138 | 1.160 | 0.128 |
| Ji'an | 11 (0.44) | 14 (0.56) | / | 0.72 | 0.28 | 0.052 | 0.403 | 1.680 | 0.322 |

*HWE, Hardy-Weinberg equilibrium; He, heterozygosity; Ne, effective population size; PIC, polymorphism information content.*

**TABLE 4 |** Genotypic frequencies of *lncFAM200B* SNP1-SNP2 combined genotypes in cattle.

| Breeds | Sample size(*N*) | SNP1-SNP2 combined genotypes numbers (frequencies) | | | | | |
|---|---|---|---|---|---|---|---|
| | | CC-CC | CA-CT | CA-CC | AA-CC | CC-CT | AA-CT |
| QC | 139 | 119 (0.86) | 15 (0.11) | 5 (0.03) | / | / | / |
| Jinnan | 121 | 44 (0.36) | 15 (0.12) | 43 (0.36) | 17 (0.14) | 1 (0.01) | 1 (0.01) |
| Ji'an | 25 | 10 (0.40) | 10 (0.40) | 1 (0.04) | / | 4 (0.16) | / |

The association analyses found that the genotypes of SNP1 were significantly associated with the hip height in Jinnan cattle ($P$ = 0.012). The hip height of the CC genotype carriers was 131.7 ± 6.7 cm, which was evidently higher than that of CA (128.6 ± 6.5 cm) and AA (127.1 ± 5.4 cm) genotype carriers, but we did not observe any significant difference between CA and AA genotype carriers (**Figure 8**). Besides, at SNP1 and SNP2 loci, the body measurement traits (hip height, body height, body length, heart girth, rump length) of CC genotype carriers were all better than the carriers with the other genotypes in Jinnan cattle (**Figure 8**). Furthermore, the combined genotypes of SNP1 and SNP2 were found to be significantly associated with hip height in Jinnan cattle ($P$ = 0.033). The hip height of the CC-CC carriers (132.0 ± 6.6 cm, $n$ = 44) was markedly higher than that of CA-CT (127.7 ± 7.9 cm, $n$ = 15), CA-CC (128.9 ± 6.0 cm, $n$ = 43), and AA-CC (127.1 ± 5.4 cm, $n$ = 17) genotype carriers (**Figure 9**). Because we only found one individual with CC-CT and one individual with AA-CT genotype, they were excluded in association analyses (**Table 4**). In Qinchuan and Ji'an cattle, no significant association was found between SNP1, SNP2, or the combined genotypes and the body measurement traits.

## Influence of the Haplotypes on the Transcriptional Activity of Bovine *lncFAM200B*

Bearing in mind the significant relationship between SNP1 and the combined genotypes with the cattle body measurement traits, we wanted to further investigate the mechanism that contributed to the phenotype. Four plasmids (pGL3-CC, pGL3-CT, pGL3-AC, pGL3-AT) of SNP1 and SNP2 haplotypes were constructed and transfected into commonly used 293T cells to detect the luciferase activity. The luciferase activity of positive control (Control) was significantly higher compared to that of the negative control (empty Basic) and we found that the relative luciferase activity of pGL3-CC was the highest among the four haplotypes in 293T cells. The luciferase activities of pGL3-CC and pGL3-AT were significantly higher than that of the pGL3-CT haplotypes ($P$ < 0.05). But no difference was found among the other haplotypes (**Figure 10**). These results suggested that the genotypes of SNP1-SNP2 haplotypes influenced the body measurement traits by regulating the expression of *lncFAM200B*.

# DISCUSSION

With the rapid development of high-throughput sequencing technology, an increasing number of lncRNAs have been discovered in many animal species. Structurally, the lncRNA resembled protein-coding gene with its own promoter, exons, and introns. The *lncFAM200B* was screened from the sequencing results obtained in an earlier study done by Sun et al. (2016). In their study, they implemented strict parameters to identify the lncRNA from the sequencing results such as the number of exons must be ≥2, the size must ≥200 nt, the read number should

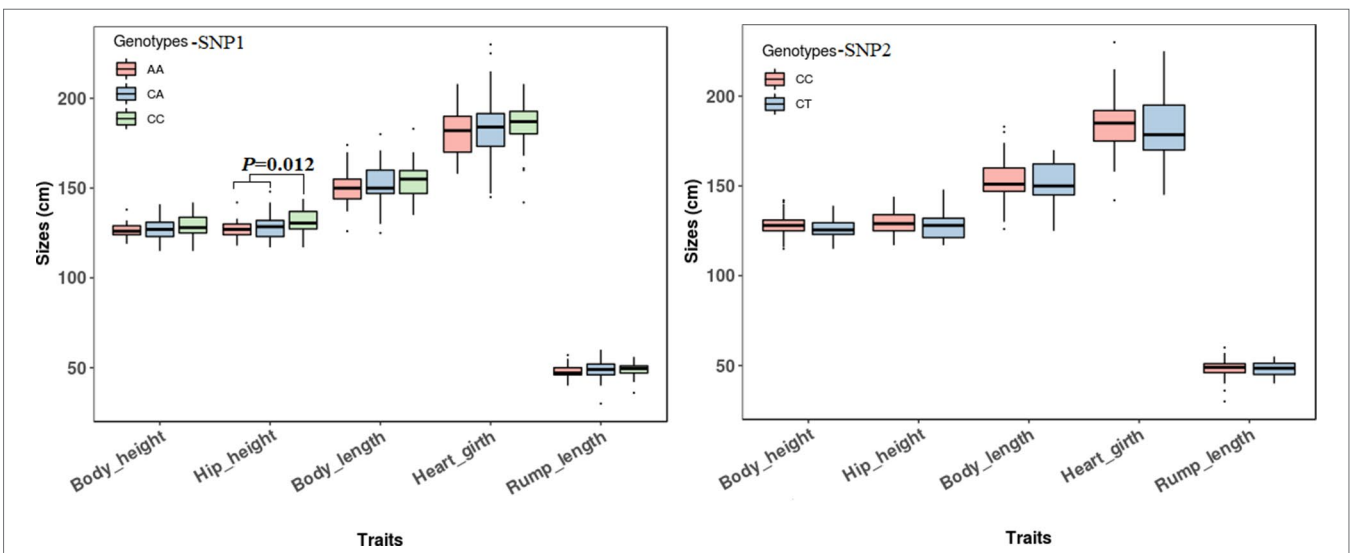


**FIGURE 8 |** Association of *lncFAM200B* SNPs (left–SNP1; right–SNP2) and body measurement traits of Jinnan cattle.

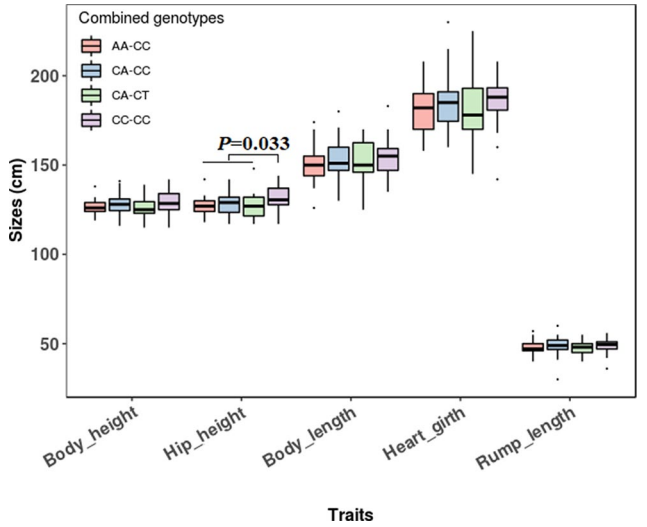


**FIGURE 9 |** Association of *lncFAM200B* SNP1-SNP2 combined genotypes and body measurement traits of Jinnan cattle.

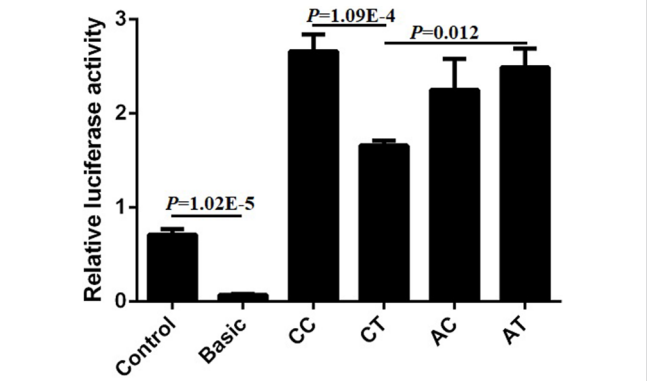


**FIGURE 10 |** Relative luciferase activity of different haplotypes of bovine *lncFAM200B* SNP1-SNP2 in 293T cell.

be >3, the ORF should be no longer than 100 amino acids, and the predicted protein-coding potential should be weak (Li et al., 2016; Sun et al., 2016). Based on their research, we used different methods (RACE, *in vitro* prokaryotic expression system, and protein-coding ability prediction analysis) to further prove that *lncFAM200B* was a novel lncRNA.

Expression analysis found that the expression of *lncFAM200B* positively correlated with the expression of *MyoG* ($P = 0.003$) and *Myf5* ($P = 0.035$). MyoG, a muscle-specific transcription factor, positively regulated the skeletal muscle fiber development, myoblast differentiation, and fusion, and was found to be indispensable for myogenic differentiation (Zammit, 2017). Myf5 is a master regulator belonging to the MRFs family and is known to play a key role in muscle differentiation or myogenesis. *Myf5* is a master gene for the determination of skeletal muscle, which pushes the myogenic precursors into myoblasts (Dimicoli-Salazar et al., 2011). The genes have the same expression pattern may have the same function, such as *MEGF10*, a myogenic regulator of satellite cells in skeletal muscle, shares a similar expression pattern with *MyoG* in muscle regeneration (Park et al., 2014). Thus, we hypothesize that *lncFAM200B* might play a positive role in muscle development.

The molecular markers based on nucleotide sequence variations among individuals, which are the directly reflection of genetic polymorphism in DNA level. Compared to the morphological markers, DNA molecular markers have many advantages. Genomic variations are extremely abundant and are the impetus of biological evolution providing rich material for animal breeding. At different stages of biological development, such as the early disease diagnosis and early animal selection for breeding, the DNA markers can be used. The detection method of DNA genetic variations is simple and rapid. Nowadays, DNA markers are widely used in biological evolution analysis, genetics analysis, diagnosis of genetic diseases and so on (Alidoust et al., 2018). In animal breeding, it is important to explore crucial markers. In cattle, numerous variations have been identified within the protein-coding genes, but only a few studies have uncovered the variations in the non-coding RNA genes (Jin et al., 2018; Yu et al., 2018). In this study, first, we analyzed the SNPs in the promoter region of *lncFAM200B* gene and found that the SNP1 was linked with the promoter active region mutation, SNP2. Importantly, the genotypes of SNP1 and combined genotypes of SNP1 and SNP2 were associated with the hip height in Jinnan cattle.

We attempted to uncover the cause of the above SNP effect on the cattle growth trait. Promoter regulates the activity of gene by affecting the binding of transcription factors and DNA promoter region sequences. Mutations in the gene promoter region will result in gene expression disorder, further resulting in phenotypic changes and disease (Lu et al., 2019). In this study, we used the dual-luciferase reporter system to detect the effects of SNP1 and SNP2 variations on gene expression. In the commonly used 293T cells, haplotype CC showed the highest fluorescence value followed by haplotype AT and both were significantly higher than haplotype CT. The haplotype CC had the highest hip height, which agreed with the luciferase activity data. These results further provided evidence proving that *lncFAM200B* is a positive regulator of muscle development.

## CONCLUSION

The lncRNA *lncFAM200B* differentially expressed in embryonic, neonatal, and adult bovine skeletal muscles. In myoblast proliferation and differential stages, the expression characteristic of *lncFAM200B* was positively correlated with the expression of *MyoG* and *Myf5*. In *lncFAM200B* active region (−403 to −139 of *lncFAM200B* transcription start site), one novel SNP (SNP2, NC_037333.1:g.110851751 C-T, ERZ990081) was discovered which linked with the nearby SNP1 (rs456951291). The genotypes of the SNP1 and the combined genotypes of SNP1 and SNP2 were significantly associated with the hip height in Jinnan cattle. Interestingly, haplotype CC had the highest luciferase activity and the highest hip height. Our results established that *lncFAM200B* is a positive regulator of muscle development and we believe that our studies will help in advancing the beef cattle MAS breeding program.

## DATA AVAILABILITY STATEMENT

The detailed information of SNP2 can be found in the European Variation Archive database after 2019/12/31. Project: PRJEB33081; Analyses: ERZ990081.

## ETHICS STATEMENT

The animal study was reviewed and approved by Faculty Animal Policy and Welfare Committee of Northwest A&F University (no. NWAFAC1008).

## AUTHOR CONTRIBUTIONS

SZ and XL conceived and designed the experiments. SZ, ZK, and CP performed the experiments. XS provided the RNA-Seq data. XC, RD, CL, HC, and XL collected the DNA samples. SZ and XL analyzed the data. XL contributed reagents, materials, and analysis tools. XL and SZ wrote the paper.

## FUNDING

This work was funded by the National Natural and Science Foundation of China (No. 31672400).

## REFERENCES

Alidoust, M., Hamzehzadeh, L., Rivandi, M., and Pasdar, A. (2018). Polymorphisms in non-coding RNAs and risk of colorectal cancer: a systematic review and meta-analysis. *Crit. Rev. Oncol. Hematol.* 132, 100–110. doi: 10.1016/j.critrevonc.2018.09.003

Aljanabi, S. M., and Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res.* 25, 4692–4693. doi: 10.1093/nar/25.22.4692

Bharathy, N., Ling, B. M., and Taneja, R. (2013). Epigenetic regulation of skeletal muscle development and differentiation. *Subcell Biochem.* 61, 139–150. doi: 10.1007/978-94-007-4525-4_7

Billerey, C., Boussaha, M., Esquerré, D., Rebours, E., Djari, A., Meersseman, C., et al. (2014). Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* 15, 499. doi: 10.1186/1471-2164-15-499

Chen, H. J., Ihara, T., Yoshioka, H., Itoyama, E., Kitamura, S., Nagase, H., et al. (2018). Expression levels of brown/beige adipocyte-related genes in fat depots of vitamin A-restricted fattening cattle. *J. Anim. Sci.* doi: 10.1093/jas/sky240

Chen, M., Wang, J., Liu, N., Cui, W., Dong, W., Xing, B., et al. (2019). Pig SOX9: expression profiles of Sertoli cell (SCs) and a functional 18 bp indel affecting testis weight. *Theriogenology* 138, 94–101. doi: 10.1016/j.theriogenology.2019.07.008

Cui, Y., Yan, H., Wang, K., Xu, H., Zhang, X., Zhu, H., et al. (2018). Insertion/Deletion within the KDM6A gene is significantly associated with litter size in goat. *Front. Genet.* 9, 91. doi: 10.3389/fgene.2018.00091

Dimicoli-Salazar, S., Bulle, F., Yacia, A., Massé, J. M., Fichelson, S., and Vigon, I. (2011). Efficient in vitro myogenic reprogramming of human primary mesenchymal stem cells and endothelial cells by Myf5. *Biol. Cell* 103, 531–542. doi: 10.1042/BC20100112

Fernandes, J. C. R., Acuña, S. M., Aoki, J. I., Floeter-Winter, L. M., and Muxel, S. M. (2019). Long non-coding RNAs in the regulation of gene eExpression: physiology and disease. *Noncoding RNA* 5, pii: E17. doi: 10.3390/ncrna5010017

Hayashi, S., Manabe, I., Suzuki, Y., Relaix, F., and Oishi, Y. (2016). Klf5 regulates muscle differentiation by directly targeting muscle-specific genes in cooperation with MyoD in mice. *Elife.* 5, e17462. doi: 10.7554/eLife.17462

Jin, C. F., Li, Y., Ding, X. B., Li, X., Zhang, L. L., Liu, X. F., et al. (2017). lnc133b, a novel, long non-coding RNA, regulates bovine skeletal muscle satellite cell proliferation and differentiation by mediating miR-133b. *Gene* 630, 35–43. doi: 10.1016/j.gene.2017.07.066

Jin, Y., Yang, Q., Zhang, M., Zhang, S., Cai, H., Dang, R., et al. (2018). Identification of a novel polymorphism in bovine lncRNA ADNCR gene and its association with growth traits. *Anim. Biotechnol.* 30, 159–165. doi: 10.1080/10495398.2018.1456446

Kang, Z., Jiang, E., Wang, K., Pan, C., Chen, H., Yan, H., et al. (2019a). Goat membrane associated ring-CH-type finger 1 (MARCH1) mRNA expression and association with litter size. *Theriogenology* 128, 8–16. doi: 10.1016/j.theriogenology.2019.01.014

Kang, Z., Zhang, S., He, L., Zhu, H., Wang, Z., Yan, H., et al. (2019b). A 14-bp functional deletion within the CMTM2 gene is significantly associated with litter size in goat. *Theriogenology* 139, 49–57. doi: 10.1016/j.theriogenology.2019.07.026

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35 (Web Server issue), W345–W349. doi: 10.1093/nar/gkm391

Li, M., Sun, X., Cai, H., Sun, Y., Plath, M., Li, C., et al. (2016). Long non-coding RNA ADNCR suppresses adipogenic differentiation by targeting miR-204. *Biochim. Biophys. Acta* 1859, 871–882. doi: 10.1016/j.bbagrm.2016.05.003

Li, Y., Chen, X., Sun, H., and Wang, H. (2018). Long non-coding RNAs in the regulation of skeletal myogenesis and muscle diseases. *Cancer Lett.* 417, 58–64. doi: 10.1016/j.canlet.2017.12.015

Liu, X. F., Ding, X. B., Li, X., Jin, C. F., Yue, Y. W., Li, G. P., et al. (2017). An atlas and analysis of bovine skeletal muscle long noncoding RNAs. *Anim. Genet.* 48, 278–286. doi: 10.1111/age.12539

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lu, V. M., Goyal, A., Lee, A., Jentoft, M., Quinones-Hinojosa, A., and Chaichana, K. L. (2019). The prognostic significance of TERT promoter mutations in meningioma: a systematic review and meta-analysis. *J. Neurooncol.* 142, 1–10. doi: 10.1007/s11060-018-03067-x

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci U S A* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Park, S. Y., Yun, Y., Kim, M. J., and Kim, I. S. (2014). Myogenin is a positive regulator of MEGF10 expression in skeletal muscle. *Biochem. Biophys. Res. Commun.* 450, 1631–1637. doi: 10.1016/j.bbrc.2014.07.061

Scotto-Lavino, E., Du, G., and Frohman, M. A. (2006). 3' end cDNA amplification using classic RACE. *Nat. Protoc.* 1, 2742–2745. doi: 10.1038/nprot.2006.481

Sun, X., Li, M., Sun, Y., Cai, H., Lan, X., Huang, Y., et al. (2016). The developmental transcriptome sequencing of bovine skeletal muscle reveals a long noncoding RNA, lncMD, promotes muscle differentiation by sponging miR-125b. *Biochim. Biophys. Acta* 1863, 2835–2845. doi: 10.1016/j.bbamcr.2016.08.014

Wang, X., Yang, Q., Wang, K., Zhang, S., Pan, C., Chen, H., et al.. (2017). A novel 12-bp indel polymorphism within the GDF9 gene is significantly associated with litter size and growth traits in goats. *Anim. Genet.* 48, 735–736. doi: 10.1111/age.12617

Wang, X., Yang, Q., Wang, K., Yan, H., Pan, C., Chen, H., et al. (2019). Two strongly linked single nucleotide polymorphisms (Q320P and V397I) in GDF9 gene are associated with litter size in cashmere goats. *Theriogenology* 125, 115–121. doi: 10.1016/j.theriogenology.2018.10.013

Xie, B., Tassi, E., Swift, M. R., McDonnell, K., Bowden, E. T., Wang, S., et al. (2006). Identification of the fibroblast growth factor (FGF)-interacting domain in a secreted FGF-binding protein by phage display. *J. Biol. Chem.* 281, 1137–1144. doi: 10.1074/jbc.M510754200

Xu, Y., Shi, T., Zhou, Y., Liu, M., Klaus, S., Lan, X., et al. (2018). A novel PAX7 10-bp indel variant modulates promoter activity, gene expression and contributes to different phenotypes of Chinese cattle. *Sci. Rep.* 8, 1724. doi: 10.1038/s41598-018-20177-8

Yan, P., Luo, S., Lu, J. Y., and Shen, X. (2017). Cis- and trans-acting lncRNAs in pluripotency and reprogramming. *Curr. Opin. Genet. Dev.* 46, 170–178. doi: 10.1016/j.gde.2017.07.009

Yang, Q., Yan, H., Li, J., Xu, H., Wang, K., Zhu, H., et al. (2017). A novel 14-bp duplicated deletion within goat GHR gene is significantly associated with growth traits and litter size. *Anim. Genet.* 48, 499–500. doi: 10.1111/age.12551

Yang, Q., Zhang, S., Li, J., Wang, X., Peng, K., Lan, X., et al. (2019). Development of a touch-down multiplex PCR method for simultaneously rapidly detecting three novel insertion/deletions (indels) within one gene: an example for goat GHR gene. *Anim. Biotechnol.* doi: 10.1080/10495398.2018.1517770

Yu, X., Zhang, Y., Li, T., Ma, Z., Jia, H., Chen, Q., et al. (2017). Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. *Nat. Commun.* 8, 14016. doi: 10.1038/ncomms14016

Yu, X., Wang, Z., Sun, H., Yang, Y., Li, K., and Tang, Z. (2018). Long non-coding MEG3 is a marker for skeletal muscle development and meat production traits in pigs. *Anim. Genet.* 49, 571–578. doi: 10.1111/age.12712

Yue, Y., Jin, C., Chen, M., Zhang, L., Liu, X., Ma, W., et al. (2017). A lncRNA promotes myoblast proliferation by up-regulating GH1. *Cell. Dev. Biol. Anim.* 53, 699–705. doi: 10.1007/s11626-017-0180-z

Zammit, P. S. (2017). Function of the myogenic regulatory factors Myf5, MyoD, Myogenin and MRF4 in skeletal muscle, satellite cells and regenerative myogenesis. *Semin Cell Dev. Biol.* 72, 19–32. doi: 10.1016/j.semcdb.2017.11.011

Zhang, S., Dang, Y., Zhang, Q., Qin, Q., Lei, C., Chen, H., et al. (2015). Tetra-primer amplification refractory mutation system PCR (T-ARMS-PCR) rapidly identified a critical missense mutation (P236T) of bovine ACADVL gene affecting growth traits. *Gene* 559, 184–188. doi: 10.1016/j.gene.2015.01.043

Zhao, H., Nettleton, D., and Dekkers, J. C. (2007). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. *Genet Res.* 89, 1–6. doi: 10.1017/S0016672307008634

Zhu, M., Liu, J., Xiao, J., Yang, L., Cai, M., Shen, H., et al. (2017). Lnc-mg is a long non-coding RNA that promotes myogenesis. *Nat. Commun.* 8, 14718. doi: 10.1038/ncomms14718

**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: 08 November 2019
doi: 10.3389/fgene.2019.01129

# DNA Sequence Variants and Protein Haplotypes of Casein Genes in German Black Pied Cattle (DSN)

*Saskia Meier, Paula Korkuć, Danny Arends and Gudrun A. Brockmann**

*Faculty of Life Sciences, Albrecht Daniel Thaer Institute for Agricultural and Horticultural Sciences, Animal Breeding Biology and Molecular Genetics, Humboldt University of Berlin, Berlin, Germany*

Casein proteins were repeatedly examined for protein polymorphisms and frequencies in diverse cattle breeds. The occurrence of casein variants in Holstein Friesian, the leading dairy breed worldwide, is well known. The frequencies of different casein variants in Holstein are likely affected by selection for high milk yield. Compared to Holstein, only little is known about casein variants and their frequencies in German Black Pied cattle ("Deutsches Schwarzbuntes Niederungsrind," DSN). The DSN population was a main genetic contributor to the current high-yielding Holstein population. The goal of this study was to investigate casein (protein) variants and casein haplotypes in DSN based on the DNA sequence level and to compare these with data from Holstein and other breeds. In the investigated DSN population, we found no variation in the alpha-casein genes *CSN1S1* and *CSN1S2* and detected only the *CSN1S1*B* and *CSN1S2*A* protein variants. For *CSN2* and *CSN3* genes, non-synonymous single nucleotide polymorphisms leading to three different β and κ protein variants were found, respectively. For β-casein protein variants $A^1$, $A^2$, and *I* were detected, with $CSN2*A^1$ (82.7%) showing the highest frequency. For κ-casein protein variants *A*, *B*, and *E* were detected in DSN, with the highest frequency of *CSN3*A* (83.3%). Accordingly, the casein protein haplotype *CSN1S1*B-CSN2*A*$^1$*-CSN1S2*A-CSN3*A* (order of genes on BTA6) is the most frequent haplotype in DSN cattle.

**Keywords: sequencing, 1000 Bull Genomes Project, bovine, SNP, comparative genomics, endangered**

OPEN ACCESS

***Edited by:***
*Gábor Mészáros, University of Natural Resources and Life Sciences Vienna, Austria*

***Reviewed by:***
*Martin Johnsson, Swedish University of Agricultural Sciences, Sweden*
*Joanna Szyda, Wroclaw University of Environmental and Life Sciences, Poland*

****Correspondence:***
*Gudrun A. Brockmann*
*gudrun.brockmann@agrar.hu-berlin.de*

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

***Received:*** *13 June 2019s*
***Accepted:*** *17 October 2019*
***Published:*** *08 November 2019*

***Citation:***
*Meier S, Korkuć P, Arends D and Brockmann GA (2019) DNA Sequence Variants and Protein Haplotypes of Casein Genes in German Black Pied Cattle (DSN). Front. Genet. 10:1129. doi: 10.3389/fgene.2019.01129*

## INTRODUCTION

The German Black Pied cattle (DSN, "Deutsches Schwarzbuntes Niederungsrind") is a dual-purpose breed for milk and beef production. DSN is considered the founder population of the high-yielding Holstein Friesian breed (Köppe-Forsthoff, 1967; Grothe, 1993). The DSN ancestors have their roots in the German and Dutch North Sea coast region. While DSN cattle produce about 2,500 kg less milk per lactation compared to German Holstein, they were almost entirely replaced by Holstein and DSN became an endangered breed with currently about 2,800 cows registered in Germany. Nevertheless, with 4.3% fat and 3.7% protein, milk from DSN cows contains more protein and fat compared to Holstein (RBB Rinderproduktion Berlin-Brandenburg GmbH, 2016). Moreover, DSN cattle are considered to be more robust and fertile.

To preserve the DSN breed and conserve the genetic diversity, farmers are financially compensated for the lower milk yield by the EU and the German government. The close genetic relationship to

Holstein makes a genetic comparison between the original DSN and Holstein interesting with respect to differences in milk yield and protein composition.

Genes known to influence protein content and composition in milk are the casein genes *CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3*, encoding the casein proteins alpha S1 ($\alpha_{S1}$), beta (β), alpha S2 ($\alpha_{S2}$), and kappa (κ), respectively (Ferretti et al., 1990; Threadgill and Womack, 1990), which are located in the given order on BTA6 in the so-called casein gene cluster, which spans ~250 kb (Boettcher et al., 2004). All caseins account for about 75% of the milk protein content (Gallinat et al., 2013); the remaining 25% are whey proteins. Several single nucleotide polymorphisms (SNPs) and insertions or deletions in exons of these casein genes are known to change their protein sequences, resulting in different casein variants. In the *Bos* genus, 10 protein variants for $\alpha_{S1}$- (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, and *J*), 15 for β- ($A^1$, $A^2$, $A^3$, *B*, *C*, *D*, *E*, *F*, *G*, $H^1$, $H^2$, *I*, *J*, *K*, and *L*), 5 for $\alpha_{S2}$- (*A*, *B*, *C*, *D*, and *E*), and 11 for κ-casein (*A*, *B*, *C*, *E*, $F^1$, $F^2$, $G^1$, $G^2$, *H*, *I*, and *J*) have been reported (**Table 1**). Additional variants in the upstream gene regions could affect the expression of the casein genes and influence the amount and ratio of different caseins in the milk (Martin et al., 2002). Casein polymorphisms were found to affect milk processing and cheese making properties as well as the digestibility in human nutrition, hypoallergenic reactivity, and the risk of cardiovascular diseases and diabetes, for example (Caroli et al., 2009).

While many studies investigated the casein gene cluster in Holstein and other breeds (Ng-Kwai-Hang et al., 1984; Velmala et al., 1995; Formaggioni et al., 1999; Boettcher et al., 2004; Gallinat et al., 2013), so far only little is known about the genetic diversity of the casein cluster in DSN cattle. In a former study of β- and κ-casein variants in DSN cattle, homozygous carriers of the β-casein variant $A^2$ showed a tendency for higher milk, fat, and protein yield with lower fat and protein percentages, while κ-casein variants tended to have an influence on the protein percentage (Freyer et al., 1999). Since DSN has not been selected for protein variants in the recent past, but for other important traits such as milk yield and udder conformation, an indirect selection for specific casein variants could have happened as a by-product. Because of the close proximity of the four casein genes in the bovine genome, the casein genes are not inherited independently, but are often transmitted from parents to offspring as a single haplotype. Therefore, it is very useful to determine the frequency not only for single protein variants but also for each "comprehensive haplotype" made by building a haplotype out of protein variants found in the four casein genes using the sequential order in which these genes are found in the casein cluster. Such haplotypes for the casein gene cluster were described for many dairy breeds using sequence variation within coding regions (Ikonen et al., 2001; Caroli et al., 2003; Boettcher et al., 2004), in promoter regions (Jann et al., 2004; Ahmed et al., 2017) or microsatellites (Velmala et al., 1995). Some studies provided evidence for a correlation between casein haplotypes and milk yield, fat, and protein percentage (Velmala et al., 1995; Braunschweig et al., 2000; Ikonen et al., 2001; Boettcher et al., 2004; Braunschweig, 2008; Nilsen et al., 2009).

In the DSN cattle, the frequencies of single casein protein variants and casein protein haplotypes recently have been investigated by isoelectric focusing of milk samples ($N$ = 1,219) (Hohmann et al., 2018). In British Friesian, a breed that has similar ancestors and a similar breeding history like DSN, casein haplotypes were examined on the basis of genotype data ($N$ = 51) (Jann et al., 2004).

In the current study, we used whole-genome sequencing data of the DSN population and additional data from the 1000 Bull Genomes Project (Daetwyler et al., 2014; http://www.1000bullgenomes.com/) to examine and compare the sequence of all casein genes including the 1-kb upstream regulatory region. Our aim is to compare the DSN population with 13 other cattle breeds. This comparison is undertaken to investigate the genetic diversity of missense variants in the casein gene cluster across these cattle breeds and might provide selectable casein variants and/or haplotypes to improve DSN breeding.

**TABLE 1 |** Known protein variants

| Gene | Protein | Variants |
|---|---|---|
| *CSN1S1* | $\alpha_{S1}$ | *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J* |
| *CSN2* | β | $A^1$, $A^2$, $A^3$, *B*, *C*, *D*, *E*, *F*, *G*, $H^1$, $H^2$, *I*, *J*, *K*, *L* |
| *CSN1S2* | $\alpha_{S2}$ | *A*, *B*, *C*, *D*, *E* |
| *CSN3* | κ | *A*, *B*, *C*, *E*, $F^1$, $F^2$, $G^1$, $G^2$, *H*, *I*, *J* |

*In this table, we list all known variants for the casein genes published in recent literature for the CSN1S1, CSN2, CSN1S2, and CSN3 genes and their corresponding proteins in Bos genus (Ibeagha-Awemu et al., 2007; Caroli et al., 2010; Gallinat et al., 2013).*

# MATERIAL AND METHODS

## Sequencing Data

In order to characterize DSN casein sequence variants, the raw sequence variants of *Bos taurus* animals available from the 1000 Bull Genomes Project Run 6.0 were used (http://www.1000bullgenomes.com/; Daetwyler et al., 2014). Animals that shared high genetic similarity (>0.99 relative Manhattan distance; Korkuć et al., 2019), which could not be explained by kinship, were removed from the dataset. Furthermore, only breeds with at least 30 animals were selected for the analyses, so that the final dataset contained 14 different *B. taurus* breeds (30 DSN, 541 Holstein Friesian, 276 Angus, 217 Simmental, 148 Brown Swiss, 127 Charolais, 82 Limousin, 75 Hereford, 66 Jersey, 56 Danish Red, 54 Montbéliarde, 53 Fleckvieh, 52 Gelbvieh, and 44 Normande).

Filtering of raw SNP data was performed as described in Daetwyler et al. (2014), except we did not apply the proximity filter, which keeps only the highest quality SNP within 3 bp to increase the number of investigated SNPs in the casein cluster. In addition, we required at least three reads mapped to the reference and/or alternative allele to be considered a trustworthy SNP call; otherwise, the SNP genotype for that animal was set to missing. Only variants were investigated which are polymorphic in at least one breed.

The 30 DSN cattle in the 1000 Bull Genomes dataset were selected to best represent the current DSN population. The DSN population submitted includes 13 cows (mostly bull mothers) and 17 artificial insemination bulls. Due to the small population size, relationships between DSN cattle exist. Animal selection criteria for the other breeds from the 1000 Bull Genomes Project are not known.

## Investigated DNA Sequence Region

Genomic positions, reference genome, and protein sequences of the casein genes were obtained from Ensembl Release 93 (Zerbino et al., 2018) based on UMD3.1 assembly (Zimin et al., 2009). Sequence variants located within the casein genes *CSN1S1*, *CSN1S2*, *CSN2*, and *CSN3* (**Supplementary Table 1**) and 1,000 bp upstream were selected for analyses. The sequence variants were examined and categorized into variant types based on their genomic locations (1,000 bp upstream, 5′-UTR, intron, synonymous, missense, splice region, 3′-UTR) using the Ensembl Variant Effect Predictor (McLaren et al., 2016).

The lowest detectable allele frequency in DSN was 1/60 (0.017) as the minimum number of animals per breed was set to 30. So an allele frequency of 0.017 implies a single heterozygous animal within the population.

A comparison of the SNP annotation of the genes in the casein cluster to the rest of the genomic SNP was performed using all SNP variants annotated by the 1000 Bull Genomes Project (Hayes and Daetwyler, 2019). However, while our analysis of the casein cluster does not include intergenic variants, we recalculated the annotation percentages in the 1000 Bull dataset after removing the "intergenic variant" category. A comparison between the casein cluster and the rest of the genome can be found in **Supplementary Table 7**.

Haplotypes and haplotype frequency of protein-coding variants were estimated if at least two protein-coding variants were present. Haplotype analysis was performed using the function haplo.group from R package haplo.stats with the default settings (Sinnwell and Schaid, 2018). In order to assess the similarity of cattle breeds with regard to their haplotypes, Euclidean distances of protein variants and haplotype frequencies between all breeds were calculated. The resulting distance matrix was used to cluster (using average linkage) the cattle breeds hierarchically and to generate a dendrogram with standard R plot routines. All other plots were generated using the R package ggplot2 (Wickham et al., 2016).

Protein variants with a minimum frequency of 5% in a single breed were used to build comprehensive haplotypes across all four casein genes. Haplotypes are named according to the ordered position of the casein genes on the chromosome (*CSN1S1-CSN2-CSN1S2-CSN3*) and the variant name of each individual casein protein, e.g., *B-A$^1$-A-A* for *CSN1S1*B–CSN2*A$^1$–CSN1S2*A–CSN3*A*. This way of coding casein variants was proposed by Caroli et al.; more information about casein (haplotype) coding can be found in their 2009 paper (Caroli et al., 2009).

# RESULTS

## Distribution of DNA Sequence Variants in Casein Genes and Upstream Regions

In total, 892 SNPs were detected within the four casein genes (*CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3*) and their 1,000-bp upstream regions. Most of the detected variants were intron variants (87.3%), followed by variants in the 1,000-bp upstream gene region (5.8%), and missense variants (2.2%). Remaining SNPs were synonymous variants (1.2%), located in the 3′-UTR (2.2%), splice region (0.7%), or in the 5′-UTR region (0.4%) (**Table 2** and **Supplementary Table 2**).

Comparison of casein SNPs to the 1000 Bull Genomes whole-genome SNP dataset showed that the percentages detected in the casein cluster are similar to the whole-genome annotation frequencies (intron variants 84.7%, upstream region 11.4%, missense variants 1.4%, synonymous variants 1.4%, 3′-UTR 0.7%, splice region 0.2%, and 5′-UTR 0.2%) (**Supplementary Figure 1** and **Supplementary Table 7**).

SNP density was calculated for the average number of SNPs per 10 kb for upstream (+1,000 bp), intron and exon regions of the four casein genes (**Table 2**). The highest SNP density over all four genes was found in the introns (14.57 SNPs per 10 kb), followed by upstream gene regions (13.00 SNPs per 10 kb) and exons (6.22 SNPs per 10 kb). *CSN3* had the highest density of intronic DNA variants (17.44 SNPs per 10 kb) and exon regions (9.46 SNPs per 10 kb), while *CSN1S1* had the lowest SNP density in the exons (3.36), but the highest in the upstream region (22.00).

In DSN, 254 of 892 sequence variants over all four casein genes were detected (**Supplementary Table 3**). Six SNPs were found to be novel. This means that these SNPs were not found in the dbSNP and/or EVA database; this was investigated using the Ensemble genome browser (Release 93) which integrates both these databases. One in intron 6 of *CSN1S1* (BTA6:87147250 G/A) found in DSN and Holstein. One in intron 14 within the splice region of *CSN1S1* (BTA6:87155332 C/T) found in DSN, Holstein, and Fleckvieh. Another novel SNP that was found in intron 2 of *CSN3* (BTA6:87382140 T/C) was segregating in most of the investigated breeds. The alternative allele frequency (AAF) of this SNP is similar in DSN and Danish Red ($AAF_{(DSN)}$ = 28.3%, $AAF_{(Danish\ Red)}$ = 21.7%), while all other breeds showed an alternative allele frequency <10%. Interestingly, in *CSN2*, three novel SNPs were found in a single DSN bull only, one of

**TABLE 2 |** SNP density.

| Gene | Upstream | Intron | Exon | Missense | Synonymous |
|---|---|---|---|---|---|
| *CSN1S1* | 22.0 | 17.3 | 3.4 | 1.7 | 1.7 |
| *CSN2* | 10.0 | 15.6 | 8.7 | 6.1 | 2.6 |
| *CSN1S2* | 8.0 | 9.6 | 5.8 | 2.5 | 3.3 |
| *CSN3* | 12.0 | 17.4 | 9.5 | 8.3 | 1.2 |
| Total | 13.0 | 14.6 | 6.2 | 4.0 | 2.2 |

*SNP density per 10 kb in the upstream (+1,000 bp), intron and exon (split into missense and synonymous variants) regions of the casein genes CSN1S1, CSN2, CSN1S2, and CSN3.*

them in intron 1 (BTA6:87186177 G/A) and two in intron 4 (BTA6:87185025 T/A and BTA6:87184912 C/G).

The alternative allele frequency of all SNPs in the four casein genes differs between the investigated breeds. Through clustering of the 892 SNPs based on the respective alternative allele frequency per breed, distinct relationships between the breeds can be observed (**Figures 1** and **2**). The alternative allele frequencies of the sequence variants across all casein genes showed breed-specific differences. Overall, the alternative allele frequencies of DSN are most similar to those of Danish Red (dual-purpose breed), Holstein (milk production breed), and Hereford (beef production breed). DSN show very low

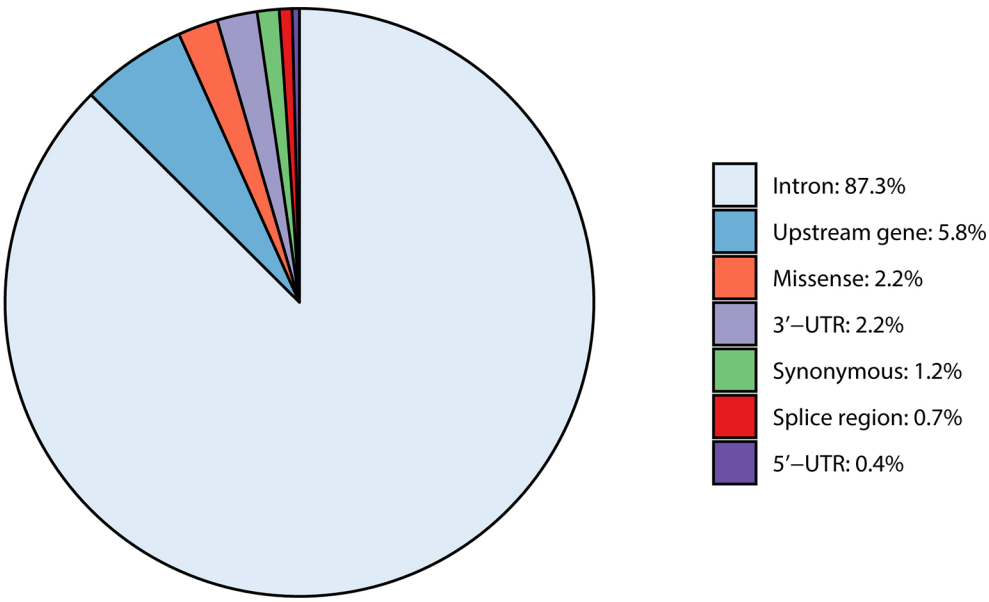


**FIGURE 1 |** Overview of variant types occurring within the four casein genes *CSN1S1*, *CSN1S2*, *CSN2*, and *CSN3* including their 1,000-bp upstream region.

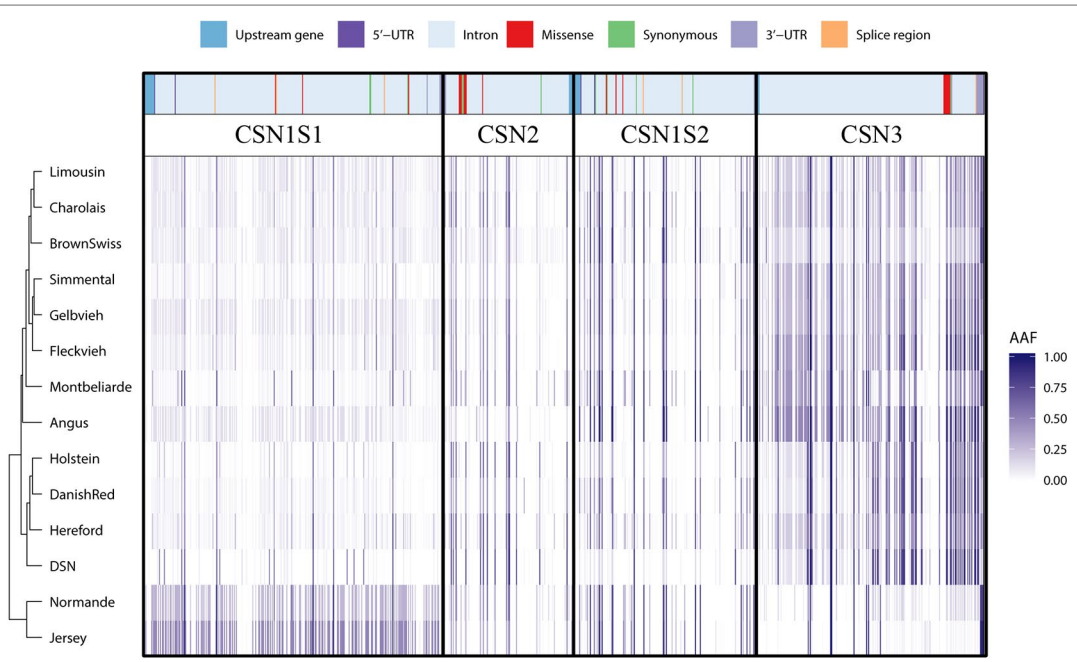


**FIGURE 2 |** Clustering of per-breed alternative allele frequency for the detected sequence variants in the casein genes *CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3* including their 1,000-bp upstream region. The respective variant types are presented above the alternative allele frequencies. It should be noted that the clustering is mainly based on intron variants (*light blue areas*) as they make up 87.3% of all detected variants.

alternative allele frequency for SNPs in *CSN1S1* and *CSN1S2*, but higher ones for SNPs in *CSN2* and *CSN3*. In contrast to all other breeds, Normande and Jersey had high and low alternative allele frequency in *CSN1S1* and *CSN3*, respectively. As such, these two breeds also cluster together on the lower side of the dendrogram (**Figure 2**). The relationship between all investigated breeds based on all genome-wide SNPs in the 1000 Bull Genomes Project showed a close relatedness between DSN, Holstein, and Danish Red (**Supplementary Figure 2**).

## Casein Protein Variants

### CSN1S1

Protein variants *CSN1S1*B* and *CSN1S1*C* were detected in at least one breed. In DSN, only the *CSN1S1*B* variant was detected (**Table 3**). Variants *CSN1S1*A* and *CSN1S1*C* were not observed among the 30 sequenced DSN animals. In Gelbvieh, Holstein, and Danish Red, the frequency of the *CSN1S1*C* variant was also low (<1%). In contrast, Limousin, Brown Swiss, and Fleckvieh had higher frequencies of the *CSN1S1*C* variant (>10%). The highest protein variant frequency of the *CSN1S2*C* variant was detected in the Jersey (44.8%) and Normande (25.6%) breeds (**Supplementary Table 4**).

### CSN2

Seven missense variants were found in the *CSN2* gene, of which five β-casein protein variants ($A^1$, $A^2$, *B*, *I*, and *F*) have a frequency of at least 5% in one breed. The distribution of those five most common β-casein protein variants differed in DSN compared to the other breeds. In DSN, the $A^1$ is the most common protein variant with a frequency of 82.7% compared to 30.0% in Holstein. The protein variants $A^2$ (15%) and *I* (2%) were found in DSN as well (**Table 3**). Variant *I* has not been described before for DSN (Jann et al., 2002; Caroli et al., 2009). The variants *B* and *F* were not detected in the examined DSN population, but were found in other breeds. Nine out of 14 breeds have a frequency of the $A^2$ variant of more than 50%, with the highest frequency in Angus (94.7%) (**Supplementary Table 5**).

### CSN1S2

In the *CSN1S2* gene, three missense variants were found which correspond to protein variants *CSN1S2*A*, *CSN1S2*C*, and *CSN1S2*D*. In DSN only variant *A* was detected (**Table 3**), similar to Jersey, Montbéliarde, Normande, Fleckvieh, and Hereford. Additionally, in Holstein, *CSN1S2*D* was found with low frequencies (0.3%). Gelbvieh has the highest frequency for variant *D,* with 12.2%. The highest frequency of the *C* variant was found in Angus, with a frequency of 7.5% (**Supplementary Table 4**).

### CSN3

Seven missense variants were found in the *CSN3* gene. The κ-casein variants *A*, *B*, and *E* have a frequency of at least 5% in one breed. In DSN, variant *A* is the most frequent (83.3%), followed by *B* (13.3%) and *E* (3.4%) (**Table 3**). *CSN3*A* is the most frequently detected variant in 10 out of the 14 breeds investigated. The highest frequency for the *B* variant was found in Jersey (96.0%), Brown Swiss (67.4%), Normande (84.6%), and Charolais (51.0%). The distribution of the *CSN3* protein variants in DSN are similar to Fleckvieh (*CSN3*A* = 84.4%, *CSN3*B* = 14.5%), although the *E* variant was not detected in Fleckvieh (**Supplementary Table 6**).

## Protein Haplotype Analysis Across the Casein Cluster

Across all casein genes, frequency of variants varied between the investigated breeds. Therefore, we performed a haplotype

**TABLE 3 |** Allele frequency of missense variants.

| Variant of casein gene | BTA position[a] | Allele | Amino acid | Protein seq. position[b] | SNP ID | Variant frequency | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | DSN | HF | All breeds |
| *CSN1S1*B* | 6:87157262 | **A**/G | **Glu**/Gly | 207 (192) | rs43703010 | 1.0 | 0.995 | 0.944 |
| *CSN2*A¹* | 6:87181619 | **T**/G | **His**/Pro | 82 (67) | rs43703011 | 0.827 | 0.340 | 0.295 |
| *CSN2*A²* | 6:87181619 | T/**G** | His/**Pro** | 82 (67) | rs43703011 | 0.156 | 0.562 | 0.592 |
| *CSN2*I* | 6:87181542 | T/**G** | Met/**Leu** | 108 (93) | rs109299401 | 0.017 | 0.059 | 0.036 |
| *CSN1S2*A* | 6:87266177 | **C**/T | **Ser**/Phe | 23 (8) | rs441966828 | 1.0 | 1.0 | 0.994 |
| *CSN3*A* | 6:87390576 | T/**C** | Ile/**Thr** | 157 (136) | rs43703015 | | | |
| | 6:87390612 | **C**/A | **Ala**/Asp | 169 (148) | rs43703016 | 0.833 | 0.752 | 0.628 |
| | 6:87390632 | **A**/G | **Ser**/Gly | 176 (155) | rs43703017 | | | |
| *CSN3*B* | 6:87390576 | **T**/C | **Ile**/Thr | 157 (136) | rs43703015 | 0.133 | 0.203 | 0.341 |
| | 6:87390612 | C/**A** | Ala/**Asp** | 169 (148) | rs43703016 | | | |
| *CSN3*E* | 6:87390632 | A/**G** | Ser/**Gly** | 176 (155) | rs43703017 | 0.034 | 0.045 | 0.030 |

*Allele frequencies of missense variants in CSN1S1, CSN2, CSN1S2, and CSN3 in DSN compared to Holstein Friesian (HF) and other breeds. For each variant, we list the alleles as ref/alt. In a bold font we highlight the SNP allele and resulting amino acid which causes the casein variant. As an example, the CSN2*A¹ and CSN2*A² variants are caused by a SNP on the same position 6:87181619. In the case of A¹, the T-allele causes a histidine to be incorporated into the protein sequence. The A² variant is defined as a G on the same position, leading to a proline in the resulting protein.*

*[a]Bos taurus autosome (BTA) CSN1S1*B (ENSBTAG00000007695), CSN2*A² (ENSBTAG00000002632), CSN1S2*A (ENSBTAG00000005005), and CSN3*A (ENSBTAG00000039787).*

*[b]Positions of amino acids according to the reference protein sequence from Ensembl Release 93 UMD3.1 assembly. Positions in the mature protein are given in parentheses.*

analysis across all protein variants of the four casein genes to position DSN relative to the other breeds.

Altogether, 37 haplotypes were constructed across all cattle breeds; 13 out of 37 haplotypes had a frequency higher than 5% in at least one breed. Out of the 13 haplotypes which met our inclusion criteria, five haplotypes showed a frequency >5%. For DSN, nine haplotypes could occur theoretically based on the number of casein protein variants across the casein cluster. Out of the expected haplotypes, seven were found. The most common haplotype in DSN was *B-A*$^{1}$*-A-A* with a frequency of 71.1%. In contrast to DSN, the most frequent haplotype in Holstein (53.1%) as well as in seven other breeds was *B-A*$^{2}$*-A-A* (**Table 4**).

Because of their similarity in their comprehensive haplotype distribution, DSN and Danish Red cattle clustered closely together (**Figure 3**). Both show the highest frequency for the *B-A*$^{1}$*-A-A* haplotype. Holstein clusters together with Hereford, Angus, Charolais, Fleckvieh, Gelbvieh, Limousin, and Simmental, which all show the highest frequency for the *B-A*$^{2}$*-A-A* haplotype. The breeds Brown Swiss (*B-A*$^{2}$*-A-B* = 50.0%), Montebéliarde (*B-A*$^{2}$*-D-B* = 35.9%), Jersey (*C-A*$^{2}$*-A-B* = 50.6%), and Normande (*C-A*$^{2}$*-A-B* = 28.4%) cluster together, showing the highest proportion of other haplotypes.

# DISCUSSION

## DNA Sequence Variants and New Alleles

Over the whole cattle genome, 0.6% of base pairs were polymorphic sequence variants in all breeds within the 1000 Bull Genomes Project (Sanchez et al., 2017). Within the casein cluster, we detected 0.4% of polymorphic sequence variants, which is an adequate result under consideration of the short region of about 250 kb on the bovine genome.

In the investigated casein region, intron variants are slightly more frequent with 87.3% in our study than in the whole cattle genome with an average of 84.7% (Hayes and Daetwyler, 2019). Upstream gene variants make up 11.4% of all SNPs in the whole cattle genome. In this study (1,000 bp upstream), only 5.8% of total SNPs were located in the upstream regions, which the authors suspect is due to the definition of what constitutes as "upstream." Missense variants are more frequent in the investigated casein region, with a proportion of 2.2% compared to the rest of the bovine genome (1.4%), which might point to more abundant genetic variation in the casein cluster compared to the whole genome. Overall, the casein cluster is very similar compared to the average cattle genome, with a few small deviations in the percentage of SNPs found in the upstream, missense, 3′-, and 5′-UTR as well as in splice sites.

In our analysis, we found 892 SNPs, of which 254 were present in DSN (28.4%). The allele frequencies across all SNPs clearly differentiate between the different cattle breeds. In upstream regulatory regions, no new variant was detected in DSN. Upstream variants in *CSN1S1*, *CSN2*, and *CSN3*, which might have regulatory effects on gene expression, have an allele frequency distribution in DSN similar to other breeds, and the allele frequencies of two variants in the upstream regions of *CSN1S2* are comparable to Danish Red. This is interesting

**TABLE 4** | Haplotype frequencies for the casein cluster *CSN1S1-CSN2-CSN1S2-CSN3* for all breeds.

| Haplotype | Total | Limousin | Angus | Hereford | Charolais | Simmental | Fleckvieh | Normande | Montbéliarde | Brown Swiss | Gelbvieh | Jersey | Danish Red | Holstein | DSN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N animals | 1821 | 82 | 276 | 75 | 127 | 217 | 53 | 44 | 54 | 148 | 52 | 66 | 56 | 541 | 30 |
| *B-A*$^{2}$*-A-A* | 0.424 | 0.419 | 0.644 | 0.359 | 0.331 | 0.446 | 0.415 | 0.100 | 0.072 | 0.115 | 0.490 | 0.045 | 0.136 | 0.531 | 0.116 |
| *B-A*$^{1}$*-A-A* | 0.147 | 0.071 | 0.032 | 0.313 | 0.041 | 0.145 | 0.270 | 0.046 | 0.055 | 0.078 | 0.091 | | 0.439 | 0.196 | 0.717 |
| *B-A*$^{2}$*-A-B* | 0.141 | 0.196 | 0.075 | 0.026 | 0.286 | 0.178 | 0.147 | 0.081 | 0.304 | 0.500 | 0.133 | 0.047 | 0.132 | 0.056 | 0.024 |
| *B-A*$^{1}$*-A-B* | 0.057 | 0.105 | 0.006 | 0.028 | 0.163 | 0.081 | | 0.062 | 0.014 | 0.020 | 0.114 | 0.124 | 0.031 | 0.063 | 0.093 |
| *B-B-A-B* | 0.051 | 0.038 | 0.002 | 0.238 | 0.013 | 0.009 | | 0.221 | | 0.079 | | 0.226 | 0.017 | 0.027 | |
| *C-A*$^{2}$*-A-B* | 0.048 | 0.094 | 0.026 | 0.007 | 0.033 | 0.032 | | 0.284 | | 0.069 | | 0.506 | 0.002 | 0.001 | |
| *B-I-A-B* | 0.035 | 0.029 | 0.007 | 0.007 | | 0.020 | | 0.162 | 0.151 | | | 0.052 | | 0.061 | 0.017 |
| *B-A*$^{2}$*-A-E* | 0.020 | | 0.105 | | | 0.003 | | | | | | | 0.028 | 0.002 | 0.013 |
| *C-A*$^{2}$*-A-A* | 0.019 | 0.018 | 0.020 | 0.023 | 0.028 | 0.049 | 0.117 | 0.020 | 0.034 | 0.051 | | | | 0.002 | |
| *B-B-A-A* | 0.017 | | | | 0.065 | 0.030 | 0.022 | | 0.359 | 0.083 | | | | 0.001 | |
| *B-A*$^{2}$*-D-B* | 0.004 | | | | | | | | | | 0.132 | | 0.054 | | |
| *B-F-A-A* | 0.005 | | | | | | | | | | | | 0.136 | 0.003 | |
| *B-A*$^{2}$*-C-A* | 0.011 | | 0.075 | | | | | | | | | | | | |
| *Residual* | 0.021 | 0.030 | 0.007 | | 0.040 | 0.006 | 0.030 | 0.023 | 0.011 | 0.005 | 0.040 | | 0.022 | 0.057 | 0.021 |

*Haplotypes with at least 5% in one breed are shown.*

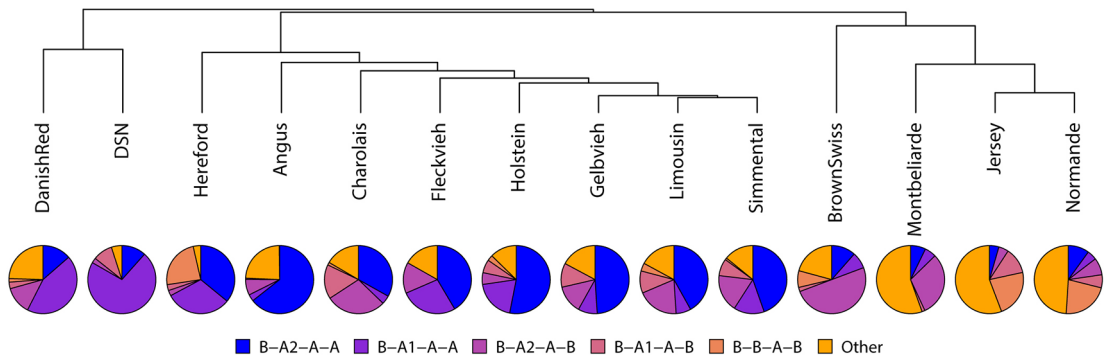


**FIGURE 3 |** Haplotype analysis across the casein proteins *CSN1S1-CSN2-CSN1S2-CSN3* for the five most common haplotypes listed using the respective protein variant names. Haplotypes with a total frequency less than 5% are summarized as "Other."

because DSN and Danish Red have similar breeding goals towards a dual-purpose phenotype and the breeds show similar fat and protein percentages in milk. As such, it could be proposed that the similarities in the *CSN1S2* upstream regions could be influencing the expression level of *CSN1S2* in both breeds, leading to similarities in the protein composition of the milk from these breeds. The expression level of the *CSN1S2* gene variant of DSN/Danish Red should be further investigated in comparison to other breeds.

Six new DNA variants were detected in the intronic regions of *CSN1S1*, *CSN2*, and *CSN3*. Three out of these six new DNA variants were detected in two different *CSN2* intron regions in a single DSN bull only. Because of the relatively stringent quality filter for sequencing data of at least three reads to one allele, we are reasonably confident that these three SNPs are real. However, a sequencing failure in this animal cannot be fully excluded. Three additional new SNPs that were detected in DSN and other breeds are reliable because of their frequencies and their occurrence in different breeds.

## Casein Protein Variants

No variation was detected in the two α-caseins in DSN. In the 30 sequenced animals, only the *CSN1S1*B* and *CSN1S2*A* variants were detected, while in Holstein the protein variants *CSN1S1*C* and *CSN1S2*D* were detected at low frequencies. However, since the investigated DSN population was small, we cannot exclude additional $\alpha_{S1}$ and $\alpha_{S2}$ protein variants; for example, variant *CSN1S1*C* has recently been detected in DSN (Hohmann et al., 2018).

In other breeds selected for high milk yield, the *CSN1S1*B* variant was reported to be fixed (Caroli et al., 2003). For DSN, which is a dual-purpose breed, *CSN1S1*B* is the only variant detected in our study. Interestingly, Jersey cattle, which were selected for high fat and protein content, showed the lowest frequency for *CSN1S1*B* (51.9%) and the highest frequency for *CSN1S1*C* (44.8%), which might mean a positive effect on protein and fat content for the *CSN1S1*C* variant. Since *CSN1S1*C* was recently detected in DSN (Hohmann et al., 2018), this might provide an opportunity for DSN breeders to increase the percentage of milk fat and protein in DSN by actively searching for and breeding with animals carrying the *CSN1S1*C* variant.

The $A^1$ variant of the β-casein has a frequency of 82.9% in DSN, which is much higher than in other breeds. Compared to earlier results from the DSN population, an overestimation of this variant (DSN Brandenburg *CSN2*A*$^1$ = 67% frequency; Hohmann et al., 2018) could result from the small sample size in our data. This overestimation goes probably to the disadvantage of the β-casein variant $A^2$, which we only detected by a frequency of 15.4% in DSN (DSN Brandenburg *CSN2*A*$^2$ = 31% frequency; Hohmann et al., 2018). The *I* variant of β-casein showed a frequency of 1.7% in our DSN population. While all casein variants that occur in DSN were also found in Holstein, the reverse situation is not true.

Since our study used SNPs to predict protein variants, we are not able to detect some known casein variants which can only be found using protein analysis. As an example, our study is unable to estimate the occurrence of *CSN2*C* since the dephosphorylation of $Ser_{35}P$ into a unphosphorylated Ser in *CSN2* happens posttranslational and can only be investigated at the protein molecule level (Gallinat et al., 2013). Other studies on the DSN population show the existence of the *CSN2*B* variant with low frequencies (DSN Brandenburg *CSN2*B* = 2% frequency; Hohmann et al., 2018). In further investigations, the sequence on protein level should be examined parallel to the DNA sequence.

With a frequency of 83.2%, the *A* variant of κ-casein is the most common in DSN, followed by *CSN3*B* (13.3%) and *CSN3*E* (3.5%). The variant frequencies agree with previous findings by Hohmann and colleagues (Hohmann et al., 2018). In contrast to Holstein, no additional κ-casein protein variant could be found in DSN. The *E* variant, which influences cheese making properties in a presumably negative way (Caroli et al., 2009), was detected in six breeds including DSN at a low frequency. A low frequency is also occurring in Holstein (4.6%) and Danish Red (3.6%). However, increasing the *E* variant in the population should be selected against in DSN.

## Casein Haplotype Frequencies in DSN Compared to Other Breeds

In DSN, *B-A*$^1$*-A-A* is the most frequent casein haplotype with a frequency of 71.7%. This is due to the very high frequency of

*CSN2*A¹* (82.9%), which might be overestimated in our results. Studies with higher sample sizes showed similar results. Also, they detected the highest frequency (57%) for the shortened *CSN1S1*B–CSN2*A¹–CSN3*A* haplotype in DSN (Hohmann et al., 2018). The 57% estimate should be considered the more reliable estimate as it is based on a larger sample size. The most common comprehensive casein haplotype in British Friesian was also *B-A¹-A-A*, with a frequency of 60% (Jann et al., 2004), which is similar to the frequency found in DSN. In contrast to DSN, the protein variants *CSN2*I* and *CSN3*E* were not detected in British Frisian.

The haplotype *B-A²-A-A* is the most common in Holstein (53.1%) and several other *B. taurus* breeds (Limousin 41.9%, Angus 64.4%, Hereford 35.9%, Charolais 33.1%, Simmental 44.6%, Fleckvieh 41.5%, and Gelbvieh 49.0%), and the estimated frequencies of the casein protein variants reported in this paper are comparable to frequencies found in the literature, e.g., for Aberdeen Angus (51.1%) (Jann et al., 2004) or Italian Holsteins (*CSN1S1*B-CSN2*A²-CSN3*A* = 48%) (Boettcher et al., 2004). For Brown Swiss, the haplotype *B-A²-A-B* with a frequency of 50% is identical to results in the literature for the shortened haplotype *CSN1S1*B-CSN2*A²-CSN3*B* in Italian Brown Swiss (Boettcher et al., 2004). The cattle populations within the 1000 Bull Genomes Project seem to adequately represent the respective cattle breeds.

Further investigation should investigate the effect of different haplotypes in DSN on milk yield and protein and fat percentage. However, the current sample size would not lead to significant results. A previous investigation of casein variants with >600 DSN found no significant results based on the β- and κ-casein genotype (Freyer et al., 1999).

## CONCLUSION

Few of the already known casein protein variants, $\alpha_{S1}$ (*B*), β (*A¹*, *A²*, and *I*), $\alpha_{S2}$ (*A*), and κ (*A*, *B*, and *E*), were detected in DSN using whole-genome sequencing data. This study is the first to find the *CSN2*I* variant in DSN. Besides the detection of this new variant, we confirm previous findings by Hohmann and colleagues that the most common casein cluster haplotype in DSN is *B-A¹-A-A*. Based on the casein haplotype, DSN clusters together with Danish Red.

DSN cattle is remarkably different from the other investigated *B. taurus* breeds by having a high frequency of the *CSN2*A¹* variant. The preferred protein variants *CSN2*A²* for potentially improving human health and *CSN3*B* for better cheese making properties were detected at low frequencies in the DSN breed. Our study found a large and untapped potential for DSN breeders to select and increase beneficial protein variants. However, selection for these variants could also (negatively) influence other important traits (e.g., protein and fat percentage or milk yield).

Because of its low variability, the $\alpha_{S2}$ protein is often omitted from casein studies. In our study of 14 breeds, we also come to the same conclusion that variability in $\alpha_{S2}$ is low and can be disregarded when investigating protein variants. However, we found a number of upstream genetic variations which show a similarity between the dual-purpose breeds DSN and Danish Red. These upstream variants might influence expression of the *CSN1S2* gene and should be investigated further.

## DATA AVAILABILITY STATEMENT

All data required to reproduce the analysis, results, and conclusions can be requested from the authors at this point in time, since access to the 1000 Bull Genomes data is currently only available to partners. However, the 1000 Bull Genomes consortium will make the whole genome sequencing data available publicly when data collection and analysis is completed.

## ETHICS STATEMENT

Ethical review and approval was not required for the animal study because samples are collected based on routine procedures on these farm animals. Ear tags were taken as part of the required registration procedure, blood samples were taken by a trained veterinarian to perform standard health recording. Semen from bulls was acquired under routine conditions as part of the normal operation of RBB as an artificial insemination company.

## AUTHOR CONTRIBUTIONS

SM, PK, DA, and GB designed the study. SM interpreted the data and drafted the manuscript. PK performed all computational and statistical analysis. DA, PK, and GB helped draft the manuscript. All authors read and approved the final manuscript.

## FUNDING

This project is funded by the German Federal Ministry of Food and Agriculture (BLE) and Federal Program of Ecological Agriculture (BÖLN) (funding number 2815NA010). We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## ACKNOWLEDGMENTS

The authors wish to acknowledge the 1000 Bull Genomes Consortium for providing the data. RBB Rinderproduktion Berlin-Brandenburg GmbH and the associated DSN farms supported the project with their expertise in animal's selection, supply of semen doses, and collecting ear tag samples of cows during farm management routine.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01129/full#supplementary-material

## REFERENCES


Ahmed, A. S., Rahmatalla, S. A., Bortfeldt, R. H., Arends, D., Reissmann, M., and Brockmann, G. A. (2017). Milk protein polymorphisms and casein haplotypes in Butana cattle. *J. Appl. Genet.* 58, 261–271. doi: 10.1007/s13353-016-0381-2

Boettcher, P. J., Caroli, A., Stella, A., Chessa, S., Budelli, E., Canavesi, F., et al. (2004). Effects of casein haplotypes on milk production traits in italian holstein and brown swiss cattle. *J. Dairy Sci.* 87 (12), 4311–4317. doi: 10.3168/jds.S0022-0302(04)73576-6

Braunschweig, M. H. (2008). Associations between 2 paternal casein haplotypes and milk yield traits of Swiss Fleckvieh cattle. *J. Appl. Genet.* 49, 69–74. doi: 10.1007/BF03195250

Braunschweig, M., Hagger, C., Stranzinger, G., and Puhan, Z. (2000). Associations between casein haplotypes and milk production traits of swiss brown cattle. *J. Dairy Sci.* 83, 1387–1395. doi: 10.3168/jds.S0022-0302(00)75007-7

Caroli, A. M., Chessa, S., and Erhardt, G. J. (2009). Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. *J. Dairy Sci*. 92 (11), 5335–5352. doi: 10.3168/jds.2009-2461

Caroli, A., Bolla, P., Vivona, G., and Gandini, G. (2003). Milk protein polymorphisms in the Reggiana cattle. *Ital. J. Anim. Sci.* 2, 52–54. doi: 10.4081/ijas.2003.11675912

Caroli, A., Rizzi, R., Lühken, G., and Erhardt, G. (2010). Short communication: Milk protein genetic variation and casein haplotype structure in the Original Pinzgauer cattle. *J. Dairy Sci.* 93, 1260–1265. doi: 10.3168/jds.2009-2521

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

Ferretti, L., Leone, P., and Sgaramella, V. (1990). Long range restriction analysis of the bovine casein genes. *Nucleic Acids Res.* 18 (23), 6829–6833. doi: 10.1093/nar/18.23.6829

Formaggioni, P., Summer, A., Malacarne, M., and Mariani, P. (1999). Milk protein polymorphism: detection and diffusion of the genetic variants in bos genus. *Ann. della Fac. di Med. Vet.*, 127–165.

Freyer, G., Liu, Z., Erhardt, G., and Panicke, L. (1999). Casein polymorphism and relation between milk production traits. *J. Anim. Breed. Genet.* 116, 87–97. doi: 10.1046/j.1439-0388.1999.00181.x

Gallinat, J. L., Qanbari, S., Drögemüller, C., Pimentel, E. C. G., Thaller, G., and Tetens, J. (2013). DNA-based identification of novel bovine casein gene variants. *J. Dairy Sci.* 96, 699–709. doi: 10.3168/jds.2012-5908

Grothe, P. O. (1993). *Holstein-Friesian, eine Rasse geht um die Welt*. Münster-Hiltrup: Landwirtschaftsverlag GmbH.

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

Hohmann, L., Scheper, C., Erhardt, G., König, S., and Weimann, C., *Diversitätsanalyse boviner Milchproteinpolymorphismen unter besonderer Berücksichtigung von Selektionslinien. in Vortragstagung der DGfZ und GfT am 12./13.* p. A15, 2018, September 2018 in Bonn.

Ibeagha-Awemu, E. M., Prinzenberg, E. M., Jann, O. C., Luhken, G., Ibeagha, A. E., Zhao, X., et al. (2007). Molecular characterization of bovine CSN1S2*B and extensive distribution of zebu-specific milk protein alleles in European cattle. *J. Dairy Sci.* 90, 3522–3529. doi: 10.3168/jds.2006-679

Ikonen, T., Bovenhuis, H., Ojala, M., Ruottinen, O., and Georges, M. (2001). Associations between casein haplotypes and first lactation milk production traits in finnish ayrshire cows. *J. Dairy Sci.* 84, 507–514. doi: 10.3168/jds.S0022-0302(01)74501-8

Jann, O. C., Ibeagha-Awemu, E. M., Özbeyaz, C., Zaragoza, P., Williams, J. L., Ajmone-Marsan, P., et al. (2004). Geographic distribution of haplotype diversity at the bovine casein locus. *Genet. Sel. Evol.* 36, 243–257. doi: 10.1186/1297-9686-36-2-243

Jann, O., Ceriotti, G., Caroli, A., and Erhardt, G. (2002). A new variant in exon VII of bovine beta-casein gene (CSN2) and its distribution among European cattle breeds. *J. Anim. Breed. Genet.* 119, 65–68. doi: 10.1046/j.1439-0388.2002.00318.x

Köppe-Forsthoff, J. (1967). *100 Jahre Deutsche Schwarzbuntzucht*. Hiltrup Landwirtschaftsverlag.

Korkuć, P., Arends, D., and Brockmann, G. A. (2019). Finding the optimal imputation strategy for small cattle populations. *Front. Genet.* 10, 1–10. doi: 10.3389/fgene.2019.00052

Martin, P., Szymanowska, M., Zwierzchowski, L., Leroux, C., and Martin, P. (2002). The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reprod. Nutr. Dev.* 42, 433–459. doi: 10.1051/rnd:2002036

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0974-4

Ng-Kwai-Hang, K. F., Hayes, J. F., Moxley, J. E., and Monardes, H. G. (1984). Association of genetic variants of casein and milk serum proteins with milk, fat, and protein production by dairy cattle. *J. Dairy Sci.* 67, 835–840. doi: 10.3168/jds.S0022-0302(84)81374-0

Nilsen, H., Olsen, H. G., Hayes, B. J., Sehested, E., Svendsen, M., Nome, T., et al. (2009). Casein haplotypes and their association with milk production traits in Norwegian Red cattle. *Genet. Sel. Evol.* 41, 1–12. doi: 10.1186/1297-9686-41-24

Rinderproduktion Berlin-Brandenburg GmbH, R. B. B. (2016). *Deutsches Schwarzbuntes Niederungsrind - lebendes Kulturerbe*. Available at: https://www.rinderzucht-bb.de/zucht/dsn-genreserve/

Sanchez, M. P., Govignon-Gion, A., Croiseau, P., Fritz, S., Hozé, C., Miranda, G., et al. (2017). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet. Sel. Evol.* 49, 1–16. doi: 10.1186/s12711-017-0344-z

Sinnwell, J., and Schaid, D. (2018). *Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous*. Available at: https://CRAN.R-project.org/package=haplo.stats

Threadgill, D. W., and Womack, J. E. (1990). Genomic analysis of the major bovine milk protein genes. *Nucleic Acids Res.* 18, 6935–6942. doi: 10.1126/science.1164266

Velmala, R., Vilkki, J., and Mäki-Tanila, A. (1995). Casein haplotypes and their association with milk production traits in the Finnish Ayrshire cattle. *Anim. Genet.* 26 (6), 419–425. doi: 10.1111/j.1365-2052.1995.tb02694.x

Wickham, H. (2016). ggplot2: *Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol.* 10, R42. doi: 10.1186/gb-2009-10-4-r42


**Conflict of Interest:** SM is an employee of the RBB Rinderproduktion Berlin-Brandenburg GmbH, a cattle breeders association which produces semen and provides services to DSN and Holstein cattle breeders in the Berlin/Brandenburg area.

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: 14 November 2019
doi: 10.3389/fgene.2019.01095

# Identification of Candidate Signature Genes and Key Regulators Associated With Trypanotolerance in the Sheko Breed

*Yonatan Ayalew Mekonnen[1], Mehmet Gültas[1,2], Kefena Effa[3], Olivier Hanotte[4,5] and Armin O. Schmitt[1,2]**

[1] Breeding Informatics Group, Department of Animal Sciences, University of Göttingen, Göttingen, Germany, [2] Center for Integrated Breeding Research (CiBreed), University of Göttingen, Göttingen, Germany, [3] Animal Biosciences, National Program Coordinator for African Dairy Genetic Gain, International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia, [4] Cells, Organisms amd Molecular Genetics, School of Life Sciences, University of Nottingham, Nottingham, United Kingdom, [5] LiveGene, International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia

OPEN ACCESS

**Edited by:**
Gábor Mészáros,
University of Natural Resources and Life Sciences Vienna, Austria

**Reviewed by:**
Nina Moravčíková,
Slovak University of Agriculture, Slovakia
Kieran G. Meade,
The Irish Agriculture and Food Development Authority, Ireland
John B. Cole,
United States Department of Agriculture (USDA), United States

***Correspondence:**
Armin O. Schmitt
armin.schmitt@uni-goettingen.de

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

**Received:** 03 June 2019
**Accepted:** 11 October 2019
**Published:** 14 November 2019

**Citation:**
Mekonnen YA, Gültas M, Effa K, Hanotte O and Schmitt AO (2019) Identification of Candidate Signature Genes and Key Regulators Associated With Trypanotolerance in the Sheko Breed. Front. Genet. 10:1095. doi: 10.3389/fgene.2019.01095

African animal trypanosomiasis (AAT) is caused by a protozoan parasite that affects the health of livestock. Livestock production in Ethiopia is severely hampered by AAT and various controlling measures were not successful to eradicate the disease. AAT affects the indigenous breeds in varying degrees. However, the Sheko breed shows better trypanotolerance than other breeds. The tolerance attributes of Sheko are believed to be associated with its taurine genetic background but the genetic controls of these tolerance attributes of Sheko are not well understood. In order to investigate the level of taurine background in the genome, we compare the genome of Sheko with that of 11 other African breeds. We find that Sheko has an admixed genome composed of taurine and indicine ancestries. We apply three methods: (i) The integrated haplotype score (*iHS*), (ii) the standardized log ratio of integrated site specific extended haplotype homozygosity between populations (*Rsb*), and (iii) the composite likelihood ratio (CLR) method to discover selective sweeps in the Sheko genome. We identify 99 genomic regions harboring 364 signature genes in Sheko. Out of the signature genes, 15 genes are selected based on their biological importance described in the literature. We also identify 13 overrepresented pathways and 10 master regulators in Sheko using the TRANSPATH database in the geneXplain platform. Most of the pathways are related with oxidative stress responses indicating a possible selection response against the induction of oxidative stress following trypanosomiasis infection in Sheko. Furthermore, we present for the first time the importance of master regulators involved in trypanotolerance not only for the Sheko breed but also in the context of cattle genomics. Our finding shows that the master regulator Caspase is a key protease which plays a major role for the emergence of adaptive immunity in harmony with the other master regulators. These results suggest that designing and implementing genetic intervention strategies is necessary to improve

the performance of susceptible animals. Moreover, the master regulatory analysis suggests potential candidate therapeutic targets for the development of new drugs for trypanosomiasis treatment.

**Keywords: trypanosomiasis, trypanotolerant, selection signature, candidate signature genes, master regulators, overrepresented pathways**

## INTRODUCTION

Trypanosomiasis is a disease caused by uni-cellular protozoan parasites which affects the health of humans and livestock. In Africa, this disease is referred to as African animal trypanosomiasis (AAT) (Kristjanson et al., 1999; Shaw et al., 2014). AAT is the major livestock production constraint especially in sub-Saharan African countries. It is mainly caused by *Trypanosoma congolense*, *Trypanosoma vivax*, and *Trypanosoma brucei brucei* (Hoare, 1972; Abebe, 2005, Batista et al., 2011; Yaro et al., 2016). Particularly, *T. congolense* is the most frequent cause of livestock disease in this region (Naessens, 2006). The disease is transmitted from infected animals to healthy animals by tsetse fly as a vector (Welburn et al., 2016). The infected animal shows symptoms such as anemia (Murray et al., 1990; Naessens, 2006), neurological symptoms (Tuntasuvan et al., 1997; Giordani et al., 2016), reduced productivity, infertility, abortion (Barrett and Stanberry, 2009), listlessness, and emaciation (Nantulya, 1986; Batista et al., 2007; Steverding, 2008; Noyes et al., 2011). If not treated, it can lead to death (Kristjanson et al., 1999; Barrett and Stanberry, 2009; Giordani et al., 2016). Hence, this disease has a major economic impact that accounts for an estimated annual loss of US$ 5 billion in sub-Saharan countries (Kristjanson et al., 1999; Giordani et al., 2016).

Ethiopia is located in the eastern part of the tsetse belt. The tsetse fly distribution in the country spans from the south western to the north western regions covering 22,000 km$^2$) between longitude 38° and 38° East and latitude 5° and 12° North along river basins (Andrew, 2004; NTTICC, 2004). About 14 million cattle, 7 million horses, 1.8 million camels, and 14 million small ruminants are kept in the infection zone (MoARD, 2004). AAT severely affects the draft power as well as meat and milk production of the animals (Chanie et al., 2013). Therefore, AAT is considered as a major challenge constraining the path toward ensuring food security and combating poverty in this region (Meyer et al., 2018).

Until now, a number of methods have been applied to control the spread of this disease such as trypanocidal drugs, insect traps, and insecticides (Slingenbergh, 1992; Leak et al., 1996; Giordani et al., 2016). But none of these controlling measures has been successful to eradicate the disease. The current situation is deteriorating because of the trypanocidal drug resistance due to inappropriate drug usage. Moreover, pharmaceutical companies are less attracted to invest in new drug discovery and development due to high cost (Codjia et al., 1993; Mulugeta et al., 1997; Kristjanson et al., 1999; Naula and Burchmore, 2003). To control the spread of this disease, Lutje et al. (1996) have suggested a cross breeding strategy between trypanotolerant and trypanosusceptible cattle, together with vector control. Accordingly, Hanotte et al. (2003) performed crossbreeding between the trypanotolerant N'Dama and trypanosusceptible Boran breeds to produce an $F_2$ population that shows heterosis. This led to the assumption that an $F_2$ cross between trypanotolerant and susceptible breeds could produce a trypanotolerant synthetic breed whose performance would exceed that of either parent. Consequently, marker assisted selection from the $F_2$ breed would be the most promising strategy to produce a breed that combines high production and trypanotolerance (Hanotte et al., 2003; Noyes et al., 2011).

In Ethiopia, Sheko shows better trypanotolerance attributes than other breeds such as Abigar and Horro (Lemecha et al., 2006). Sheko is found in the southern region of the Bench Maji Zone, the adjoining areas of Keffa and Shaka and is considered as an endangered breed due to extensive interbreeding with local indicine and sanga breeds (DAGRIS, 2007). Sheko cattle are kept in the tsetse infested regions likely explaining their degree of trypanotolerance (Hanotte et al., 2003; Bahbahani et al., 2018). In order to address the tolerance attributes of the Sheko breed at the molecular level, this study analyzes the genotyping data of the breed to explore the genome for candidate signature genes. The rationale is that natural or artificial selection targets the genome in response to environmental pressures or stresses as shaping adaptation and evolution. This implies that if the new allele of a mutation is beneficial (increases the fitness of their carriers) under certain environmental pressure or stress, then the frequency of these alleles will rapidly increase in the population (Charlesworth, 2007). Under positive selection, strong and long range linkage disequilibrium (LD) and unexpectedly high local haplotype homozygosity might occur in the genome (Gautier and Vitalis, 2012; Bomba et al., 2015).

Likewise, trypanosomiasis is considered as an environmental pressure which plays a major role to create selection signatures in the genome and which is thus leading to breed formation (Kristjanson et al., 1999; Abebe, 2005; Yaro et al., 2016). These signs or traces of selection in the genome could be detected by using a "bottom-up" or a "from genotype to phenotype" approach (McGuire and McGuire, 2008). This study provides traces or signs of positive selection in the genome of Sheko against trypanosomiasis using the "bottom-up" approach. In response to trypanosomiasis as the environmental pressure, the genome of Sheko could undergo changes at the molecular level. With the aim to identify the molecular mechanism of Sheko tolerance, we use extended haplotype homozygosity (EHH; *iHS* and *Rsb*) and spatial distribution of allele frequency [composite

likelihood ratio (CLR)] based methods to identify genes that are associated with this selection pressure in the Sheko breed. Combining methods for the detection of selection signature regions has been suggested as a means of increasing the power of the study compared to single analysis (e.g. Ma et al., 2015; Vatsiou et al., 2016).

## Summary of the Analysis Workflow

Our workflow can be divided into two major steps as described below (see also **Figure 1**): 1) We analyzed the genetic relationship and structure of Sheko and 11 other indigenous African breeds using Plink 1.9 and the ADMIXTURE 1.3 software. 2) The identified candidate signature genes were then used in the analysis pipeline comprising the following four sub-steps: i) First, we identified genomic regions and signature genes under positive selection toward trypanotolerance in Sheko using *iHS*, CLR and *Rsb* analyses. As an intermediate result, we present the 15 genes resulting from a literature survey; ii) in the second step, we applied enrichment analysis in gene ontology (GO) terms in the combined gene sets of the three methods and made clusters of enriched GO terms in the form of a treemap using the geneXplain platform; iii) we then identified overrepresented pathways based upon the significant genes found in (ii) using the TRANSPATH database in the geneXplain platform; iv) finally, we identified the master regulators 10 steps upstream in the regulatory hierarchy using the significant genes found in (ii) using the TRANSPATH database in the geneXplain platform.

# RESULT AND DISCUSSION

## The Genetic Relationship and Structure of Cattle Populations

In order to understand the genetic structure of Sheko in comparison with 11 other African breeds, principal component analysis (PCA) was used. The result shows that the first two principal components (PCs), which explain 30.3% and 4.6% of the total variation, distinguishes the African taurine (N'Dama and Muturu) from the African indicine breeds [Benshangul, Serere, Karamojong, East African Shorthorn Zebu (EASZ), Fogera, and Gindeberet] (**Figure 2A**). Moreover, the Sheko, Nganda, Ankole, and Nuer are positioned between the African taurine and the African indicine clusters. These breeds are close to the indicine cluster and thereby support the admixture of more indicine than taurine type genomes in these breeds. The PCA result also shows the highest level of genetic heterogeneity in the Nganda breed which might be caused by ongoing crossbreeding of Nganda with exotic breeds to enhance their productivity (Mwai et al., 2015). We also conducted PCA exclusively for indigenous Ethiopian breeds. The result shows that the Sheko and Nuer form separate groups while the indicine type breeds (Benshangul, Fogera and Gindeberet) form a cluster in both PCs (**Figure 2B**).

For the further understanding of the degree of admixture in the populations, the ADMIXTURE 1.3 (Alexander et al., 2009) software was used for K = 2 to 7 hypothetical ancestral populations (**Figure 3**). We start from two hypothetical ancestral populations with the aim to determine the degree of indicine and taurine genetic background in the cattle breeds.

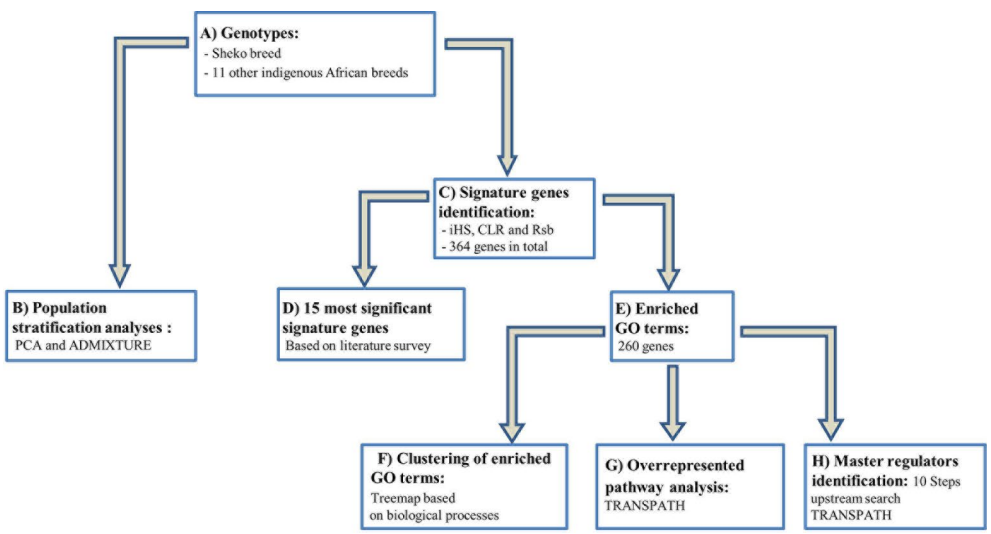


**FIGURE 1 |** Workflow for the study to identify candidate genes and key regulators that are associated with trypanotolerance in Sheko breed. **(A)** The genotypes of the Sheko and 11 other indigenous African breeds are obtained and quality control filtering is performed. **(B)** The genomic structure of Sheko in comparison to 11 other indigenous African breeds is analyzed using principal component analysis (PCA) and ADMIXTURE. **(C)** The identification of 364 signature genes is performed by *iHS*, CLR, and *Rsb* analyses. **(D)** Among 364 genes, the 15 most significant genes that are associated with trypanotolernace attributes are identified and disclosed. **(E)** Significantly functionally enriched terms [gene ontology (GO) terms] are identified for the 364 signature genes. 260 genes are identified as significantly enriched. **(F)** Using the functionally enriched 260 genes, a treemap is produced based on the biological processes. **(G)** Functionally enriched signature genes (260 genes) are analyzed to identify overrepresented pathways. **(H)** A master regulator network is generated up to 10 steps upstream using functionally enriched signature genes. The treemap, overrepresented pathway, and master regulator analyses were performed in the geneXplain platform.

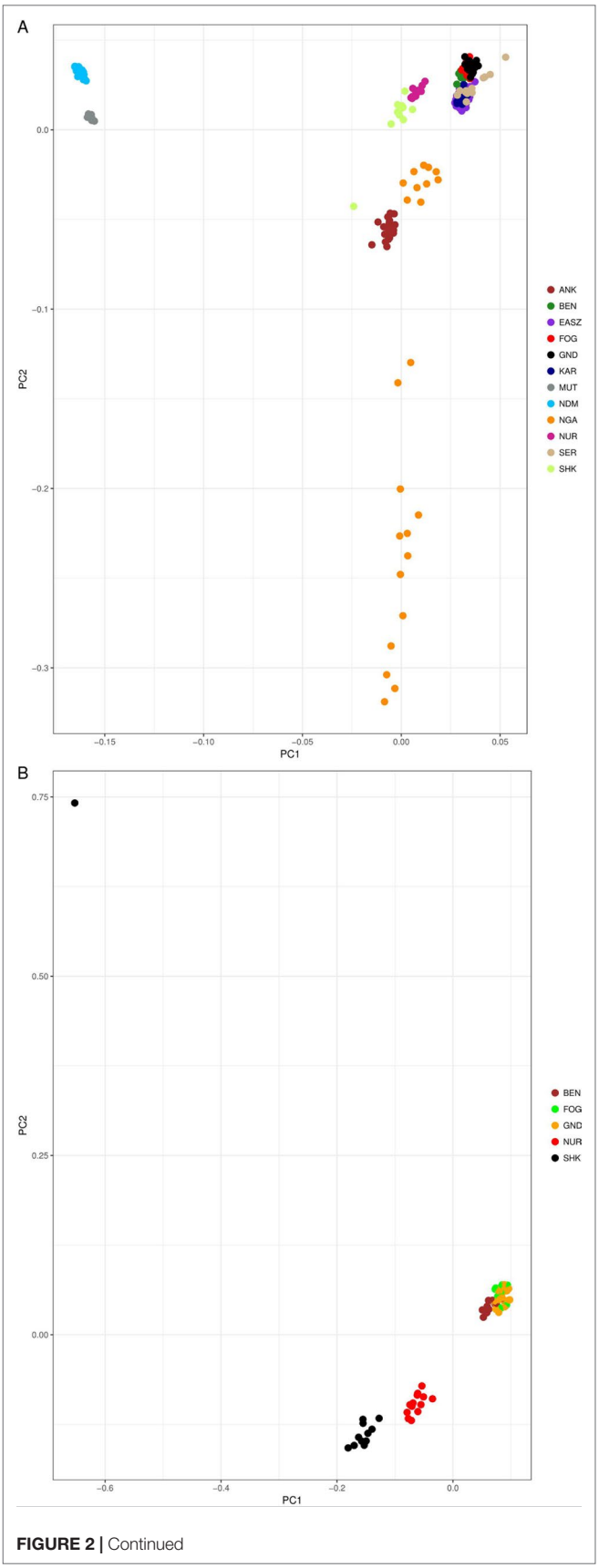


**FIGURE 2 |** Continued

**FIGURE 2 |** PCA plots of the first two principal components showing the genetic relationship between cattle breeds. **(A)** PCA plot for all cattle breeds included in this study, and **(B)** PCA plot for the Ethiopian cattle breeds. ANK, Ankole; BEN, Benshangul; FOG, Fogera; GND, Gindeberet; KAR, Karamojong; MUT, Muturu; NDM, N'Dama; NGA, Nganda; NUR, Nuer; SER, Serere; SHK, Sheko.

Since the CV errors from K = 3 to K = 6 have not exceeded the cross-validation (CV) errors of K = 2, we extend the hypothetical population up to K = 7 which has the highest CV error (**Supplementary Figure 1**). At K = 2, the two ancestries taurine and indicine are revealed. The genomes of Ankole, Nganda, Nuer, and Sheko are mainly of indicine origin but have substantial taurine admixture, a result supporting our interpretation of the first PC of **Figure 2A**, that African taurine are separated from the East African indicine breeds and the mixed taurine-indicine type populations. At K = 3, Ankole, Nuer and Sheko show genetic heterogeneity with a considerable level of taurine admixture. EASZ, Karamojong, Serere, Benshangul, Fogera, and Gindeberet also show minor levels of taurine admixture whereas Nganda reveals a high level of within breed genetic differentiation. This is also in agreement with the second PC coordinate analysis in showing genetic heterogeneity within the cattle breeds (**Figure 2A**). Moreover, with the increment of the value of K, Sheko and Nuer show a higher level of genetic heterogeneity than the other east African breeds. Furthermore, at K = 6 and K = 7, the African taurine breeds N'Dama and Muturu show separate genetic backgrounds. In general, Sheko shows the highest level of African taurine genomic contribution for all values of K among East African breeds. The proportions of admixture in each of the analyzed breeds are presented for K = 7 in **Supplementary Table 1**.

Consistent with the previous findings and the origins of the genetic backgrounds of the cattle breeds worldwide (Mbole-Kariuki et al., 2014; Bahbahani et al., 2018), K = 2 highlights best the ancient divergence between indicine and taurine cattle. However, the three optimal genetic clusters suggested by the minimal CV error (**Supplementary Figure 1**) reflect the common genetic background unique to East Africa besides taurine and indicine ancestral genetic admixture. In agreement with our study, Bahbahani et al. (2018) reported east African genetic background unique to East African cattle breeds. Moreover, the admixture plots show two individuals of Sheko with a high level of taurine introgression. One of these individuals with higher taurine introgression is also detected by the PCA (**Figure 1B**, upper left corner). This could be due to the recent crossbreeding of Sheko with European dairy breeds. There were similar observations in Butana, and it was speculated that farmers might have been involved in the crossbreeding with European dairy breeds in order to increase milk production (Bahbahani et al., 2018). We believe that the introgression of the European dairy breeds into the genome of indigenous breeds such as Sheko and Butana might distort their adaptive evolutionary responses against their natural environmental stresses. In this regard, future studies should assess the impact of European dairy breeds on the genome of

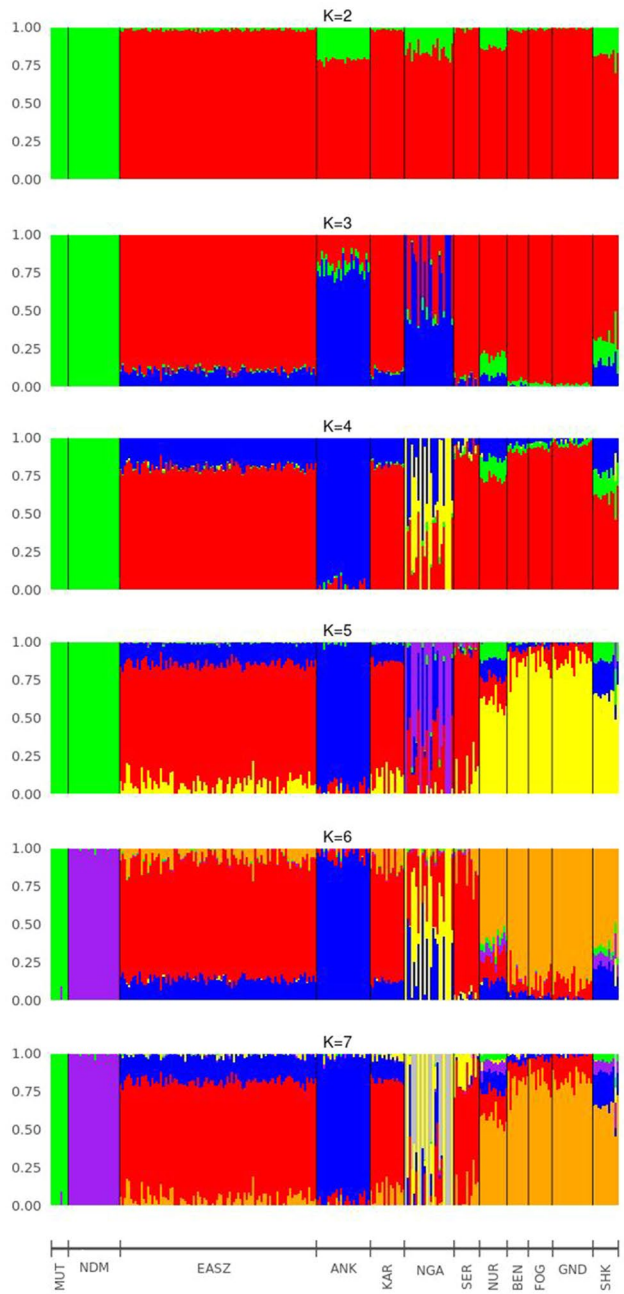


**FIGURE 3 |** Admixture bar plots of each individual assuming different numbers of ancestral breeds (K = 2 to K = 7). ANK, Ankole; BEN, Benshangul; FOG, Fogera; GND, Gindeberet; KAR, Karamojong; MUT, Muturu; NDM, N'Dama; NGA, Nganda; NUR, Nuer; SER, Serere; SHK, Sheko.

the indigenous African breeds with respect to their natural adaptation and tolerance attributes.

It is believed that the taurine background of the Sheko is linked to its trypanotolerance characteristics (Lemecha et al., 2006; Gibbs et al., 2009). This taurine admixture is likely a legacy of the first taurine occurrence on the African continent (Hanotte et al., 2000; Salim et al., 2014). A study on mtDNA indicates that all African cattle breeds analyzed so far carried taurine mtDNA haplotypes which suggests that these waves of indicine arrival into Africa were male-mediated (Bradley et al., 1996; Bonfiglio et al., 2012).

## Identification of Candidate Signature Genes Associated With Trypanotolerance

A total of 20, 14, and 65 genomic regions harboring 109, 64, and 202 candidate signature genes were identified by *iHS*, CLR, and *Rsb* analyses in 22, 10, and 27 autosomes in Sheko, respectively (**Figure 4** and **Supplementary Tables 2–4**). Among the 364 unique candidate signature genes identified by *iHS*, CLR, and *Rsb* analyses, 260 disposed of enriched GO terms ($\alpha = 0.05$) (**Supplementary Tables 5–7**). Moreover, 96, 323, and 463 intergenic variants were identified in gene desert regions by *iHS*, CLR, and *Rsb* analyses in all candidate regions, respectively (**Supplementary Tables 8–10**).

Mainly focusing on the top 10 candidate signature genes of each of the three methods, we performed a literature survey and identified 15 (4 genes identified by *iHS*, 3 genes identified by CLR, 7 genes identified by *Rsb*, and 1 gene identified by both *iHS* and CLR) candidate signature genes that are associated with trypanotolerant attributes which have been reported in previous studies (**Table 1**). Notably, polymorphisms in or nearby the MIGA1, CDAN1, HSPA9, and PCSK6 genes in the genome of Sheko might be associated with the evolutionary response against anemia. The MIGA1 gene is associated with iron deficiency anemia and immunity (Moura et al., 2001; Rouault, 2006). This gene also plays a major role for the development and proliferation of lymphocyte since defective T- and B-cell activation is caused by inadequate iron uptake (Rouault, 2006; Jabara et al., 2016). Another interesting candidate signature gene related with anemia is CDAN1. Polymorphisms in this gene are associated with congenital dyserythropoietic anemia type 1 (Dgany et al., 2002; Renella et al., 2011). Moreover, the hsp70 protein family and the heat shock 70kDa protein 9 (HSPA9) gene play a role as a downstream mediator of erythropoietin signaling and contribute to normal erythropoiesis (Singh et al., 1997; Ran et al., 2000; Ohtsuka et al., 2007; Chen et al., 2011). The mutation in this gene is associated with sideroblastic anemia (Schmitz-Abe et al., 2015), while the PCSK6 gene is involved in iron homeostasis and hence related with iron deficiency anemia (Guillemot and Seidah, 2015). In agreement with our findings, it has been reported by several studies that trypanotolerant N'Dama do better control anemia, a process mediated by hematopoietic cells differentiation, than trypanosusceptible breeds (Berthier et al., 2016; Naessens, 2006).

In previous studies, trypanotolerant animals were reported to switch from innate immune response to adaptive immune response with the induction of active macrophages (M2) following trypanosome infection (Stijlemans et al., 2010; Bosschaerts et al., 2011). For instance, humoral response differences between trypanosusceptible (Boran) and trypanotolerant (N'Dama) cattle corresponding to the amount of antibody (Ab) titers have been observed. There is a difference in trypanosome-specific antiparasite Ab secreting cells in spleen and B cell activation between trypanotolerant and trypanosusceptible cattle (La Greca et al., 2014; Mamoudou et al., 2016; Morrison et al., 2016). In agreement with this, we identified the SPAG11B, RAET1G, PPP1R14C, and

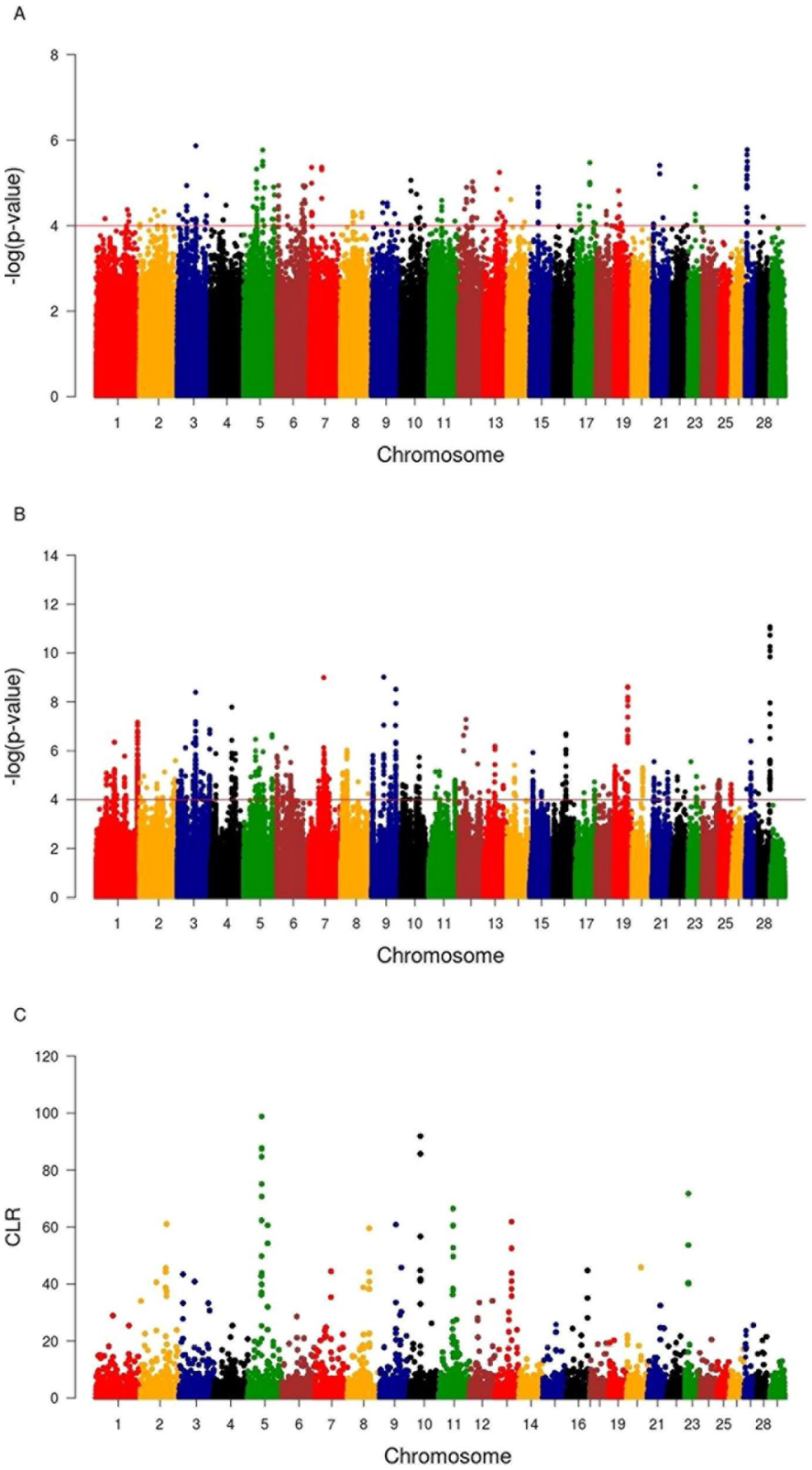


**FIGURE 4 |** Manhattan plots of genome-wide *iHS* **(A)**, *Rsb* **(B)**, and CLR **(C)** analyses. The x-axis shows the autosomal chromosomes and the y-axis shows –log transformed *P*-values **(A** and **B)** and CLR values **(C)**.

**TABLE 1 |** Summary of major candidate signature regions identified by CLR, *iHS*, and *Rsb* analyses.

| Genes | Method | CHR | Association | Position (UMD3.1) Start-End (bp) |
|---|---|---|---|---|
| MIGA1 | Rsb | 3 | Anemia, immune tolerance and neurological dysfunction (Moura et al., 2001; Rouault, 2006; Jabara et al., 2016) | 6706504–67137909 |
| CDAN1 | CLR | 10 | Anemia (Dgany et al., 2002; Renella et al., 2011) | 38138863–38151656 |
| HSPA9 | Rsb | 7 | Anemia (Singh et al., 1997; Ran et al., 2000; Ohtsuka et al., 2007; Chen et al., 2011; Schmitz-Abe et al., 2015) | 51506219–51521515 |
| PCSK6 | iHS | 21 | Anemia (Guillemot and Seidah, 2015) | 29553201–29673109 |
| SPAG11B | iHS | 27 | Immune tolerance (Yang et al., 1999; Ganz, 2003) | 4920083–4942958 |
| RAETIG | Rsb | 9 | Immune tolerance (Eagle and Trowsdale, 2007; Tomasec et al., 2007; Lanier, 2015) | 88232044–88408862 |
| PPP1R14C | Rsb | 9 | Immune tolerance, anemia and neurological dusfunction (Hanke et al., 1996; Linnekin et al., 1997; Liu et al., 2002; Haynes et al., 2003; Vignali et al., 2008) | 88384683–88500749 |
| TTC3 | Rsb | 1 | Immune tolerance and neurological dysfunction (Chen et al., 2003; Liu et al., 2009; Pulst, 2016) | 151034217–151141015 |
| ERN1 | Rsb | 19 | Immune tolerance and neurological dysfunction (Leach and Treacher, 1998; Oosthuyse et al., 2001; Rius et al., 2008; Liao et al., 2009; Minchenko et al., 2015; Singh et al., 2016) | 48924511–48971838 |
| CAPG | CLR | 11 | Immune tolerance and neurological dysfunction (Leach and Treacher, 1998; Oosthuyse et al., 2001; Rius et al., 2008; Liao et al., 2009; Zhang et al., 2009; Minchenko et al., 2015; Singh et al., 2016) | 49423731–49438680 |
| TTBK2 | CRL | 10 | Neurological dysfunction (Jackson, 2012; Matilla-Duenas, 2012) | 38159317–38248606 |
| POLR3B | iHS | 5 | Neurological dysfunction (Schiffmann and van der Knaap, 2009; Daoud et al., 2013) | 70062608–70178439 |
| GNAS | iHS and CLR | 13 | Neurological dysfunction (Tuntasuvan et al., 1997; Bastepe, 2008; Giordani et al., 2016) | 58010287–58049012 |
| CHAT | Rsb | 28 | Listlessness (Johnson et al., 2016) | 44143245–44187239 |
| AP1M1 | iHS | 7 | Listlessness (Molenaar et al., 1982) | 7820650–7850254 |

TTC3 genes which are involved in immune tolerance in Sheko. Interestingly, the PPP1R14C gene could play an important role in the tolerance mechanisms of Sheko with PP1, a competitive inhibitor of ATP binding of Src tyrosine kinase family members (Hanke et al., 1996; Liu et al., 2002). The inhibition of Src kinase is associated with the termination of stem cell factor induced proliferation of hemopoietic cells (Linnekin et al., 1997). It was also reported that Src kinases are involved as a primary activator of AKT (serine/threonine kinase family). AKT plays a critical role in adaptive immunity through the inhibition of regulatory T-cells ($T_{reg}$ cells), which could play a key role in maintaining the immune tolerance (Liu et al., 2002; Haynes et al., 2003; Vignali et al., 2008). In addition, activated AKT is a mediator of neuronal cell survival (Liu et al., 2002; Chen et al., 2003; Pulst, 2016).

Furthermore, the TTC3 gene is also involved in the regulation of AKT signaling and is related with immune tolerance and neuronal cell survival (Chen et al., 2003; Liu et al., 2009; Pulst, 2016). Therefore, the mutation in the PPP1R14C gene is associated with three tolerance attributes (immune tolerance, neurological dysfunction, and anemia). Remarkably, the candidate signature gene RAET1G is one of the few genes that could encode a ligand recognized by NKG2D proteins in response to stress and infections (Eagle and Trowsdale, 2007; Tomasec et al., 2007; Lanier, 2015). Furthermore, the isoforms of the SPAG11B gene encode defensine-like peptides which are expressed by phagocytic cells (Yang et al., 1999). These structurally diverse peptides make multimeric forms during infection and disrupt the membrane of the pathogen (Ganz, 2003). They are also involved in the recruitment of T- and dendritic cells to facilitate the adaptive immunity (Yang et al., 1999). Therefore, the mutations or the differential expression of these genes are critical for the immune tolerance of Sheko to combat anemia and neurological dysfunction caused by trypanosome infection.

Trypanosomiasis is also reported to affect the nervous system of the animal. Fatihu et al. (2009) and Allam et al. (2011) reported causes of thyroid and parathyroid gland dysfunction following trypanosome infection in cattle. The dysfunctioning of thyroid and parathyroid glands often result in neurological complications or cerebral pathology (Jaggy et al., 1994; Wu and Hersh, 1994). Therefore, mutations in the POLR3B, MIGA1, TTC3, ERN1, CAPG, GNAS, and TTBK2 genes might be associated with the response to the presence of the parasite in the brain white matter, cerebral fluid, thyroid, and parathyroid glands. The endoplasmic reticulum to nucleus signaling 1 (ERN1) and capping protein gelsolin-like (CAPG) genes are involved in the regulation of hypoxia (a state of a cell with inadequate or reduced oxygen availability) (Leach and Treacher, 1998). The reduction of the hypoxic response element in the spinal cord results in the progressive degradation of the motor neuron (Oosthuyse et al., 2001; Minchenko et al., 2015). Therefore, mutations in the ERN1 and CAPG genes are associated with neurological dysfunction (Liao et al., 2009; Minchenko et al., 2015). The ERN1 and CAPG genes might also be involved in the innate immune response since hypoxia triggers innate immunity responses through the activation of the hypoxia induced factor α1 (HIF-1α) (Oosthuyse et al., 2001; Rius et al., 2008; Singh et al., 2016).

Trypanosome parasites are also known for their ability to manipulate the host immune responses. One of the mechanisms of innate immune evasion by these parasites is the reduction of HIF-1α by indolepyruvate. Therefore, the reduction of hypoxic response elements in the spinal cord results in the progressive degradation of the motor neuron (Oosthuyse et al., 2001). Therefore, the mutation in the ERN1 and CAPG genes in particular would be related to the host innate immune evasion of the parasite. Another reported candidate signature gene related with neurological dysfunction is the TTBK2 gene. A mutation

in the TTBK2 gene is associated with spinocerebellar ataxia which is a genetic syndrome causing progressive degeneration of the cerebellum and the spinal cord (Jackson, 2012; Matilla-Duenas, 2012). Moreover, a mutation in the POLR3B gene is associated with hypomyelinating leukodystrophy which is characterized by a deficiency in myelin deposition of the white matter of the brain (Schiffmann and van der Knaap, 2009; Daoud et al., 2013). In addition, the POLR3B gene is also involved in positive regulation of the interferon-beta production and the innate immune response (GO:0032728, GO:0045089).

Strikingly, a mutation in the GNAS gene is associated with pseudohypoparathyroidism which is characterized by a low level of calcium and a high phosphate level in the blood (Bastepe, 2008). Allam et al. (2011) reported a similar profile during trypanosome infection in cattle that could be associated with neurological dysfunction such as muscle spasm (Tuntasuvan et al., 1997; Bastepe, 2008; Giordani et al., 2016). Furthermore, during trypanosome infection, listlessness and emaciation are some of the clinical signs of the infection (Nantulya, 1986; Steverding, 2008; Noyes et al., 2011). These clinical signs might be associated with the destruction of the thyroid gland by trypanosome parasites in cattle (Fatihu et al., 2009). The candidate signature genes AP1M1 and CHAT are related with these clinical signs. Most importantly, the AP1M1 gene is a member of the adapter protein complex which is involved in thyroid abnormalities (Johnson et al., 2016). Due to the thyroid gland dysfunction (hypothyroidism), the nerves are unable to conduct electrical impulses properly. This leads to general weakness, lethargy, and listlessness (Jaggy et al., 1994). The mutation in the CHAT gene is associated with myasthenia gravis which is an autoimmune disease characterized by load dependent muscle weakness (Molenaar et al., 1982).

Our findings show strong selective sweeps (**Figures 4A–C**) in the genomic regions around the selected signature genes of **Table 1** (**Supplementary Table 11**). This might indicate that the mutations in these genes have reached fixation or are near fixation. Therefore, the identified candidate signature genes in **Table 1** might play a major role in the natural tolerance attributes of Sheko against trypanosomiasis. Moreover, the comparison of candidate signature genes identified by the *iHS*, CLR, and *Rsb* methods show more overlaps between *iHS* and CLR than between *Rsb* and *iHS* or CLR analyses (**Figures 5A, B**), in agreement with *Rsb* being a powerful method to detect selection signature when the selected allele has reached fixation (Tang et al., 2007; Oleksyk et al., 2010; Bahbahani et al., 2018).

Among the 15 identified candidate signature genes (**Table 1**), the MIGA1, RAETG, and PPP1R1AC genes are not significantly functionally enriched ($\alpha = 0.05$). This might indicate that these candidate signature genes in Sheko could be specific to the environmental pressure in the region such as trypanosomiasis. Moreover, the identified signature regions of the three methods were compared with trypanotolerant QTL regions which were reported by Hanotte et al. (2003). Among the 55 trypanotolerant QTL, which were identified by crossing trypanotolerant N'Dama and susceptible Boran, 6 regions were overlapping with trypanotolerant QTL in N'Dama (**Supplementary Table 12**). Interestingly, among the identified candidate signature genes in **Table 1**, the AP1M1 and GNAS genes are found in these overlapping regions. The overlapping regions and genes of Sheko and N'Dama might indicate occurrence of selection at the same genes in these two breeds against the same environmental pressures.

## Functional Annotation of Candidate Signature Genes

In order to characterize the biological functions of functionally enriched candidate genes, a treemap was produced using the geneXplain platform (Krull et al., 2006). The treemap shows the clusters of 30 functional terms. Most of these terms are associated with cellular transport, metabolic processes and biological regulation (**Figure 6**). Interestingly, among the 30

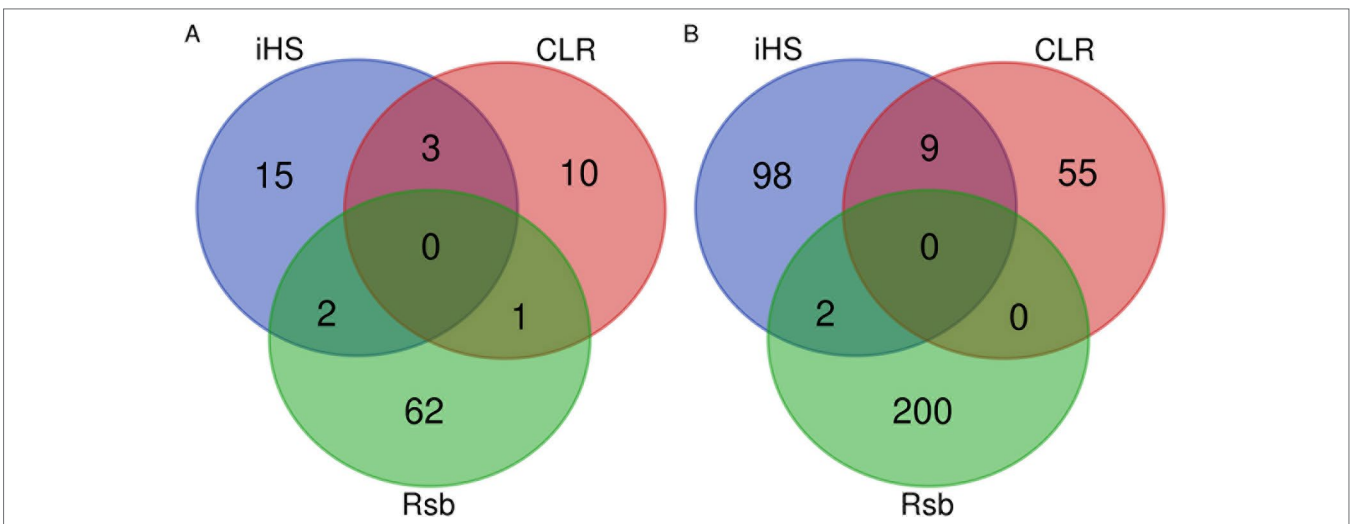


**FIGURE 5 |** Venn diagrams of the overlapping **(A)** genomic regions and **(B)** candidate genes identified by *iHS*, CLR, and *Rsb*.

enriched terms, two GO-terms are T-cell chemotaxis and cell–cell adhesion which play a critical role in the immune system (Springer, 1990; Gerard and Rollins, 2001; Bach et al., 2007). T-cell chemotaxis (chemoattractant cytokines) is a process that requires the movement of T-cells in response to a certain signal or external stimulus. The movement or circulation of immune cells in the blood and lymph as non-adherent cells and in tissues as adherent cells is critical for patrolling the body against infectious organisms effectively (Springer, 1990). For instance, β defensin is chemotactic for chemokine receptors of macrophages, natural killer cells, immature dendritic cells, and memory T-cells. Therefore, the recruitment of these cells to the site of a microbial invasion provides a link between innate and adaptive immunity (Yang et al., 1999). Likewise, T-cell mediated migration of thymocyte toward chemokines was observed following trypanosome infection in human (Mendes-da Cruz et al., 2006). In the presence of infectious organisms (foreign antigens), the immune cells aggregate at the site of the infection and through their adhesion receptors they adhere to cells bearing a foreign antigen (Springer, 1990).

## Identification of Overrepresented Pathways in the Candidate Signature Gene Sets

Pathway analysis has become a powerful tool in order to refine the molecular mechanisms of disease tolerance. The rationale of pathway analysis lies in the detection of overrepresentation of biologically defined pathways based upon the functionally enriched candidate selected genes. We performed pathway analysis using the TRANSPATH database on the geneXplain platform. The TRANSPATH pathway analysis identified 15 genes out of 260 functionally enriched genes that are involved in 13 overrepresented TRANSPATH pathways (**Table 2**). Among these genes, the immunoproteasome PSMD7 gene is involved in most of the overrepresented pathways. This gene is involved in the processes of presenting antigens by the major histocompatibility complex (MHC) class I proteins to CD8+ T-lymphocytes (Morrot and Zavala, 2004; Goldszmid and Sher, 2010; Jordan and Hunter, 2010). Sufficient induction of CD8+ during infection leads to pathogen elimination. It has been reported that immunoproteasome subunits are key determinants of the CD8+ T-cell level and quality involved in host resistance to trypanosomes infection (Ersching et al., 2016). This gene plays a critical role in the development of adaptive immunity or tolerance (Doolan and Hoffman, 1999).

However, adaptive immunity also plays a key role for the emergence of auto-immunity. Previous studies indicate that trypanosome infection could deplete thymocytes. As a result, immature T-lymphocytes are released from the thymic central tolerance and differentiate into mature T-helper cells in the lymph nodes (Flávia Nardy et al., 2015). This process would induce auto-immunity against self-antigens. Moreover, during trypanosome infection, the red blood cell membrane might be damaged by

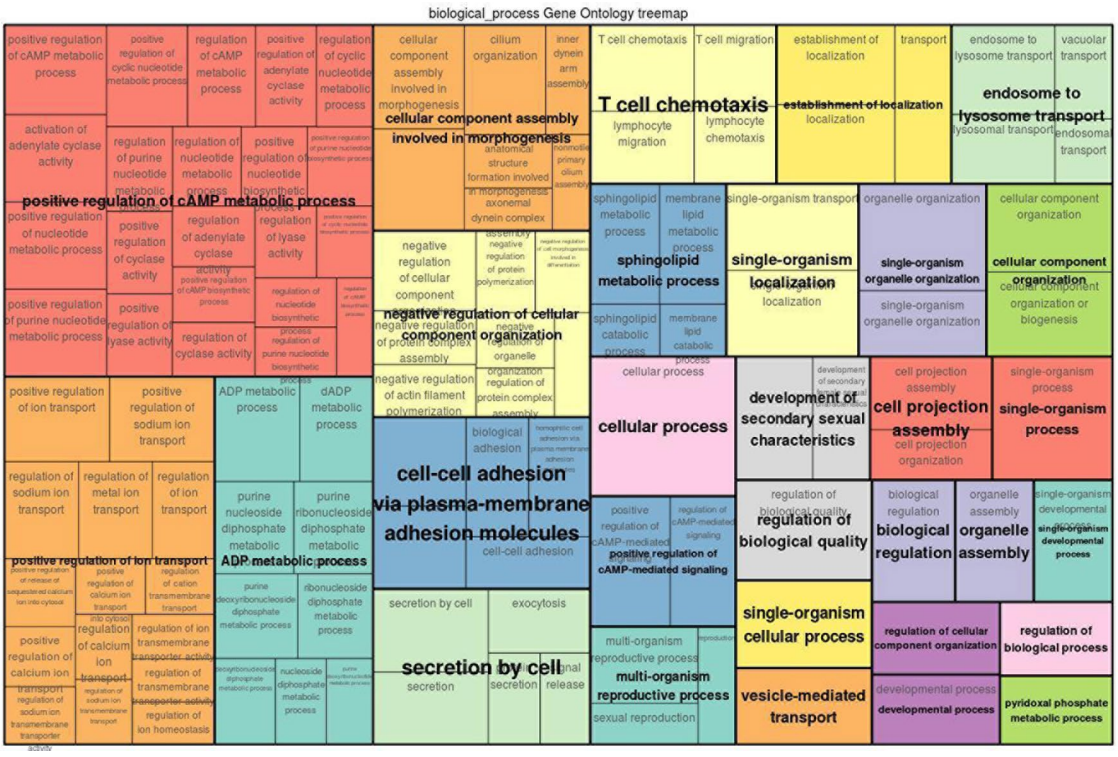


**FIGURE 6 |** GO treemap for the 260 functionally enriched ($P < 0.05$) genes. The size of the boxes corresponds to the –log10 $P$-value of the GO-term. The boxes are grouped together based on the upper-hierarchy GO-term which is written in bold letters.

**TABLE 2 |** Overrepresented pathways for the identified candidate signature genes.

| Pathway | Raw *P*-value | Genes |
|---|---|---|
| PDGF B —> STATs | 0.003 | STAT3, STAT5A |
| Stress-associated pathways | 0.007 | MBP, MEF2A, PSMD7, RAF1, RBX1, STAT3 |
| E2F network | 0.008 | AKT3, CDC25C, PPP2R5A, PSMD7, RAF1, RBX1 |
| G2/M phase (cyclin B:Cdk1) | 0.015 | AKT3, CDC25C, PSMD7, RBX1 |
| IMP —> ADP | 0.025 | AK5, AMPD3 |
| ARIP1 —> atrophin1 | 0.034 | AKT3, APBA1 |
| p38 pathway | 0.039 | MBP, MEF2A, STAT3 |
| Plk1 cell cycle regulation | 0.039 | CDC25C, PSMD7, RBX1 |
| IL-3 signaling | 0.043 | MBP, RAF1, STAT5A |
| Aurora-B cell cycle regulation | 0.045 | CENPE, PSMD7, RBX1 |
| Oxygen independent HIF-1alpha degradation | 0.045 | PSMD7, RBX1, UBE2R2 |
| Cul3 —/Nrf2 | 0.047 | PSMD7, RBX1 |
| S phase (Cdk2) | 0.048 | CDC25C, RAF1, RBX1 |

*The names of the pathways are provided by the TRANSPATH database on the geneXplain platform.*

parasite enzymes such as proteases or phospholipases. This could expose epitopes which are not recognized as self-antigens and would trigger immune-mediated hemolysis due to antibody response against these self-antigens (Taylor, 1998). This could be controlled by suppressing the development of auto-reactive immune cells through ubiquitination which is a degradative tag to be recognized by a proteasome complex such as PSMD7 (Lodish et al., 2004; Zinngrebe et al., 2014). Furthermore, some of the identified candidate signature genes are also associated with protein ubiquitination processes which might indicate that these genes are also involved in the functions described above (**Supplementary Tables 2–4**). To the best of our knowledge, our study is the first to show the potential of a molecular mechanism for controlling auto-reactive immune cells caused by trypanosomiasis in cattle. In agreement with our finding, Kierstein et al. (2006) reported that a trypanotolerant mouse strain showed overexpression of several genes encoding proteases.

In general, most of the overrepresented pathways (PDGFB –> STATs, stress associated pathways, IMP –> ADP, ARIP1 –> atrophin 1, p38 pathway, IL-3 signaling, oxygen independent HIF-1alpha degradiation and Cul3 –/Nrf2) pathways are activated by cellular stresses and antigens while others [E2F network, G2/M phase (cyclin B:Cdk1), S phase (Cdk2), Plk1 cell cycle regulation and Aurora-B cell cycle regulation] pathways are involved in cell cycle processes.

The first two pathways in **Table 2** (PDGFB –> STATs and stress associated pathways) are related to the immune system and anemia. Especially, in stress associated pathways we find MBP, RAF1, MEF2A, and STAT3 genes that are involved in the immune and nervous systems. In the MBP gene, there are eight different mRNAs due to alternative splicing of exons (Zelenika et al., 1993). Three of the eight splice variants are expressed in the brain, macrophages and hemolymphopoietic tissues such as spleen, bone marrow, and thymus (Zelenika et al., 1993). This gene is also involved in the interleukin (IL)-3 signaling pathway. IL-3 is a T-cell-derived hematopoiesis stimulating cytokine involved in the production, differentiation and function of granulocytes and macrophages (Ymer et al., 1985; Dorssers et al., 1987). This suggested that the expression of alternatively spliced MBP mRNAs is related with the immune system in response to trypansome infection or the presence of a pathogen in the central nervous system. The serine/threonine kinase proto-oncogene RAF1 is also related with the stress associated pathway and is involved in inducing adaptive immunity by regulating the expression of cytokines that are important for the differentiation of T-helper cells (Gringhuis et al., 2009).

Moreover, STATs family members are also involved in the activation of various cytokines and in the promotion of cell survival by inducing the expression of antiapoptotic BCL2L1/BCL-X(L) genes (Benito et al., 1996; Packham et al., 1998; Yuan et al., 2004). For instance, STAT3 activation by trypomastigotes was associated with the survival of cardiomyocytes during infection (Ponce et al., 2012; Stahl et al., 2013). The other gene involved in defense response is MEF2A which is associated with promoting antimicrobial peptide expression during infection (Clark et al., 2013). This gene is also involved in neuronal cell survival and loss of function (Gong et al., 2003; She et al., 2011). As reported by She et al. (2012), neurotoxins induce ubiquitination of MEF2A in response to toxic stress which leads to the loss of neuronal viability. Furthermore, He et al. (2015) reported that increased platelet-derived growth factor (PDGF)-B related signaling is associated with induced chemokine secretion which is a mediator of innate and adaptive immune responses (Kim and Broxmeyer, 1999). In addition, knock-out mice for PDGF-B develop anemia (Kaminski et al., 2001) which indicates that the PDGFB –> STATs pathway is also involved in this disease.

The E2F network as well as the Cdk1 and Cdk2 related pathways are also associated with anemia which is the most prominent and consistent clinical sign of trypanosome infection (Kaminski et al., 2001; Dimova and Dyson, 2005; Noyes et al., 2011, Hu and Sun, 2016). The tumor suppressor retinoblastoma (Rb) is the inhibitor of E2Fs. When Rb binds to E2Fs, it prevents E2F mediated activation of transcriptional genes. In quiescent cells, E2F is required for the cell differentiation through a series of signal transduction cascades, including Cdks activation and phosphorylation. The Aurora-B and Plk1 pathways are involved in the activation and phosphorylation of Cdks, respectively. As a result of these and several other signaling cascades, E2Fs is activated while inactivating Rb. The activated E2F mediates quiescent cells for S phase entry and cell cycle progression (Dyson, 1998; Nevins, 1998; Trimarchi and Lees, 2002; Dimova and Dyson, 2005; Song et al., 2007). Hu et al. (2012) reported that mice deficient for both E2F8 (i.e., E2F gene family) and Rb show severe anemia.

Furthermore, the hypoxia inducible factor (HIF) and the nuclear factor-erythroid 2-related factor 2 (NRF2) pathways are related with anemia (Lee et al., 2004; Silva and Faustino, 2015). During hypoxia, HIF facilitates a high production of red blood cell (erythropoiesis) in order to overcome shortage of oxygen (Silva and Faustino, 2015). The other pathway, NRF2, regulates the expression of antioxidant responsive element-driven genes and plays a critical role in the antioxidant

responsive element-driven cellular protection (Cho et al., 2002). In addition, knockout mice for NRF2 show regenerative immune-mediated hemolytic anemia which indicates that this pathway is involved in erythrocyte maintenance during oxidative stress (Lee et al., 2004).

Intriguingly, serine/threonine kinase family isoforms of the AKT gene are involved in the E2F, Cdk1, IMP-ADP, and ARIP1-atrophin1 pathways. This gene is activated in the host cells during trypanosome infection (Woolsey et al., 2003; Chuenkova and PereiraPerrin, 2009). The host kinase AKT promotes infected host cell survival and restricts the growth of intracellular parasites (Caradonna et al., 2013). AKT3 is also a key mediator of down stream signaling pathways of activated receptor tyrosine kinases which play a role in STAT3 activation (Yuan et al., 2004; Chuenkova and PereiraPerrin, 2009. The different isoforms of the kinase AKT regulate the development of immunity and autoimmuniy. Zhang et al. (2013) reported that AKT is predominantly expressed in the innate immune cells. The isoforms of AKT are primarily involved in regulating inflammatory responses although it has been reported that AKT also modulates adaptive immune responses (Liu et al., 2002).

Moreover, the AKT related pathway Atrophin-1 plays a role in erythroid and lymphoid cell differentiation and in E3 ubiquitin ligase atrophin-1 interacting protein 4 (ITCH) signaling cascades. Atrophin-1 is involved in the regulation of immune responses through Notch-mediated signaling pathways (Qiu et al., 2000; You et al., 2009; Aki et al., 2015). It is also associated with spinocerebellar degeneration caused by extended CAG repeats encoding several glutamine units (polyglutamine tract) in the atrophin-1 protein (Kanazawa, 1999). The disease is characterized by neurological symptoms such as ataxia which is one of the clinical signs of trypanosome infection (Tuntasuvan et al., 1997; Suzuki and Yazawa, 2011; Giordani et al., 2016).

Further important pathways are p38, IMP –> ADP, and the aurora B-cell cycle regulation pathways that are involved in the host defense mechanism. The p38 pathway is a MAPK-related pathway which is activated by various physical and chemical stresses, such as hypoxia and various cytokines. The activation of the p38 pathway is critical for normal immunity and inflammatory responses (Roux and Blenis, 2004). Moreover, the AK5 and AMPD3 genes are involved in the IMP –> ADP pathway and play a central role in the regulation of inflammation and red blood cell homeostasis (Tavazzi et al., 2000; Mabley and Szabo, 2008). AK5 is associated with double positive thymocyte and auto-immunity regulation in the brain and pancreatic tissues (Stanojevic et al., 2008) while the AMPD3 gene is involved in the regulation of the energy state of red blood cells during oxidative stress (hypoxia) (Tavazzi et al., 2000). In addition to that, the aurora B-cell cycle regulation pathway is involved in the progression of T-lymphocytes which play a critical role for the development of innate and adaptive immunity (Song et al., 2007; Paul et al., 2011). To this end, the HIF and NRF2 related pathways are directly associated with the induction of host innate and adaptive immunity under oxidative stress (Singh et al., 1997; Cramer et al., 2003; Jantsch et al., 2011; McNamee et al., 2013; Battino et al., 2018).

In summary, our findings of the search for signature genes appear to be well substantiated by the results of the overrepresented pathways analysis. This implies that most of the overrepresented pathways are mainly associated with host defense mechanisms against pathogens and anemia. Particularly, stress-associated, HIF and NRF2 related pathways are involved in oxidative stress responses. Interestingly, trypanosome infection induces the production of superoxide, hydrogen peroxide, peroxyl radicals, and hydroxyl radicals which are known to cause oxidative stress followed by tissue damage and hemolysis (Saleh et al., 2009). Under oxidative stress (hypoxia), erythrocytes are important mobile oxidative sinks (antioxidant) for themselves, other cells, and tissues. However, these properties of the red blood cells during oxidative stress contribute to its susceptibility toward hemolysis which leads to anemia (Chan et al., 2001; Sangokoya et al., 2010). In order to overcome the shortage of oxygen, stress-associated, HIF, and NRF2 related pathways play a critical role in the production of red blood cells in which hemoglobin acts as oxygen repository for red blood cells and other cells (Chan et al., 2001; Sangokoya et al., 2010; Silva and Faustino, 2015).

None of the most significant candidate signature genes (**Table 1**) was contained in the overrepresented pathway gene list (**Table 2**). This indicates that the candidate signature genes might be involved in the evolutionary gear particularly toward trypanotolerance in Sheko. For instance, candidate signature genes involved in the regulation of hypoxia (ERN1 and CAPG) are not identified in the overrepresented hypoxia related pathways. This might indicate that these candidate signature genes might be specific to oxidative stress tolerance attributes in Sheko. Hence, trypanotolerance of Sheko could be controlled by some major selected genes whose major effect close to fixation in the breed (become breed characteristic) and cohorts of genes with minor effects.

## Identification of Master Regulators Based on Candidate Signature Genes

To gain more insight into the regulatory mechanisms of the identified candidate signature genes, we performed a master regulatory network analysis using the TRANSPATH database in the geneXplain platform. Applying the maximum radius of 10 steps upstream in the regulatory hierarchy, we identified ten master regulators (**Figure 7**). Remarkably, the master regulator Caspase, which is a family of protease enzymes, is associated mainly with regulating the reduction of the load of intracellular parasites, induction of nitric oxide production, increasing the level of CD4 and CD8+ T-cells, secretion of IFNγ, and control of trypanosome infection by macrophages (Gonçalves et al., 2013). This master regulator is involved in programmed cell death such as pyroptosis and necroptosis. These types of programmed cell deaths play a role for protecting an organism against oxidative stress (stress signals) and pathogenic attack (Shalini et al., 2015). In addition, Caspase also plays a role in the normal erythroid differentiation in the terminal stages (Zermati et al., 2001).

Most of the regulatory molecules (Syk, Lck, Lyn, Jak1, Jak2, and Jak3) are protein tyrosine kinases while others (VHR and PTP1B) are protein tyrosine phosphatases and activated kinase (PAK1). These master regulators are mainly associated with innate and adaptive immune responses and are critical for the

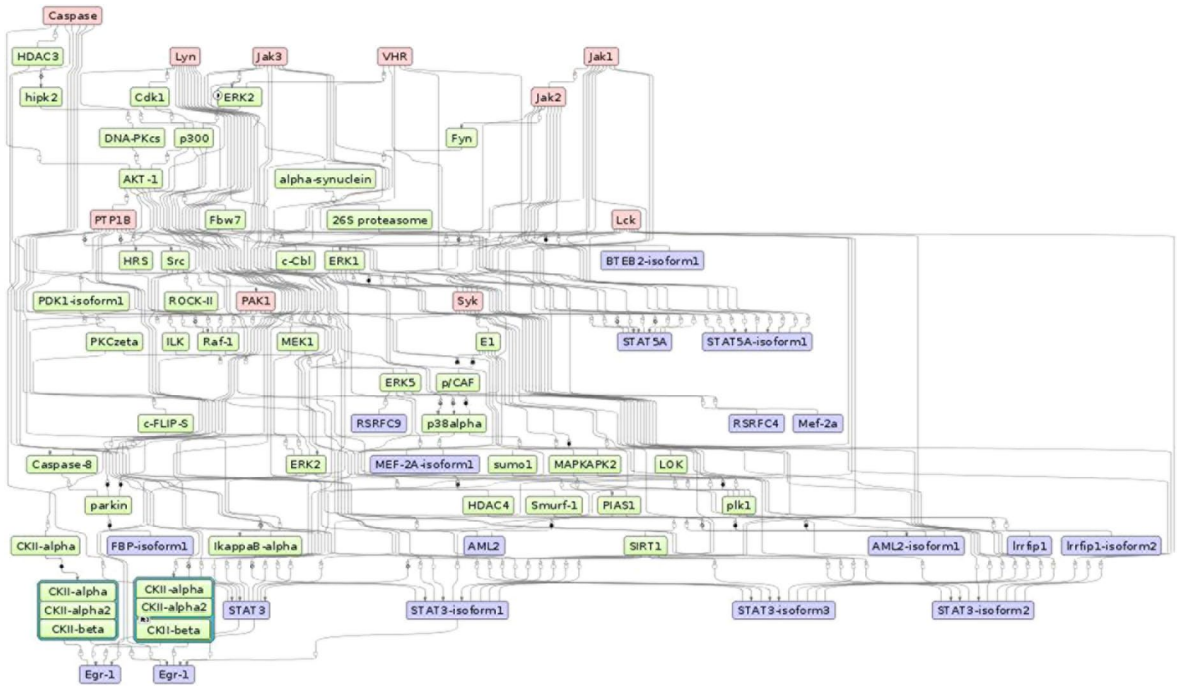


**FIGURE 7 |** The master regulatory networks for Sheko (Caspase, Lyn, Jak1, Jak2, Jak3, VHR, PTP1B, PAK1, Lck, and Syk). Red, blue, and green indicate master regulators, regulated proteins, and connecting molecules, respectively.

functioning of the nervous and immune systems. For instance, the activation of the regulatory molecule Syk requires the regulatory molecule Lck to phosphorylate immunoreceptor tyrosine-based activation motifs. Then, the phosphorylated immunoreceptor tyrosine-based activation motif modulates T-cell proliferation and differentiation by recruiting Syk protein tyrosine kinases (Acuto et al., 2008; Au-Yeung et al., 2009). In addition, coupling of the other master molecules JAK1 and JAK3 occurs on the cell surface receptor of IFNγ, followed by phosphorylation of the IFNγ receptor 1. This process leads to the activation of the STAT1 protein. The STAT1 protein binds to the target element of the IFNγ inducible gene in the nucleus and facilitates the transcription of the target regions during immunity responses (Rosenzweig and Holland, 2005; Casanova and Abel, 2007). Another reported regulator molecule VHR is also involved in the phosphorylation of STAT proteins and in the T-lymphocyte physiology (Alonso et al., 2001; Hoyt et al., 2007). Moreover, the master molecule JAK2 plays a critical role in the maintenance of hematopoiesis. It has been shown that selective deletion of JAK2 results in lethal anemia in adult mice (Grisouard et al., 2014).

Furthermore, a related master molecule, the protein tyrosine phosphatase 1B (PTP1B), is reported to modulate the activation of macrophages and plays a key role in mediating the central dendritic cell function of bridging innate and adaptive immunity (Heinonen et al., 2006; Martin-Granados et al., 2015). The kinase family regulator molecule Lyn is also involved in the regulation of innate and adaptive immune responses (Ingley, 2012). Lyn is also known for mediating the production of type I interferone (IFN-I) which is involved in host defense mechanisms against invading pathogens (Kawai and Akira, 2007; Blasius and Beutler, 2010; McNab et al., 2015). The related kinase regulatory molecule PAK1 is highly expressed in most leukocytes that are involved in immune responses. PAK1 also plays an important role in the activation of MAP-kinase pathways which are involved in all aspects of immune responses, from innate immunity to the activation of adaptive immune responses (Yi et al., 1991; Adachi et al., 1992; Zhang et al., 1995; Dong et al., 2002; Wang et al., 2002; Traves et al., 2014). In general, these proteins and master regulatory molecules are a large family of signaling enzymes that are expressed in various immune cells and regulate immune cell differentiation, cytokine production, and immune responses. Therefore, to maintain the tolerance against a pathogen, the regulation of these signaling pathways is critical (Manning et al., 2002; Salmond et al., 2009).

Strikingly, stress-induced protein kinases could also induce or aggravate auto-immunity by phosphorylating self-antigens to be recognized by auto-antibodies (Utz et al., 1997; Patterson et al., 2014). However, Caspase-mediated apoptosis plays an important role in arresting the development of auto-immunity by eliminating auto-reactive and pro-inflammatory cells (Eguchi, 2001). Moreover, the activation of Caspase and JAK2 is essential for the processes of erythroid differentiation and for the maintenance of hematopoiesis (Zermati et al., 2001). On the other hand, the inhibition of Caspase dependent mechanisms contributes to cell survival (Lamkanfi et al., 2007). We believe that the candidate signature genes involved in anemia, neurological dysfunction, listlessness, and immune tolerance might be governed by the top master regulator Caspase in harmony with other regulatory molecules. In general, our study provides a first report on the top

master regulators for trypanotolerance of Sheko and the overall analysis framework might be helpful to understand the underlying mechanisms of different cattle diseases in future works.

# MATERIALS AND METHODS

## SNP Genotyping and Quality Control

sDNA was extracted from 67 blood and tissue samples according to the QIAGEN DNA extraction protocol (**Supplementary Table 13**). 19 samples from Gindeberet, 12 from Sheko, 13 from Nuer, 12 from Benshangul and 11 from Fogera breeds were collected. All samples were taken randomly from unrelated animals based on the information given by livestock keepers at the time of sampling. All samples were genotyped for 777,962 SNPs using the Illumina BovineHD Genotyping Bead chip. In addition, the genotyping data of two west African breeds (24 N'Dama and 8 Muturu), and five east African breeds (92 EASZ, 25 Ankole, 16 Karamojong, 23 Nganda, and 12 Serere) were obtained from the International Livestock Reaserch Institute (ILRI, Addis Ababa, Ethiopia; Bahbahani et al. (2017)). For quality control, Plink1.9 (Purcell et al., 2007) was used on 735,293 autosomal SNPs. SNPs with minor allele frequency of less than 1% were excluded (19,581 SNPs). Minimum genotyping call rate (<95%) and maximum identity-by-state (IBS) (≥95%) were also used as filtering criteria. Two Benshangul samples failed the genotyping call rate criteria and were excluded from the analysis but no pair of samples was excluded due to the IBS filtering criterion. The total sample size for the down stream analysis consisted of 265 samples and 715,712 SNPs. BEAGLE 4 (Browning and Browning, 2007) was used for inferring haplotype phasing and imputing the missing alleles. The imputation was performed by fitting 83 sliding windows across the autosomes in which on average 8600 markers were included. With in each window 12 iterations were executed. Since our samples consist of indigenous African breeds, the total of 264 (n − 1) animals included in this study are used as a background to impute the missing alleles in the context of indigenous African cattle genome (i.e., without using the reference genome).

## Genetic Background of the Cattle Population

In the eastern part of Africa, the mixture of African taurine and indicine cattle populations is common which reflects the wave of these two different ancestral aurochs in the region (Hanotte et al., 2000; Salim et al., 2014; Bahbahani et al., 2017). Regarding these two ancestral populations, the N'Dama and Muturu breeds are considered as African taurine whereas the Fogera, EASZ, Ankole, Karamojong, and Serere breeds are referred to as African zebu (Bahbahani et al., 2017). The Nuer and Ankole breeds are classified as African sanga (DAGRIS, 2007) while the Nganda breed is assigned to African zenga (Bahbahani et al., 2017). The sanga and zenga cattle are crossbreds between the indigenous humpless cattle and zebu. The latter have higher zebu genetic introgression than the former (Rege, 1999). Interestingly, the Sheko breed is considered as the last oddments of the primordial *Bos taurus* cattle in eastern Africa. However, some animals in the present population of Sheko display small humps which indicates the genetic introgression of zebu cattle (DAGRIS, 2007). Yet, there is no research publication or documentation available on the genetic background of the Benshangul and Gindeberet breeds which are included in this study. The breed type and origin of the cattle samples included in this study are presented in **Table 3**.

## Breed Differentiation, Genetic Relationship, and Structure

In order to understand the genomic structure of Sheko, we considered in total 12 indigenous African breeds genotyped with the Illumina BovineHD Genotyping BeadChip. To assess the within and between population genetic structure and admixture, PCA and admixture analyses were conducted. PCA was performed using Plink 1.9 to estimate the eigenvectors of the variance-standardized relationship matrix of all samples. In order to refine the genetic structure of the indigenous Ethiopian cattle breeds, separate PCA calculation were made for samples that were collected in Ethiopia (Sheko, Benshangul, Gindeberet, Fogera, and Nuer). Admixture analysis was performed using the ADMIXTURE 1.3 software with CV and 200 bootstraps for the hypothetical number of ancestries K ($2 \leqslant K \leqslant 7$). Both PCA and admixture analyses were used to determine the level of admixture and genetic differentiation of the populations. Furthermore, admixture analysis was used to determine the level of indicine and taurine ancestries of each breed at the genome-wide level. In particular, PCA and admixture analyses were performed to show the taurine background of Sheko.

## Analysis of Signatures of Positive Selection

In general, methods for the detection of selection signatures are based on the spatial distribution of allele frequencies and the property of segregating haplotypes in the population (Hayes et al., 2010). As suggested by Ma et al. (2015) and

**TABLE 3 |** Cattle breeds included in the study.

| Breed name | Breed category* | Breed origin |
|---|---|---|
| N'Dama | African taurine | Guinea |
| Muturu | African taurine | Nigeria |
| Ankole | Sanga | Uganda |
| Karamojong | African zebu | Uganda |
| Serere | African zebu | Uganda |
| Nganda | Zenga | Uganda |
| EASZ | African zebu | Kenya |
| Sheko | African taurine and zebu | Ethiopia |
| Nuer | Sanga | Ethiopia |
| Gindeberet | Not available | Ethiopia |
| Benshangul | Not available | Ethiopia |
| Fogera | African zebu | Ethiopia |

**Breed category according to DAGRIS (2007).*
*EASZ, East African Shorthorn Zebu.*

Vatsiou et al. (2016), combining these methods would help to reach a higher power than with single analysis. In this paper, we used EHH and spatial distribution of allele frequency-based methods to identify signatures of positive selection in the genome of the Sheko breed. This denotes that integrated haplotype score (*iHS*) and CLR analyses were performed on Sheko (12) while the ratio of site-specific EHH (EHHS) between populations (*Rsb*) analysis were performed between Sheko (12) and combined trypanosusceptible reference cattle populations (179) [EASZ (92) (Muhanguzi et al., 2014; Van Wyk et al., 2014), Ankole (25) (Magona et al., 2004), Karamojong (16) (Muhanguzi et al., 2017), Nganda (23) (FAO, 2004), Serere (12) (Ocaido et al., 2005) and Fogera (11) (Sinshaw et al., 2006)]. The results of these tests were combined into one gene set.

### Extended Haplotype Homozygosity Based Methods

*Rsb* and *iHS* are LD based approaches which are implemented in R package *rehh*. Both *Rsb* and *iHS* are used to identify genome-wide signatures of selection (Gautier and Vitalis, 2012). These tests start with a core haplotype (i.e., a set of closely linked SNPs in which recombination does not take place) identification (Sabeti et al., 2002; Skipper, 2002). Then, the decay of LD as a function of the distance from the core haplotypes is analyzed (Sabeti et al., 2002). The *Rsb* analysis was performed between Sheko and the combined group of trypanosusceptible breeds. For each group, integrated site-specific EHH of each SNP (iES) was calculated. Standardized log-ratio between iES of the two groups was used to calculate *Rsb* values. The *iHS* values were calculated for Sheko as the natural log ratio of integrated EHH (iHH) between reference and alternative alleles for each SNP (Gautier and Vitalis, 2012; Bahbahani et al., 2018). The bovine reference genome (UMD3.1) is used as the reference allele while the study population (Sheko) is considered as the alternative allele. The *iHS* values were standardized based on the calculated mean and standard deviation values. This allows direct comparisons among different SNPs regardless of their allele frequencies (Gautier and Vitalis, 2012). For the standardization of *Rsb* values, median and standard deviation values were used. One-tailed Z-tests for *Rsb* and two-tailed Z-tests for *iHS* were applied on the standardized and normally distributed *Rsb* and *iHS* values (**Supplementary Figures 2A, B**) to identify statistically significant SNPs that are under positive selection. For one-tailed Z-tests, $P = 1 - \Phi(\text{Rsb})$, whereas $P = 1 - 2|\Phi(\text{iHS}) - 0.5|$ was used for the two sided tests with $\Phi$ being the Gaussian cumulative density function. For both *Rsb* and *iHS* *P*-values, the significance threshold of $\alpha = 10^{-4}$ was applied following the study of Bahbahani et al. (2018) to identify candidate regions.

### Spatial Distribution of Allele Frequency Based Method

The CLR test is an LD based selective sweep searching algorithms using the information from the spatial distribution of allele frequencies (Charlesworth, 2012). CLR is used to identify skewed patterns of the allele frequency spectrum toward excess of rare alleles and high frequency alternative alleles due to the hitchhiking effect (Kim and Stephan, 2002; Nielsen et al., 2005; Qanbari et al., 2014). The *P*-values were calculated by the rank of the genome wide scan of CLR values. As suggested by Wilches et al.(2014), the 95$^{th}$) quantile of the distribution of the top CLR *P*-values was used to identify a significance threshold of $\alpha=10^{-5}$ (**Supplementary Figure 3**). For CLR analysis, the Sweepfinder2 (DeGiorgio et al., 2016) software was used for each chromosome with a window size of 50kb including on average 226 SNPs per window. Sweepfinder2 estimates CLRs in the context of background selection to identify sweeps (DeGiorgio et al., 2016; Huber et al., 2016).

## Functional Annotation of Selected Candidate Regions

Genes found within 25 kb around the most significant SNP were considered as candidate genes (Bahbahani et al., 2018). Protein-coding and RNA genes found within the candidate regions were retrieved using the BioMart tool (Kinsella et al., 2011). The R package *Enrichr* (Kuleshov et al., 2016) was used to determine the candidate signature genes that are functionally enriched in GO terms with respect to the whole bovine reference genome background ($\alpha = 0.05$). These functionally enriched candidate signature genes were used to produce a treemap which shows clusters of functional terms based on the biological functions of the candidate signature genes.

To gain more insight into the functional properties and molecular mechanisms involved in trypanotolerance, overrepresented pathways were analyzed using the TRANSPATH database (Krull et al., 2006) of the geneXplain platform (http://genexplain.com/). Furthermore, to understand the regulatory mechanisms of the candidate signature genes and the signaling cascades in the regulatory hierarchy involved in trypanotolerance, the identification of master regulators was conducted using the TRANSPATH database.

# CONCLUSION

For generations, African animal trypansomiasis has been the major selection pressure in the region. We have identified the candidate causative genes, pathways, and master regulators associated with the adaptation of the Sheko breed to its natural environmental pressure. Most of the identified candidate signature genes, overrepresented pathways, and master regulator molecules were involved in immune tolerance, neurological dysfunction, and anemia. This entails that the genome of Sheko was targeted by these environmental pressures which are associated with trypanosomiasis. Therefore, this study helps as an input for designing and implementing genetic intervention strategies to improve the performance of susceptible as well as animals which are relatively tolerant toward higher trypanotolerance.

The improvement of the cattle health contributes to increase the production of milk and meat. The improvement of the cattle health enhances the draft power of the animal which is associated with increasing crop production. This implies that, increasing animal and crop production significantly contributes to eradicate poverty in the area. In general, this study contributes to the existing literature in two ways: 1) The genetic controls of Sheko against trypanosomiasis have not been well studied and this study examines the genomic signatures in response to trypanosomiasis in detail; 2) this study presents pathways and master regulators which could help to understand the upstream biological processes involved in trypanotolerance. Particularly, this study for the first time identifies the master regulators involved in the regulatory mechanisms of trypanotolerance in relation to signatures of selection not only for Sheko breed but also in the context of cattle genomics, which can be used for the development of effective new drugs. However, additional studies such as differential expressions of targeted genes and regulatory molecules may be required to further confirm the validity of the results reported in this paper.

## DATA AVAILABILITY STATEMENT

The SNP data in this study can be found in the European Variation Archive (EVA): PRJEB34751.

## ETHICS STATEMENT

Standard techniques were used to collect blood. The procedure was reviewed and approved by the University of Edinburgh Ethics Committee (reference number OS 03-06) and also by the Institute Animal Care and Use Committee of the International Livestock Research Institute, Nairobi.

## AUTHOR CONTRIBUTIONS

YM, MG, and AS participated in the design of the study. YM conducted computational and statistical analyses as well as identified the signature genes. AS and MG supervised the computational and statistical analyses. YM interpreted the results with MG. YM carried out the literature survey and prepared the first draft of the manuscript. OH and KE were involved in the interpretation of the results. YM and KE were involved in collecting blood and tissue samples for this study. YM prepared the DNA samples. YM and AS were involved in the preparation of the genotyping data. YM and MG wrote the final version of the manuscript. YM, AS, and MG conceived and managed the project. All authors read and approved the final manuscript.

## ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation and Open Access Publication Funds of the Göttingen University. We thank Henner Simianer for his helpful advise and insights at the early stages of this project and for comments on the methods. We would like to thank Tariku Abena for his support during the preparation of DNA samples. We would also like to thank Felix Heinrich and Faisal Ramzan for proofreading the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01095/full#supplementary-material

**SUPPLEMENTARY FIGURE 1 |** Cross validation error in relation to the number of hypothetical ancestral populations for the Admixture analyses.

**SUPPLEMENTARY FIGURE 2 |** Histogram of standardized *Rsb* and *iHS* values.

**SUPPLEMENTARY FIGURE 3 |** Box plot of CLR -log (*P*-values).

**SUPPLEMENTARY TABLE 1 |** Proportion of admixture within each of the analyzed breeds.

**SUPPLEMENTARY TABLE 2 |** Functionally annotated gene list identified by *iHS* analysis.

**SUPPLEMENTARY TABLE 3 |** Functionally annotated gene list identified by CLR analysis.

**SUPPLEMENTARY TABLE 4 |** Functionally annotated gene list identified by *Rsb* analysis.

**SUPPLEMENTARY TABLE 5 |** Functionally enriched gene list identified by *iHS* analysis.

**SUPPLEMENTARY TABLE 6 |** Functionally enriched gene list identified by CLR analysis.

**SUPPLEMENTARY TABLE 7 |** Functionally enriched gene list identified by *Rsb* analysis.

**SUPPLEMENTARY TABLE 8 |** List of intergenic variants identified by *iHS* analysis.

**SUPPLEMENTARY TABLE 9 |** List of intergenic variants identified by CLR analysis.

**SUPPLEMENTARY TABLE 10 |** List of intergenic variants identified by *Rsb* analysis.

**SUPPLEMENTARY TABLE 11 |** Summary of the genomic regions identified by *iHS*, CLR and *Rsb*.

**SUPPLEMENTARY TABLE 12 |** QTL regions overlapping between Sheko and N'Dama.

**SUPPLEMENTARY TABLE 13 |** Summary of the blood and tissue samples collected from indigenous Ethiopian cattle breeds for DNA extraction.

## REFERENCES


Abebe, G. (2005). Current situation of Trypanosomiasis: In review article on: Trypanosomosis in Ethiopia. *Ethiop. J. Biol. Sci.* 4, 75–121. doi: 10.4314/ejbs.v4i1.39017

Acuto, O., Di Bartolo, V., and Michel, F. (2008). Tailoring T-cell receptor signals by proximal negative feedback mechanisms. *Nat. Rev. Immunol.* 8, 699–712. doi: 10.1038/nri2397

Adachi, M., Sekiya, M., Arimura, Y., Takekawa, M., Itoh, F., and Hinoda, Y. (1992). Protein-tyrosine phosphatase expression in pre-B cell NALM-6. *Cancer Res.* 52, 737–740.

Aki, D., Zhang, W., and Liu, Y. C. (2015). The E3 ligase Itch in immune regulation and beyond. *Immunol. Rev.* 266, 6–26. doi: 10.1111/imr.12301

Alexander, D., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Allam, L., Ogwu, D., Agbede, R., and Sackey, A. (2011). Hematological and serum biochemical changes in gilts experimentally infected with xperimentally infected with Trypanosoma brucei. *Vet. Arhiv.* 81, 597–609. doi: 10.1016/j.tvjl.2011.04.021

Alonso, A., Saxena, M., Williams, S., and Mustelin, T. (2001). Inhibitory role for dual specificity phosphatase VHR in T cell antigen receptor and CD28-induced Erk and Jnk activation. *J. Biol. Chem.* 276, 4766–4771. doi: 10.1074/jbc.M006497200

Andrew, A. H. (2004).*Bovine Medicine Diseases and Husbandry of Cattle*. 9600 Garsington Road, Oxford OX42DQ, UKs: Blackwell Science Ltd.

Au-Yeung, B. B., Deindl, S., Hsu, L. Y., Palacios, E. H., Levin, S. E., and Kuriyan, J. (2009). The structure, regulation, and function of ZAP-70. *Immunol. Rev.* 228, 41–57. doi: 10.1111/j.1600-065X.2008.00753.x

Bach, T. L., Chen, Q. M., Kerr, W. T., Wang, Y., Lian, L., and Choi, J. K. (2007). Phospholipase cbeta is critical for T cell chemotaxis. *J. Immunol.* 179, 2223–2227. doi: 10.4049/jimmunol.179.4.2223

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

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

Barrett, A. D., and Stanberry, L. R. (2009).*Vaccines for biodefense and emerging and neglected diseases*. (London UK: Academic Press).

Bastepe, M. (2008). The GNAS locus and pseudohypoparathyroidism. *Adv. Exp. Med. Biol.* 626, 27–40. doi: 10.1007/978-0-387-77576-0_3

Batista, J. S., Riet-Correa, F., Teixeira, M. M., Madruga, C. R., Simoes, S. D., and Maia, T. F. (2007). Trypanosomiasis by Trypanosoma vivax in cattle in the Brazilian semiarid: description of an outbreak and lesions in the nervous system. *Vet. Parasitol.* 143, 174–181. doi: 10.1016/j.vetpar.2006.08.017

Batista, J. S., Rodrigues, C. M., Garcia, H. A., Bezerra, F. S., Olinda, R. G., and Teixeira, M. M. (2011). Association of Trypanosoma vivax in extracellular sites with central nervous system lesions and changes in cerebrospinal fluid in experimentally infected goats. *Vet. Res.* 42, 63. doi: 10.1186/1297-9716-42-63

Battino, M., Giampieri, F., Pistollato, F., Sureda, A., de Oliveira, M. R., and Pittala, V. (2018). Nrf2 as regulator of innate immunity: a molecular Swiss army knife!. *Biotechnol. Adv.* 36, 358–370. doi: 10.1016/j.biotechadv.2017.12.012

Benito, A., Silva, M., Grillot, D., Nunez, G., and Fernandez-Luna, J. L. (1996). Apoptosis induced by erythroid differentiation of human leukemia cell lines is inhibited by Bcl-XL. *Blood* 87, 3837–3843. doi: 10.1182/blood.V87.9.3837.bloodjournal8793837

Berthier, D., Brenière, S. F., Bras-Gonçalves, R., Lemesre, J.-L., Jamonneau, V., and Solano, P. (2016). Tolerance to trypanosomatids: a threat, or a key for disease elimination?. *Trends In Parasitology* 32, 157–168. doi: 10.1016/j.pt.2015.11.001

Blasius, A. L., and Beutler, B. (2010). Intracellular toll-like receptors. *Immun.* 32, 305–315. doi: 10.1016/j.immuni.2010.03.012

Bomba, L., Nicolazzi, E. L., Milanesi, M., Negrini, R., Mancini, G., and Biscarini, F. (2015). Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genet. Selection Evol.* 47, 25. doi: 10.1186/s12711-015-0113-9

Bonfiglio, S., Ginja, C., De Gaetano, A., Achilli, A., Olivieri, A., and Colli, L. (2012). Origin and spread of Bos taurus: new clues from mitochondrial genomes belonging to haplogroup T1. *PloS One* 7, e38601. doi: 10.1371/journal.pone.0038601

Bosschaerts, T., Morias, Y., Stijlemans, B., Herin, M., Porta, C., and Sica, A. (2011). IL-10 limits production of pathogenic TNF by M1 myeloid cells through induction of nuclear NF-kB p50 member in Trypanosoma congolense infection-resistant C57BL/6 mice. *Eur. J. Immunol.* 41, 3270–3280. doi: 10.1002/eji.201041307

Bradley, D. G., MacHugh, D. E., Cunningham, P., and Loftus, R. T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proc. Natl. Acad. Sci. U.S.A.* 93, 5131–5135. doi: 10.1073/pnas.93.10.5131

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Caradonna, K. L., Engel, J. C., Jacobi, D., Lee, C. H., and Burleigh, B. A. (2013). Host metabolism regulates intracellular growth of Trypanosoma cruzi. *Cell Host Microbe* 13, 108–117. doi: 10.1016/j.chom.2012.11.011

Casanova, J. L., and Abel, L. (2007). Human genetics of infectious diseases: a unified theory. *EMBO J.* 26, 915–922. doi: 10.1038/sj.emboj.7601558

Chan, K., Han, X. D., and Kan, Y. W. (2001). An important function of Nrf2 in combating oxidative stress: detoxification of acetaminophen. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4611–4616. doi: 10.1073/pnas.081082098

Chanie, M., Adula, D., and Bogale, B. (2013). Socio-economic assessment of the impacts of trypanosomiasis on cattle in girja district, southern oromia region, southern ethiopia. *Acta Parasitol. Globalis* 4, 80–85. doi: 10.5829/idosi.apg.2013.4.3.7523

Charlesworth, B. (2007). A hitch-hiking guide to the genome: a commentary on 'The hitch-hiking effect of a favourable gene' by John Maynard Smith and John Haigh. *Genet. Res.* 89, 389–390. doi: 10.1017/S0016672308009580

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. *Genet.* 190, 5–22. doi: 10.1534/genetics.111.134288

Chen, H. K., Fernandez-Funez, P., Acevedo, S. F., Lam, Y. C., Kaytor, M. D., and Fernandez, M. H. (2003). Interaction of Akt-phosphorylated ataxin-1 with 14-3-3 mediates neurodegeneration in spinocerebellar ataxia type 1. *Cell* 113, 457–468. doi: 10.1016/S0092-8674(03)00349-0

Chen, T. H., Kambal, A., Krysiak, K., Walshauser, M. A., Raju, G., and Tibbitts, J. F. (2011). Knockdown of Hspa9, a del(5q31.2) gene, results in a decrease in hematopoietic progenitors in mice. *Blood* 117, 1530–1539. doi: 10.1182/blood-2010-06-293167

Cho, H. Y., Jedlicka, A. E., Reddy, S. P., Kensler, T. W., Yamamoto, M., and Zhang, L. Y. (2002). Role of NRF2 in protection against hyperoxic lung injury in mice. *Am. J. Respir. Cell Mol. Biol.* 26, 175–182. doi: 10.1165/ajrcmb.26.2.4501

Chuenkova, M. V., and PereiraPerrin, M. (2009). Trypanosoma cruzi targets Akt in host cells as an intracellular antiapoptotic strategy. *Sci. Signal* 2, ra74. doi: 10.1126/scisignal.2000374

Clark, R. I., Tan, S. W. S., Péans, C. B., Roostalu, U., Vivancos, V., and Bronda, K. (2013). MEF2 Is an In Vivo Immune-Metabolic Switch. *Cell* 155, 435–447. doi: 10.1016/j.cell.2013.09.007

Codjia, V., Mulatu, W., Majiwa, P. A., Leak, S. G., Rowlands, G. J., and Authie, E. (1993). Epidemiology of bovine trypanosomiasis in the Ghibe valley, southwest Ethiopia. 3. Occurrence of populations of Trypanosoma congolense resistant to diminazene, isometamidium and homidium. *Acta Trop.* 53, 151–163. doi: 10.1016/0001-706X(93)90026-8

Cramer, T., Yamanishi, Y., Clausen, B. E., Forster, I., Pawlinski, R., and Mackman, N. (2003). HIF-1alpha is essential for myeloid cell-mediated inflammation. *Cell* 112, 645–657. doi: 10.1016/S0092-8674(03)00154-5

DAGRIS. (2007).*Domestic Animal Genetic Resources Information System (DAGRIS)*. Rege, JHanotte, OMamo, YAsrat, BDessie, T. Addis Ababa, Ethiopia: International Livestock Research Institute (ILRI).

Daoud, H., Tetreault, M., Gibson, W., Guerrero, K., Cohen, A., and Gburek-Augustat, J. (2013). Mutations in POLR3A and POLR3B are a major cause of hypomyelinating leukodystrophies with or without dental abnormalities and/or hypogonadotropic hypogonadism. *J. Med. Genet.* 50, 194–197. doi: 10.1136/jmedgenet-2012-101357

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinf.* 32, 1895–1897. doi: 10.1093/bioinformatics/btw051

Dgany, O., Avidan, N., Delaunay, J., Krasnov, T., Shalmon, L., and Shalev, H. (2002). Congenital dyserythropoietic anemia type I is caused by mutations in codanin-1. *Am. J. Hum. Genet.* 71, 1467–1474. doi: 10.1086/344781

Dimova, D. K., and Dyson, N. J. (2005). The E2F transcriptional network: old acquaintances with new faces. *Oncogene* 24, 2810–2826. doi: 10.1038/sj.onc.1208612

Dong, C., Davis, R. J., and Flavell, R. A. (2002). MAP kinases in the immune response. *Annu. Rev. Immunol.* 20, 55–72. doi: 10.1146/annurev.immunol.20.091301.131133

Doolan, D. L., and Hoffman, S. L. (1999). IL-12 and NK cells are required for antigen-specific adaptive immunity against malaria initiated by CD8+ T cells in the Plasmodium yoelii model. *J. Immunol.* 163, 884–892.

Dorssers, L., Burger, H., Bot, F., Delwel, R., Geurts van Kessel, A. H., and Lowenberg, B. (1987). Characterization of a human multilineage-colony-stimulating factor cDNA clone identified by a conserved noncoding sequence in mouse interleukin-3. *Gene* 55, 115–124. doi: 10.1016/0378-1119(87)90254-X

Dyson, N. (1998). The regulation of E2F by pRB-family proteins. *Genes Dev.* 12, 2245–2262. doi: 10.1101/gad.12.15.2245

Eagle, R. A., and Trowsdale, J. (2007). Promiscuity and the single receptor: NKG2D. *Nat. Rev. Immunol.* 7, 737–744. doi: 10.1038/nri2144

Eguchi, K. (2001). Apoptosis in autoimmune diseases. *Internal Med.* 40, 275–284. doi: 10.2169/internalmedicine.40.275

Ersching, J., Vasconcelos, J. R., Ferreira, C. P., Caetano, B. C., Machado, A. V., Bruna-Romero, O., et al. (2016). The combined deficiency of immunoproteasome subunits affects both the magnitude and quality of Pathogen- and genetic vaccination- induced CD8+ T Cell responses to the human Protozoan parasite Trypanosoma cruzi. *PLoS Pathog.* 12 (4), e1005593

FAO. (2004).*The uganda country report as part of the state of the world's animal genetic resources (sow-angr) report*. Food and Agriculture Organization of the United Nations, Rome, Italy.

Fatihu, M., Adamu, S., Ibrahim, N., Euvie, L., and Esievo, K. (2009). The effect of experimental Trypanosoma vivax infection on the thyroid gland in Zebu bulls. *Vet. Arhiv.* 79, 429–437. doi: 10.4103%2Fijem.IJEM_12_17

Flávia Nardy, A., Freire-de Lima, C. G., and Morrot, A. (2015). Immune evasion strategies of trypanosoma cruzi. *J. Immunol. Res.* 2015. 1–7. doi: 10.1155/2015/178947

Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.* 3, 710–720. doi: 10.1038/nri1180

Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinf.* 28, 1176–1177. doi: 10.1093/bioinformatics/bts115

Gerard, C., and Rollins, B. J. (2001). Chemokines and disease. *Nat. Immunol.* 2, 108–115. doi: 10.1038/84209

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., and Gill, C. A. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Sci.* 324, 528–532. doi: 10.1126/science.1167936

Giordani, F., Morrison, L. J., Rowan, T. G., DE Koning, H. P., and Barrett, M. P. (2016). The animal trypanosomiases and their chemotherapy: a review. *Parasitology* 143, 1862–1889. doi: 10.1017/S0031182016001268

Goldszmid, R. S., and Sher, A. (2010). Processing and presentation of antigens derived from intracellular protozoan parasites. *Curr. Opin. Immunol.* 22, 118–123. doi: 10.1016/j.coi.2010.01.017

Gonçalves, V. M., Matteucci, K. C., Buzzo, C. L., Miollo, B. H., Ferrante, D., and Torrecilhas, A. C. (2013). NLRP3 controls Trypanosoma cruzi infection through a caspase-1-dependent IL-1R-independent NO production. *PloS Negl. Trop. Dis.* 7, e2469. doi: 10.1371/journal.pntd.0002469

Gong, X., Tang, X., Wiedmann, M., Wang, X., Peng, J., and Zheng, D. (2003). Cdk5-mediated inhibition of the protective effects of transcription factor MEF2 in neurotoxicity-induced apoptosis. *Neuron* 38, 33–46. doi: 10.1016/S0896-6273(03)00191-0

Gringhuis, S. I., den Dunnen, J., Litjens, M., van der Vlist, M., Wevers, B., and Bruijns, S. C. (2009). Dectin-1 directs T helper cell differentiation by controlling noncanonical NF-kappaB activation through Raf-1 and Syk. *Nat. Immunol.* 10, 203–213. doi: 10.1038/ni.1692

Grisouard, J., Hao-Shen, H., Dirnhofer, S., Wagner, K.-U., and Skoda, R. C. (2014). Selective deletion of jak2 in adult mouse hematopoietic cells leads to lethal anemia and thrombocytopenia. *Haematologica.* 99, e52–e54. doi: 10.3324/haematol.2013.100016

Guillemot, J., and Seidah, N. G. (2015). PACE4 (PCSK6): another proprotein convertase link to iron homeostasis?. *Haematologica.* 100, e377. doi: 10.3324/haematol.2015.127175

Hanke, J. H., Gardner, J. P., Dow, R. L., Changelian, P. S., Brissette, W. H., and Weringer, E. J. (1996). Discovery of a novel, potent, and Src family-selective tyrosine kinase inhibitor. Study of Lck- and FynT-dependent T cell activation. *J. Biol. Chem.* 271, 695–701. doi: 10.1074/jbc.271.2.695

Hanotte, O., Ronin, Y., Agaba, M., Nilsson, P., Gelhaus, A., and Horstmann, R. (2003). Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7443–7448. doi: 10.1073/pnas.1232392100

Hanotte, O., Tawah, C. L., Bradley, D. G., Okomo, M., Verjee, Y., and Ochieng, J. (2000). Geographic distribution and frequency of a taurine Bos taurus and an indicine Bos indicus Y specific allele amongst sub-saharan African cattle breeds. *Mol. Ecol.* 9, 387–396. doi: 10.1046/j.1365-294x.2000.00858.x

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

Haynes, M. P., Li, L., Sinha, D., Russell, K. S., Hisamoto, K., and Baron, R. (2003). Src kinase mediates phosphatidylinositol 3-kinase/Akt-dependent rapid endothelial nitric-oxide synthase activation by estrogen. *J. Biol. Chem.* 278, 2118–2123. doi: 10.1074/jbc.M210828200

He, C., Medley, S. C., Hu, T., Hinsdale, M. E., Lupu, F., and Virmani, R. (2015). PDGFRI2 signalling regulates local inflammation and synergizes with hypercholesterolaemia to promote atherosclerosis. *Nat. Commun.* 6, 7770. doi: 10.1038/ncomms8770

Heinonen, K. M., Dube, N., Bourdeau, A., Lapp, W. S., and Tremblay, M. L. (2006). Protein tyrosine phosphatase 1B negatively regulates macrophage development through CSF-1 signaling. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2776–2781. doi: 10.1073/pnas.0508563103

Hoare, C. A. (1972). “The trypanosomes of mammals,” in *A zoological monograph* (Oxford and Edinburgh: UKs: Publications BS).

Hoyt, R., Zhu, W., Cerignoli, F., Alonso, A., Mustelin, T., and David, M. (2007). Cutting edge: selective tyrosine dephosphorylation of interferon-activated nuclear STAT5 by the VHR phosphatase. *J. Immunol.* 179, 3402–3406. doi: 10.4049/jimmunol.179.6.3402

Hu, H., and Sun, S. C. (2016). Ubiquitin signaling in immune responses. *Cell Res.* 26, 457–483. doi: 10.1038/cr.2016.40

Hu, T., Ghazaryan, S., Sy, C., Wiedmeyer, C., Chang, V., and Wu, L. (2012). Concomitant inactivation of Rb and E2f8 in hematopoietic stem cells synergizes to induce severe anemia. *Blood* 119, 4532–4542. doi: 10.1182/blood-2011-10-388231

Huber, C. D., DeGiorgio, M., Hellmann, I., and Nielsen, R. (2016). Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol. Ecol.* 25, 142–156. doi: 10.1111/mec.13351

Ingley, E. (2012). Functions of the Lyn tyrosine kinase in health and disease. *Cell Commun. Signal* 10, 21. doi: 10.1186/1478-811X-10-21

Jabara, H. H., Boyden, S. E., Chou, J., Ramesh, N., Massaad, M. J., and Benson, H. (2016). A missense mutation in TFRC, encoding transferrin receptor 1, causes combined immunodeficiency. *Nat. Genet.* 48, 74–78. doi: 10.1038/ng.3465

Jackson, P. K. (2012). TTBK2 kinase: linking primary cilia and cerebellar ataxias. *Cell.* 151, 697–699. doi: 10.1016/j.cell.2012.10.027

Jaggy, A., Oliver, J. E., Ferguson, D. C., Mahaffey, E. A., and Glaus, T. (1994). Neurological manifestations of hypothyroidism: a retrospective study of 29 dogs. *J. Vet. Intern. Med.* 8, 328–336. doi: 10.1111/j.1939-1676.1994.tb03245.x

Jantsch, J., Wiese, M., Schodel, J., Castiglione, K., Glasner, J., and Kolbe, S. (2011). Toll-like receptor activation and hypoxia use distinct signaling pathways to stabilize hypoxia-inducible factor 1α (HIF1A) and result in differential HIF1A-dependent gene expression. *J. Leukoc. Biol.* 90, 551–562. doi: 10.1189/jlb.1210683

Johnson, K. R., Gagnon, L. H., and Chang, B. (2016). A hypomorphic mutation of the gamma-1 adaptin gene (Ap1g1) causes inner ear, retina, thyroid, and testes abnormalities in mice. *Mamm. Genome* 27, 200–212. doi: 10.1007/s00335-016-9632-0

Jordan, K. A., and Hunter, C. A. (2010). Regulation of CD8+ T cell responses to infection with parasitic protozoa. *Exp. Parasitol.* 126, 318–325. doi: 10.1016/j.exppara.2010.05.008

Kaminski, W. E., Lindahl, P., Lin, N. L., Broudy, V. C., Crosby, J. R., and Hellstrom, M. (2001). Basis of hematopoietic defects in platelet-derived growth factor (PDGF)-B and PDGF beta-receptor null mice. *Blood* 97, 1990–1998. doi: 10.1182/blood.V97.7.1990

Kanazawa, I. (1999). Molecular pathology of dentatorubral-pallidoluysian atrophy. *Philos. Trans. R. Soc Lond. B. Biol. Sci.* 354, 1069–1074. doi: 10.1098/rstb.1999.0460

Kawai, T., and Akira, S. (2007). Antiviral signaling through pattern recognition receptors. *J. Biochem.* 141, 137–145. doi: 10.1093/jb/mvm032

Kierstein, S., Noyes, H., Naessens, J., Nakamura, Y., Pritchard, C., and Gibson, J. (2006). Gene expression profiling in a mouse model for African trypanosomiasis. *Genes Immun.* 7, 667–679. doi: 10.1038/sj.gene.6364345

Kim, C. H., and Broxmeyer, H. E. (1999). Chemokines: signal lamps for trafficking of T and B cells for development and effector function. *J. Leukoc. Biol.* 65, 6–15. doi: 10.1002/jlb.65.1.6

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genet.* 160, 765–777.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., and Spudich, G. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011, bar030. doi: 10.1093/database/bar030

Kristjanson, P. M., Swallow, B.M., Rowlands, G.J., Kruska, R.L., and De Leeuw, P.N. (1999). Measuring the costs of African animal trypanosomosis, the potential benefits of control and returns to research. *Agric. Syst*. 59 (7), 79–98. doi: 10.1016/S0308-521X(98)00086-9

Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., and Kronenberg, D. (2006). Transpath®: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* 34, D546–D551. doi: 10.1093/nar/gkj107

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., and Wang, Z. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

La Greca, F., Haynes, C., Stijlemans, B., De Trez, C., and Magez, S. (2014). Antibody-mediated control of Trypanosoma vivax infection fails in the absence of tumour necrosis factor. *Parasite Immunol.* 36, 271–276. doi: 10.1111/pim.12106

Lamkanfi, M., Festjens, N., Declercq, W., Vanden Berghe, T., and Vandenabeele, P. (2007). Caspases in cell survival, proliferation and differentiation. *Cell Death Differ.* 14, 44–55. doi: 10.1038/sj.cdd.4402047

Lanier, L. L. (2015). NKG2D Receptor and Its Ligands in Host Defense. *Cancer Immunol. Res.* 3, 575–582. doi: 10.1158/2326-6066.CIR-15-0098

Leach, R. M., and Treacher, D. F. (1998). Oxygen transport-2. Tissue hypoxia. *BMJ* 317, 1370–1373. doi: 10.1136/bmj.317.7169.1370

Leak, S. G., Peregrine, A. S., Mulatu, W., Rowlands, G. J., and D'Ieteren, G. (1996). Use of insecticide-impregnated targets for the control of tsetse flies (Glossina spp.) and trypanosomiasis occurring in cattle in an area of south-west Ethiopia with a high prevalence of drug-resistant trypanosomes. *Trop. Med. Int. Health* 1, 599–609. doi: 10.1111/j.1365-3156.1996.tb00085.x

Lee, J. M., Chan, K., Kan, Y. W., and Johnson, J. A. (2004). Targeted disruption of Nrf2 causes regenerative immune-mediated hemolytic anemia. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9751–9756. doi: 10.1073/pnas.0403620101

Lemecha, H., Mulatu, W., Hussein, I., Rege, E., Tekle, T., and Abdicho, S. (2006). Response of four indigenous cattle breeds to natural tsetse and trypanosomosis challenge in the Ghibe valley of Ethiopia. *Vet. Parasitol.* 141, 165–176. doi: 10.1016/j.vetpar.2006.04.035

Liao, S. H., Zhao, X. Y., Han, Y. H., Zhang, J., Wang, L. S., and Xia, L. (2009). Proteomics-based identification of two novel direct targets of hypoxia-inducible factor-1 and their potential roles in migration/invasion of cancer cells. *Proteomics* 9, 3901–3912. doi: 10.1002/pmic.200800922

Linnekin, D., DeBerry, C. S., and Mou, S. (1997). Lyn associates with the juxtamembrane region of c-Kit and is activated by stem cell factor in hematopoietic cell lines and normal progenitor cells. *J. Biol. Chem.* 272, 27450–27455. doi: 10.1074/jbc.272.43.27450

Liu, G., Burns, S., Huang, G., Boyd, K., Proia, R. L., and Flavell, R. A. (2009). The receptor S1P1 overrides regulatory T cell-mediated immune suppression through Akt-mTOR. *Nat. Immunol.* 10, 769–777. doi: 10.1038/ni.1743

Liu, Q. R., Zhang, P. W., Zhen, Q., Walther, D., Wang, X. B., and Uhl, G. R. (2002). KEPI, a PKC-dependent protein phosphatase 1 inhibitor regulated by morphine. *J. Biol. Chem.* 277, 13312–13320. doi: 10.1074/jbc.M107558200

Lodish, H., Berk, A., Matsudaira, P., C., K., Krieger, M., and Scott, M. (2004). *Molecular Cell Biology*. 41 Madison Avenue, New York, USA: W. H. Freeman and Company.

Lutje, V., Taylor, K. A., Kennedy, D., Authie, E., Boulange, A., and Gettinby, G. (1996). Trypanosoma congolense: a comparison of T-cell-mediated responses in lymph nodes of trypanotolerant and trypanosusceptible cattle during primary infection. *Exp. Parasitol.* 84, 320–329. doi: 10.1006/expr.1996.0120

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity (Edinb)* 115, 426–436. doi: 10.1038/hdy.2015.42

Mabley, J. G., and Szabo, C. (2008). Inflammatory disease and sunlight: the vitamin D-poly (ADP-ribose) polymerase connection. *Future Rheumatol.* 3, 169–181. doi: 10.2217/17460816.3.2.169

Magona, J. W., Walubengo, J., and Odimim, J. J. (2004). Differences in susceptibility to trypanosome infection between Nkedi Zebu and Ankole cattle, under field conditions in Uganda. *Ann. Trop. Med. Parasitol.* 98, 785–792. doi: 10.1179/000349804225021532

Mamoudou, A., Njanloga, A., Hayatou, A., Suh, P. F., and Achukwi, M. D. (2016). Animal trypanosomosis in clinically healthy cattle of north Cameroon: epidemiological implications. *Parasit Vectors.* 9, 206. 5600, 1912–1934 doi: 10.1186/s13071-016-1498-1

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298 (5600), 1912–1934. doi: 10.1126/science.1075762

Martin-Granados, C., Prescott, A. R., Le Sommer, S., Klaska, I. P., Yu, T., and Muckersie, E. (2015). A key role for PTP1B in dendritic cell maturation, migration, and T cell activation. *J. Mol. Cell Biol.* 7, 517–528. doi: 10.1093/jmcb/mjv032

Matilla-Duenas, A. (2012). The ever expanding spinocerebellar ataxias. *Editorial Cerebellum* 11, 821–827. doi: 10.1007/s12311-012-0376-4

Mbole-Kariuki, M. N., Sonstegard, T., Orth, A., Thumbi, S. M., Bronsvoort, B. M., and Kiara, H. (2014). Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb).* 113, 297–305. doi: 10.1038/hdy.2014.31

McGuire, S. E., and McGuire, A. L. (2008). Don't throw the baby out with the bathwater: enabling a bottom-up approach in genome-wide association studies. *Genome Res.* 18, 1683–1685. doi: 10.1101/gr.083584.108

McNab, F., Mayer-Barber, K., Sher, A., Wack, A., and O'Garra, A. (2015). Type I interferons in infectious disease. *Nat. Rev. Immunol.* 15, 87–103. doi: 10.1038/nri3787

McNamee, E. N., Korns Johnson, D., Homann, D., and Clambey, E. T. (2013). Hypoxia and hypoxia-inducible factors as regulators of T cell development, differentiation, and function. *Immunol. Res.* 55, 58–70. doi: 10.1007/s12026-012-8349-8

Mendes-da Cruz, D. A., Silva, J. S., Cotta-de Almeida, V., and Savino, W. (2006). Altered thymocyte migration during experimental acute Trypanosoma cruzi infection: combined role of fibronectin and the chemokines CXCL12 and CCL4. *Eur. J. Immunol.* 36, 1486–1493. doi: 10.1002/eji.200535629

Meyer, A., Holt, H. R., Oumarou, F., Chilongo, K., Gilbert, W., and Fauron, A. (2018). Integrated cost-benefit analysis of tsetse control and herd productivity to inform control programs for animal african trypanosomiasis. *Parasites Vectors* 11, 154. doi: 10.1186/s13071-018-2679-x

Minchenko, O. H., Tsymbal, D. O., Minchenko, D. O., Kovalevska, O. V., Karbovskyi, L. L., and Bikfalvi, A. (2015). Inhibition of ERN1 signaling enzyme affects hypoxic regulation of the expression of E2F8, EPAS1, HOXC6, ATF3, TBX3 and FOXF1 genes in U87 glioma cells. *Ukr Biochem. J.* 87, 76–87. doi: 10.15407/ubj87.02.076

MoARD. (2004). Ministry Of Agriculture and Rural Development of the Government of Ethiopia (MoARD). Tsetse and trypanosomiasis prevention and control strategies. Paper presented on Farming In Tsetse Controlled Areas (FITCA), Ethiopia final workshop. Adama, Ethiopia .

Molenaar, P. C., Newsom-Davis, J., Polak, R. L., and Vincent, A. (1982). Eaton-Lambert syndrome: acetylcholine and choline acetyltransferase in skeletal muscle. *Neurol.* 32, 1061–1065. doi: 10.1212/WNL.32.9.1061

Morrison, L. J., Vezza, L., Rowan, T., and Hope, J. C. (2016). Animal African trypanosomiasis: time to increase focus on clinically relevant parasite and host species. *Trends Parasitol.* 32, 599–607. doi: 10.1016/j.pt.2016.04.012

Morrot, A., and Zavala, F. (2004). Effector and memory CD8+ T cells as seen in immunity to malaria. *Immunol. Rev.* 201, 291–303. doi: 10.1111/j.0105-2896.2004.00175.x

Moura, I. C., Centelles, M. N., Arcos-Fajardo, M., Malheiros, D. M., Collawn, J. F., and Cooper, M. D. (2001). Identification of the transferrin receptor as a novel immunoglobulin (Ig)A1 receptor and its enhanced expression on mesangial cells in IgA nephropathy. *J. Exp. Med.* 194, 417–425. doi: 10.1084/jem.194.4.417

Muhanguzi, D., Mugenyi, A., Bigirwa, G., Kamusiime, M., Kitibwa, A., and Akurut, G. G. (2017). African animal trypanosomiasis as a constraint to livestock health and production in Karamoja region: a detailed qualitative and quantitative assessment. *BMC Vet. Res.* 13, 355. doi: 10.1186/s12917-017-1285-z

Muhanguzi, D., Picozzi, K., Hatendorf, J., Thrusfield, M., Welburn, S. C., and Kabasa, J. D. (2014). Improvements on restricted insecticide application protocol for control of Human and Animal African Trypanosomiasis in eastern Uganda. *PloS Negl. Trop. Dis.* 8, e3284. doi: 10.1371/journal.pntd.0003284

Mulugeta, W., Wilkes, J., Mulatu, W., Majiwa, P. A., Masake, R., and Peregrine, A. S. (1997). Long-term occurrence of Trypanosoma congolense resistant to diminazene, isometamidium and homidium in cattle at Ghibe, Ethiopia. *Acta Trop.* 64, 205–217. doi: 10.1016/S0001-706X(96)00645-6

Murray, M., Trail, J. C., and D'Ieteren, G. D. (1990). Trypanotolerance in cattle and prospects for the control of trypanosomiasis by selective breeding. *Rev. - Off. Int. Epizoot.* 9, 369–386. doi: 10.20506/rst.9.2.506

Mwai, O., Hanotte, O., Kwon, Y.-J., and Cho, S. (2015). African indigenous cattle: unique genetic resources in a rapidly changing world. *Asian-Australasian J. Anim. Sci.* 28, 911. doi: 10.5713/ajas.15.0002R

Naessens, J. (2006). Bovine trypanotolerance: A natural ability to prevent severe anaemia and haemophagocytic syndrome?. *Int. J. Parasitol.* 36, 521–528. doi: 10.1016/j.ijpara.2006.02.012

Nantulya, V. M. (1986). Immunological approaches to the control of animal trypanosomiasis. *Parasitol. Today (Regul. Ed.)* 2, 168–173. doi: 10.1016/0169-4758(86)90148-1

Naula, C., and Burchmore, R. (2003). A plethora of targets, a paucity of drugs: progress towards the development of novel chemotherapies for human African trypanosomiasis. *Expert Rev. Anti Infect. Ther.* 1, 157–165. doi: 10.1586/14787210.1.1.157

Nevins, J. R. (1998). Toward an understanding of the functional complexity of the E2F and retinoblastoma families. *Cell Growth Differ.* 9, 585–593.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575. doi: 10.1101/gr.4252305

Noyes, H., Brass, A., Obara, I., Anderson, S., Archibald, A. L., and Bradley, D. G. (2011). Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9304–9309. doi: 10.1073/pnas.1013486108

NTTICC. (2004). National Tsetse and Trypanosomosis Investigation and Control Center. Report for the period 7th June 2003 to 6th July 2004. Bedele, Ethiopia. 21-24 .

Ocaido, M., Otim, C. P., Okuna, N. M., Erume, J., Ssekitto, C., and Wafula, R. Z. O., et al (2005). Socio-economic and livestock disease survey of agro-pastoral communities in Serere County, Soroti District, Uganda. Livestock Research for Rural Development 17 .

Ohtsuka, R., Abe, Y., Fujii, T., Yamamoto, M., Nishimura, J., and Takayanagi, R. (2007). Mortalin is a novel mediator of erythropoietin signaling. *Eur. J. Haematol.* 79, 114–125. doi: 10.1111/j.1600-0609.2007.00870.x

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc Lond. B. Biol. Sci.* 365, 185–205. doi: 10.1098/rstb.2009.0219

Oosthuyse, B., Moons, L., Storkebaum, E., Beck, H., Nuyens, D., and Brusselmans, K. (2001). Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat. Genet.* 28, 131–138. doi: 10.1038/88842

Packham, G., White, E. L., Eischen, C. M., Yang, H., Parganas, E., and Ihle, J. N. (1998). Selective regulation of Bcl-XL by a Jak kinase-dependent pathway is bypassed in murine hematopoietic malignancies. *Genes Dev.* 12, 2475–2487. doi: 10.1101/gad.12.16.2475

Patterson, H., Nibbs, R., McInnes, I., and Siebert, S. (2014). Protein kinase inhibitors in the treatment of inflammatory and autoimmune diseases. *Clin. Exp. Immunol.* 176, 1–10. doi: 10.1111/cei.12248

Paul, W. E., Steinman, R., Beutler, B., and Hoffmann, J. (2011). Bridging innate and adaptive immunity. *Cell* 147, 1212–1215. doi: 10.1016/j.cell.2011.11.036

Ponce, N. E., Cano, R. C., Carrera-Silva, E. A., Lima, A. P., Gea, S., and Aoki, M. P. (2012). Toll-like receptor-2 and interleukin-6 mediate cardiomyocyte protection from apoptosis during Trypanosoma cruzi murine infection. *Med. Microbiol. Immunol.* 201, 145–155. doi: 10.1007/s00430-011-0216-z

Pulst, S. M. (2016). Genetics of neurodegenerative diseases. *Neurol. Genet.* 2, e52. doi: 10.1212/NXG.0000000000000052

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., and Bender, D. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

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

Qiu, L., Joazeiro, C., Fang, N., Wang, H. Y., Elly, C., and Altman, Y. (2000). Recognition and ubiquitination of Notch by Itch, a hect-type E3 ubiquitin ligase. *J. Biol. Chem.* 275, 35734–35737. doi: 10.1074/jbc.M007300200

Ran, Q., Wadhwa, R., Kawai, R., Kaul, S. C., Sifers, R. N., and Bick, R. J. (2000). Extramitochondrial localization of mortalin/mthsp70/PBP74/GRP75. *Biochem. Biophys. Res. Commun.* 275, 174–179. doi: 10.1006/bbrc.2000.3237

Rege, J. (1999). The state of african cattle genetic resources i. classification framework and identification of threatened and extinct breeds. *Anim. Genet. Resources/Resources génétiques Animales/Recursos genéticos animales* 25, 1–25. doi: 10.1017/S1014233900003448

Renella, R., Roberts, N. A., Brown, J. M., De Gobbi, M., Bird, L. E., and Hassanali, T. (2011). Codanin-1 mutations in congenital dyserythropoietic anemia type 1 affect HP1alpha localization in erythroblasts. *Blood* 117, 6928–6938. doi: 10.1182/blood-2010-09-308478

Rius, J., Guma, M., Schachtrup, C., Akassoglou, K., Zinkernagel, A. S., and Nizet, V. (2008). NF-kappaB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1alpha. *Nat.* 453, 807–811. doi: 10.1038/nature06905

Rosenzweig, S. D., and Holland, S. M. (2005). Defects in the interferon-gamma and interleukin-12 pathways. *Immunol. Rev.* 203, 38–47. doi: 10.1111/j.0105-2896.2005.00227.x

Rouault, T. A. (2006). The role of iron regulatory proteins in mammalian iron homeostasis and disease. *Nat. Chem. Biol.* 2, 406–414. doi: 10.1038/nchembio807

Roux, P. P., and Blenis, J. (2004). ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions. *Microbiol. Mol. Biol. Rev.* 68, 320–344. doi: 10.1128/MMBR.68.2.320-344.2004

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

Saleh, M. A., Al-Salahy, M. B., and Sanousi, S. A. (2009). Oxidative stress in blood of camels (Camelus dromedaries) naturally infected with Trypanosoma evansi. *Vet. Parasitol.* 162, 192–199. doi: 10.1016/j.vetpar.2009.03.035

Salim, B., Taha, K. M., Hanotte, O., and Mwacharo, J. M. (2014). Historical demographic profiles and genetic variation of the East African Butana and Kenana indigenous dairy zebu cattle. *Anim. Genet.* 45, 782–790. doi: 10.1111/age.12225

Salmond, R. J., Filby, A., Qureshi, I., Caserta, S., and Zamoyska, R. (2009). T-cell receptor proximal signaling via the Src-family kinases, Lck and Fyn, influences T-cell activation, differentiation, and tolerance. *Immunol. Rev.* 228, 9–22. doi: 10.1111/j.1600-065X.2008.00745.x

Sangokoya, C., Telen, M. J., and Chi, J. T. (2010). microRNA miR-144 modulates oxidative stress tolerance and associates with anemia severity in sickle cell disease. *Blood* 116, 4338–4348. doi: 10.1182/blood-2009-04-214817

Schiffmann, R., and van der Knaap, M. S. (2009). Invited article: an MRI-based approach to the diagnosis of white matter disorders. *Neurol.* 72, 750–759. doi: 10.1212/01.wnl.0000343049.00540.c8

Schmitz-Abe, K., Ciesielski, S. J., Schmidt, P. J., Campagna, D. R., Rahimov, F., and Schilke, B. A. (2015). Congenital sideroblastic anemia due to mutations in the mitochondrial HSP70 homologue HSPA9. *Blood* 126, 2734–2738. doi: 10.1182/blood-2015-09-659854

Shalini, S., Dorstyn, L., Dawar, S., and Kumar, S. (2015). Old, new and emerging functions of caspases. *Cell Death Different.* 22, 526. doi: 10.1038/cdd.2014.216

Shaw, A. P., Cecchi, G., Wint, G. R., Mattioli, R. C., and Robinson, T. P. (2014). Mapping the economic benefits to livestock keepers from intervening against bovine trypanosomosis in Eastern Africa. *Prev. Vet. Med.* 113, 197–210. doi: 10.1016/j.prevetmed.2013.10.024

She, H., Yang, Q., and Mao, Z. (2012). Neurotoxin-induced selective ubiquitination and regulation of MEF2A isoform in neuronal stress response. *J. Neurochem.* 122, 1203–1210. doi: 10.1111/j.1471-4159.2012.07860.x

She, H., Yang, Q., Shepherd, K., Smith, Y., Miller, G., and Testa, C. (2011). Direct regulation of complex I by mitochondrial MEF2D is disrupted in a mouse model of Parkinson disease and in human patients. *J. Clin. Invest.* 121, 930–940. doi: 10.1172/JCI43871

Silva, B., and Faustino, P. (2015). An overview of molecular basis of iron metabolism regulation and the associated pathologies. *Biochim. Biophys. Acta* 1852, 1347–1359. doi: 10.1016/j.bbadis.2015.03.011

Singh, B., Soltys, B. J., Wu, Z. C., Patel, H. V., Freeman, K. B., and Gupta, R. S. (1997). Cloning and some novel characteristics of mitochondrial Hsp70 from Chinese hamster cells. *Exp. Cell Res.* 234, 205–216. doi: 10.1006/excr.1997.3609

Singh, Y., Garden, O. A., Lang, F., and Cobb, B. S. (2016). MicroRNAs regulate T-cell production of interleukin-9 and identify hypoxia-inducible factor-2Iα as an important regulator of T helper 9 and regulatory T-cell differentiation. *Immunol.* 149, 74–86. doi: 10.1111/imm.12631

Sinshaw, A., Abebe, G., Desquesnes, M., and Yoni, W. (2006). Biting flies and Trypanosoma vivax infection in three highland districts bordering lake Tana, Ethiopia. *Vet. Parasitol.* 142, 35–46. doi: 10.1016/j.vetpar.2006.06.032

Skipper, M. (2002). Human genetics: Tracking positive selection. *Nat. Rev. Genet.* 3, 824. doi: 10.1038/nrg942

Slingenbergh, J. (1992). Tsetse control and agricultural development in Ethiopia. *World Anim. Rev.* 70-71, 30–36.

Song, J., Salek-Ardakani, S., So, T., and Croft, M. (2007). The kinases aurora B and mTOR regulate the G1-S cell cycle progression of T lymphocytes. *Nat. Immunol.* 8, 64–73. doi: 10.1038/ni1413

Springer, T. A. (1990). Adhesion receptors of the immune system. *Nat.* 346, 425. doi: 10.1038/346425a0

Stahl, P., Ruppert, V., Meyer, T., Schmidt, J., Campos, M. A., and Gazzinelli, R. T. (2013). Trypomastigotes and amastigotes of Trypanosoma cruzi induce apoptosis and STAT3 activation in cardiomyocytes in vitro. *Apoptosis* 18, 653–663. doi: 10.1007/s10495-013-0822-x

Stanojevic, V., Habener, J. F., Holz, G. G., and Leech, C. A. (2008). Cytosolic adenylate kinases regulate K-ATP channel activity in human beta-cells. *Biochem. Biophys. Res. Commun.* 368, 614–619. doi: 10.1016/j.bbrc.2008.01.109

Steverding, D. (2008). The history of African trypanosomiasis. *Parasit Vectors* 1, 3. doi: 10.1186/1756-3305-1-3

Stijlemans, B., Vankrunkelsven, A., Caljon, G., Bockstal, V., Guilliams, M., and Bosschaerts, T. (2010). The central role of macrophages in trypanosomiasis-associated anemia: rationale for therapeutical approaches. *Endocr. Metab. Immune Disord. Drug Targets* 10, 71–82. doi: 10.2174/187153010790827966

Suzuki, Y., and Yazawa, I. (2011). Pathological accumulation of atrophin-1 in dentatorubralpallidoluysian atrophy. *Int. J. Clin. Exp. Pathol.* 4, 378–384.

Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PloS Biol.* 5, e171. doi: 10.1371/journal.pbio.0050171

Tavazzi, B., Di Pierro, D., Amorini, A. M., Fazzina, G., Tuttobene, M., and Giardina, B. (2000). Energy metabolism and lipid peroxidation of human erythrocytes as a function of increased oxidative stress. *Eur. J. Biochem.* 267, 684–689. doi: 10.1046/j.1432-1327.2000.01042.x

Taylor, K. A. (1998). Immune responses of cattle to African trypanosomes: protective or pathogenic?. *Int. J. Parasitol.* 28, 219–240. doi: 10.1016/S0020-7519(97)00154-9

Tomasec, P., Wang, E. C., Groh, V., Spies, T., McSharry, B. P., and Aicheler, R. J. (2007). Adenovirus vector delivery stimulates natural killer cell recognition. *J. Gen. Virol.* 88, 1103–1108. doi: 10.1099/vir.0.82685-0

Traves, P. G., Pardo, V., Pimentel-Santillana, M., Gonzalez-Rodriguez, A., Mojena, M., and Rico, D. (2014). Pivotal role of protein tyrosine phosphatase 1B (PTP1B) in the macrophage response to pro-inflammatory and anti-inflammatory challenge. *Cell Death Dis.* 5, e1125. doi: 10.1038/cddis.2014.90

Trimarchi, J. M., and Lees, J. A. (2002). Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.* 3, 11–20. doi: 10.1038/nrm714

Tuntasuvan, D., Sarataphan, N., and Nishikawa, H. (1997). Cerebral trypanosomiasis in native cattle. *Vet. Parasitol.* 73, 357–363. doi: 10.1016/S0304-4017(97)00128-3

Utz, P. J., Hottelet, M., Schur, P. H., and Anderson, P. (1997). Proteins phosphorylated during stress-induced apoptosis are common targets for autoantibody production in patients with systemic lupus erythematosus. *J. Exp. Med.* 185, 843–854. doi: 10.1084/jem.185.5.843

Van Wyk, I. C., Goddard, A., de C Bronsvoort, B. M., Coetzer, J. A., Handel, I. G., and Hanotte, O. (2014). The impact of co-infections on the haematological profile of East African Short-horn Zebu calves. *Parasitology* 141, 374–388. doi: 10.1017/S0031182013001625

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

Vignali, D. A., Collison, L. W., and Workman, C. J. (2008). How regulatory T cells work. *Nat. Rev. Immunol.* 8, 523–532. doi: 10.1038/nri2343

Wang, D., Sai, J., Carter, G., Sachpatzidis, A., Lolis, E., and Richmond, A. (2002). PAK1 kinase is required for CXCL1-induced chemotaxis. *Biochem.* 41, 7100–7107. doi: 10.1021/bi025902m

Welburn, S. C., Molyneux, D. H., and Maudlin, I. (2016). Beyond tsetse-implications for research and control of human african trypanosomiasis epidemics. *Trends Parasitol.* 32, 230–241. doi: 10.1016/j.pt.2015.11.008

Wilches, R., Voigt, S., Duchen, P., Laurent, S., and Stephan, W. (2014). Fine-mapping and selective sweep analysis of QTL for cold tolerance in Drosophila melanogaster. *G3: Genes, Genomes, Genetics.* 4, 1635–1645. doi: 10.1534/g3.114.012757

Woolsey, A. M., Sunwoo, L., Petersen, C. A., Brachmann, S. M., Cantley, L. C., and Burleigh, B. A. (2003). Novel PI 3-kinase-dependent mechanisms of trypanosome invasion and vacuole maturation. *J. Cell. Sci.* 116, 3611–3622. doi: 10.1242/jcs.00666

Wu, D., and Hersh, L. B. (1994). Choline acetyltransferase: celebrating its fiftieth year. *J. Neurochem.* 62, 1653–1663. doi: 10.1046/j.1471-4159.1994.62051653.x

Yang, D., Chertov, O., Bykovskaia, S. N., Chen, Q., Buffo, M. J., and Shogan, J. (1999). Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Sci.* 286, 525–528. doi: 10.1126/science.286.5439.525

Yaro, M., Munyard, K. A., Stear, M. J., and Groth, D. M. (2016). Combatting African Animal Trypanosomiasis (AAT) in livestock: The potential role of trypanotolerance. *Vet. Parasitol.* 225, 43–52. doi: 10.1016/j.vetpar.2016.05.003

Yi, T., Cleveland, J. L., and Ihle, J. N. (1991). Identification of novel protein tyrosine phosphatases of hematopoietic cells by polymerase chain reaction amplification. *Blood* 78, 2222–2228. doi: 10.1182/blood.V78.9.2222.2222

Ymer, S., Tucker, W. Q., Sanderson, C. J., Hapel, A. J., Campbell, H. D., and Young, I. G. (1985). Constitutive synthesis of interleukin-3 by leukaemia cell line WEHI-3B is due to retroviral insertion near the gene. *Nat.* 317, 255–258. doi: 10.1038/317255a0

You, F., Sun, H., Zhou, X., Sun, W., Liang, S., and Zhai, Z. (2009). PCBP2 mediates degradation of the adaptor MAVS via the HECT ubiquitin ligase AIP4. *Nat. Immunol.* 10, 1300–1308. doi: 10.1038/ni.1815

Yuan, Z. L., Guan, Y. J., Wang, L., Wei, W., Kane, A. B., and Chin, Y. E. (2004). Central role of the threonine residue within the p+1 loop of receptor tyrosine kinase in STAT3 constitutive phosphorylation in metastatic cancer cells. *Mol. Cell. Biol.* 24, 9390–9400. doi: 10.1128/MCB.24.21.9390-9400.2004

Zelenika, D., Grima, B., and Pessac, B. (1993). A new family of transcripts of the myelin basic protein gene: expression in brain and in immune system. *J. Neurochem.* 60, 1574–1577. doi: 10.1111/j.1471-4159.1993.tb03325.x

Zermati, Y., Garrido, C., Amsellem, S., Fishelson, S., Bouscary, D., and Valensi, F. (2001). Caspase activation is required for terminal erythroid differentiation. *J. Exp. Med.* 193, 247–254. doi: 10.1084/jem.193.2.247

Zhang, R., Zhou, L., Li, Q., Liu, J., Yao, W., and Wan, H. (2009). Up-regulation of two actin-associated proteins prompts pulmonary artery smooth muscle cell migration under hypoxia. *Am. J. Respir. Cell Mol. Biol.* 41, 467–475. doi: 10.1165/rcmb.2008-0333OC

Zhang, S., Han, J., Sells, M. A., Chernoff, J., Knaus, U. G., and Ulevitch, R. J. (1995). Rho family GTPases regulate p38 mitogen-activated protein kinase through the downstream mediator Pak1. *J. Biol. Chem.* 270, 23934–23936. doi: 10.1074/jbc.270.41.23934

Zhang, Y., Wang, X., Yang, H., Liu, H., Lu, Y., and Han, L. (2013). Kinase AKT controls innate immune cell development and function. *Immunol.* 140, 143–152. doi: 10.1111/imm.12123

Zinngrebe, J., Montinaro, A., Peltzer, N., and Walczak, H. (2014). Ubiquitin in the immune system. *EMBO Rep.* 15, 28–45. doi: 10.1002/embr.201338025

**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 November 2019
doi: 10.3389/fgene.2019.01172

# The Local South American Chicken Populations Are a Melting-Pot of Genomic Diversity

*Agusto Luzuriaga-Neira[1], Lucía Pérez-Pardal[1], Sean M. O'Rourke[2], Gustavo Villacís-Rivas[3], Freddy Cueva-Castillo[3], Galo Escudero-Sánchez[4], Juan Carlos Aguirre-Pabón[1], Amarilis Ulloa-Núñez[5], Makarena Rubilar-Quezada[5], Marcelo Vallinoto[6], Michael R. Miller[2,7] and Albano Beja-Pereira[1,8]**

[1] Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO-InBIO), Universidade do Porto, Vairão, Portugal, [2] Department of Animal Science, University of California, Davis, CA, United States, [3] Centro De Biotecnología, Universidad Nacional de Loja, Loja, Ecuador, [4] Universidad Nacional de Loja, Loja, Ecuador, [5] Facultad de Ciencias Veterinarias, Universidad de Concepción, Chillán, Chile, [6] Laboratório de Evolução (LEVO), Instituto de Estudos Costeiros (IECOS), Universidade Federal do Pará, Pará, Bragança, Brazil, [7] Center for Watershed Sciences, University of California, Davis, CA, United States, [8] Departamento de Geociências, Ambiente e Ordenamento do Território (DGAOT), Faculdade de Ciências, University of Porto, Porto, Portugal

OPEN ACCESS

**Edited by:**
*Gábor Mészáros, University of Natural Resources and Life Sciences Vienna, Austria*

**Reviewed by:**
*Martin Johnsson, Swedish University of Agricultural Sciences, Sweden*
*Steffen Weigend, Friedrich Loeffler Institute (FLI), Germany*

***Correspondence:**
*Albano Beja-Pereira*
*albanobp@fc.up.pt*

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

**Received:** *30 July 2019*
**Accepted:** *23 October 2019*
**Published:** *19 November 2019*

**Citation:**
*Luzuriaga-Neira A, Pérez-Pardal L, O'Rourke SM, Villacís-Rivas G, Cueva-Castillo F, Escudero-Sánchez G, Aguirre-Pabón JC, Ulloa-Núñez A, Rubilar-Quezada M, Vallinoto M, Miller MR and Beja-Pereira A (2019) The Local South American Chicken Populations Are a Melting-Pot of Genomic Diversity. Front. Genet. 10:1172. doi: 10.3389/fgene.2019.01172*

Chicken have a considerable impact in South American rural household economy as a source of animal protein (eggs and meat) and a major role in cultural traditions (e.g., cockfighting, religious ceremonies, folklore). A large number of phenotypes and its heterogeneity are due to the multitude of environments (from arid to tropical rain forest and high altitude) and agricultural systems (highly industrialized to subsistence agriculture). This heterogeneity also represents the successive introduction of domestic chicken into this continent, which some consider predating Columbus' arrival to South America. In this study, we have used next-generation restriction site-associated DNA sequencing to scan for genome-wide variation across 145 South American chickens representing local populations from six countries of South America (Colombia, Brazil, Ecuador, Peru, Bolivia, and Chile). After quality control, the genotypes of 122,801 single nucleotide polymorphisms (SNPs) were used to assess the genomic diversity and interpopulation genetic relationship between those populations and their potential sources. The estimated population genetic diversity displayed that the gamefowl has the least diverse population ($\theta\pi = 0.86$; $\theta S = 0.70$). This population is also the most divergent ($\mathbf{F}_{ST} = 0.11$) among the South American populations. The allele-sharing analysis and the admixture analysis revealed that the current diversity displayed by these populations resulted from multiple admixture events with a strong influence of the modern commercial egg-layer chicken (ranging between 44% and 79%). It also revealed an unknown genetic component that is mostly present in the Easter Island population that is also present in local chicken populations from the South American Pacific fringe.

**Keywords: *Gallus gallus*, RADseq, population genetics, local resources, single nucleotide polymorphisms**

## INTRODUCTION

The domestic chicken, *Gallus gallus domesticus,* is a major source of animal protein (eggs and meat) and owes its popularity to low-cost production and the inexistence of any cultural or religious prohibition to its consumption. Chicken production is even more important in rural areas with economies based on subsistence agriculture. Additionally, besides being a source of food, in some regions of the globe, the chicken has been also used for cultural, religious, and entertainment proposes (Lawler, 2014).

The initiation of molecular genetic studies in the early 1990s has answered many questions regarding the origin, dispersal, and genetic diversity of many modern domestic chickens. It is now widely accepted that the red junglefowl (*Gallus gallus*) from jungles in South and Southeast Asia is considered the most probable ancestor of the domestic chicken (Fumihito et al., 1994; Fumihito et al., 1996). Historical and archaeological sources point to early domestication of the chicken, around 5,400 BC (West and Zhou, 1988; Underhill, 1997), although recent work on ancient DNA (aDNA) suggests northern China as the earliest chicken domestication site, around 8,000 BC (Xiang et al., 2014). Also, several recent genetic studies based on the mitochondrial DNA (mtDNA) variation have suggested the additional contributions of the red junglefowl from the Indian Subcontinent, South and East of China, Thailand, Myanmar, and Indonesia (e.g., see for more detail Liu et al., 2006; Miao et al., 2013).

The history of domestic animals in South America is similar to the rest of the "new world," in which the majority of the livestock species have been introduced by European colonizers from the 15$^{th}$ century onwards. Although the indigenous guinea pig and the South American camelid species have been always considered a South American domestication, some authors, mostly based on archaeological evidence (Carter, 1971; Fitzpatrick and Callaghan, 2009; Ramírez-Aliaga, 2010), have been arguing for a pre-Colombian introduction of the chicken in SA. Recently, the sequencing of the region of the mitochondrial genome from a Chilean bone dated from Ca. 1,304 to 1,424 AD suggested a pre-Columbian origin of the South American chicken (Storey et al., 2007). However, this work was contested by other authors (Gongora et al., 2008) as the mtDNA haplotype found at this site, and on which the authors argued as evidence of a Pacific origin of chicken in SA, belongs to a ubiquitous haplogroup (E) that can be found in chicken from all over the world. More recently, a study on the contemporary mtDNA diversity of several South American populations have found that although the Iberian Peninsula (European) chicken might have been the main source of the modern South American chicken, it also identified the presence of a genetic component in the Easter Island chickens that cannot be attributed to the introduction of chickens from Europe (through the Iberian Peninsula), and which is phylogenetically closer to the Southeast Asia populations (Luzuriaga-Neira et al., 2017).

Throughout time, successive waves of European colonizers have brought to South America their chicken stocks from their places of origin. With the intensification of chicken production in the twentieth century, new and highly selected and specialized breeds (e.g., egg-layers, broilers) have been created (Crawford, 1990), which have been spread worldwide at a much faster pace. However, the introgression of these highly selected and performant lineages of chicken into the local breeds has been impeded by the lower capacity of adaptation to most of the environmental conditions (e.g., temperature, parasites, predators). Most of the gene flow from the highly selected lineages has been made through F1s, in which a high performant lineage is crossed with a locally adapted breed.

In the last decade, access to next-generation sequencing (NGS) has permitted the development of more cost-effective and efficient techniques to measure variation at a genome-wide scale. NGS has permitted major advances in demographic parameters estimation as well as on the identification of genes underlying adaptation and production traits, and this in combination with phenotype data can accelerate breeding in plants and animals (e.g., review by Daetwyler et al., 2013). Thus, genome-wide variation studies can not only identify genomic regions underpinning the adaptation of certain populations to extreme environments (e.g., Zhang et al., 2016) as well as help conserving these regions while improving the productive performances of the local breeds (Thornton, 2010; Kristensen et al., 2015).

In this study, we used RADseq to scan and genotype hundreds of thousands of single-nucleotide polymorphism (SNPs) throughout the genome to characterize six SA local chicken populations from Bolivia, Brazil, Colombia, Chile (continental and Easter Island), Ecuador, and Peru. As cock-fighting has an important socio-cultural role in South America in the last centuries (Finsterbusch, 1990; Lawler, 2014), this region possesses a large number of gamefowls that have been bred separately from the others for many generations. Like the rest of the local populations, information on the origin and genetic structure of this population is very limited or unknown and for this reason we have included samples representing this population and three other populations representing old (Iberian Peninsula population) and two contemporary sources [a cosmopolitan meat production breed (broiler) and cosmopolitan egg-layers (Isa Brown)] that might have contributed for the current genetic architecture of the current South American local populations.

## MATERIALS AND METHODS

### Tissue Sampling and DNA Extraction

Approximately 2 mm$^2$ of the comb of 145 local domestic chickens were collected from six SA local populations representing: Bolivia (N = 6), Brazil (N = 4), Chile ((N = 35; 21 Mainland + 14 Easter Island), Colombia (N = 17), Ecuador (N = 16), Peru (N = 17), and gamefowl (N = 14). Individuals representing local Iberian Peninsula chicken (N = 17) as well as individuals representing commercial egg layers (N = 5; Isa Brown endproducts) and broiler (N = 15) were also sampled. Samples were stored in 95% ethanol at -20°C.

Genomic DNA was extracted using a JetQuick™ Tissue DNA Spin Kit (Genomed, GmbH) and quantified using a Qubit Fluorimeter (Thermo Fisher Scientific). RADseq sequencing libraries were prepared using the eight base-pair recognition site restriction enzyme *SbfI* (New England Biolabs, cat.# R3642L)

using a new RAD protocol (Ali et al., 2016). In brief, DNA was normalized to 5 ng/µl and 10 µl of each sample was arrayed into a well in a 96-well plate. The DNA was cut using the eight base-pair recognition site restriction enzyme *SbfI* (New England Biolabs, cat.# R3642L). After cleavage, unique barcodes were ligated on and the samples were pooled, sheared in a Bioruptor NGS (Diagenode, Belgium), and used as input for NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, USA). The libraries were sequenced on an Ilumina Hiseq 2500 using paired end 100 bp reads.

### Data Analysis

We demultiplexed the libraries filtering solely the reads having a full barcode match and a partial restriction site match. Sequences were aligned to the Galgal4 Chicken Genome assembly (International Chicken Genome Sequencing Consortium, 2004), using the BWA algorithm (Li and Durbin, 2009), with the default parameters. Ambiguously mapped and/or clonal sequences were removed using the filters for proper pairs and PCR duplicates included in the SAMtools package (Li et al., 2009). The consensus sequences were constructed and the Binary sequence/Alignment Map format files (BAM) indexed using the same software package. To avoid bias caused by variable sequencing depth, we created subsampled BAM files using the random sampling option from SAMtools. We chose 180,000 alignments from each BAM file for the subsampled set. Genotype calls were performed using ANGSD (Korneliussen et al., 2014) with a minimum map quality score (minMapQ) and a minimum base quality score (minQ) of 20. For the variant calls, we used the SAMtools genotype likelihood model (Li, 2011) and selected sites present in at least 50% of the samples (minInd). To verify the performance of our SNP calling method, we have searched the public databases (the National Center for Biotechnology Information NCBI, dbSNP database, available at https://ftp.ncbi.nih.gov/snp/organisms/archive/chicken_9031/) for matches between our variants and those already identified in genome-wide studies. SNP annotation was performed using the SnpEff 3.0 program (Cingolani et al., 2012), using the galGal4 genome version as the reference.

### Genetic Diversity

The two most common indexes of molecular genetic variation (θ)—mean pairwise differences between sequences (π; Tajima, 1989) and Watterson segregating sites (S; Watterson, 1975)—were calculated using thetaStat (ANGSD). Pairwise weighted $F_{ST}$ windows were used to measure genetic differentiation between populations (Weir and Cockerham, 1984) using the VCFtools program (Danecek et al., 2011). Additionally, for estimating the genetic relationships between the potential population sources—i.e., samples representing Iberian Peninsula, broiler, egg-layer, South American gamefowl populations—and the South American chicken populations, we have also performed variance analyses (one-way ANOVA model) by comparing each pair of populations as a factor and the weighted $F_{ST}$ value (per 50 kb sliding window) for the same pair of populations as the variable. Averages, standard error, and plots were generated using the R software (R Core Team, 2013).

To count the number of shared SNPs among South American chicken populations, we created Variant Call Format Files (VCF) for four groups of samples according to their geographical location. One group, composed by the individuals from South American countries located at Pacific fringe (Ecuador, Peru, Chile, and Bolivia), another group formed by individuals from the Atlantic fringe (Colombia and Brazil), and the potential source populations were kept in two separated groups. The number of shared variants between the groups was determined using the module vcf-compare included also in the VCFtools software, which conducts simple comparisons between VCF files. Venn diagrams (Caminsky et al., 2016; Feichtinger et al., 2016) were used to visualize private/shared variants per group. Those variants were then represented in pie charts (**Figure 1A**) representing variants in different categories: i) shared between the Pacific and Atlantic groups, ii) shared with any of the possible source populations (Egg layer, Broiler, gamefowl or Iberian Peninsula), and iii) unique to a group. Only variants displaying a ≥5% frequency per population were considered.

## Population Structure and Genetic Relationships

The $r^2$ parameter was estimated to identify SNPs in linkage disequilibrium (LD) using the software PLINK v1.9 (Purcell and Chang, 2015) for 50 kb sliding windows, over a phased file excluding SNPs with allele frequencies <0.05 and an $r^2 > 0.5$. A second filter was applied to remove all SNPs that significantly deviated from the expected neutrality. For this, we have used a Bayesian Fst-outliers based method that identifies loci, which the $F_{ST}$ significantly depart from the average ($F_{ST}$-outlier) (BayeScan v.2.1; Foll and Gaggiotti, 2008). After removal of all significantly linked SNPs, the dataset was phased using Beagle v3.3.2 (Browning and Browning, 2007).

The population structure and the pairwise genetic relationship between individuals from different populations were investigated using a principal component analysis (PCA) implemented in the ngsTools package (Fumagalli et al., 2014) and the resulting principal components (PCs) were plotted using the R script provided at the package website (available at https://github.com/mfumagalli/ngsPopGen/tree/master/scripts). The method implemented takes into account the genotype uncertainty and uses the output of the analyses performed in ANGSD to identify the polymorphic sites (SNP_pval $1x10^{-6}$), estimate the major and minor alleles (doMajorMinor 1), and infer the minimum allele frequencies (doMaf 2). Finally, we only retained loci with a minor allele frequency of <0.05 (minMaf). The posterior genotype probabilities were calculated with uniform *a priori* (doPost 2). The covariance matrix between individuals was calculated weighting each genotype for its posterior probability (Fumagalli et al., 2014).

To explore the relatedness among the chicken populations, we used the admixture model implemented in NGSadmix (Skotte et al., 2013). This method uses the genotype likelihood, taking into account the uncertainty of the genotype callings typical of the low-sequencing depth methods (Foote et al., 2016). For this analysis, we used the genotypes likelihoods determined in ANGSD and used the same set of filters as in previous analysis

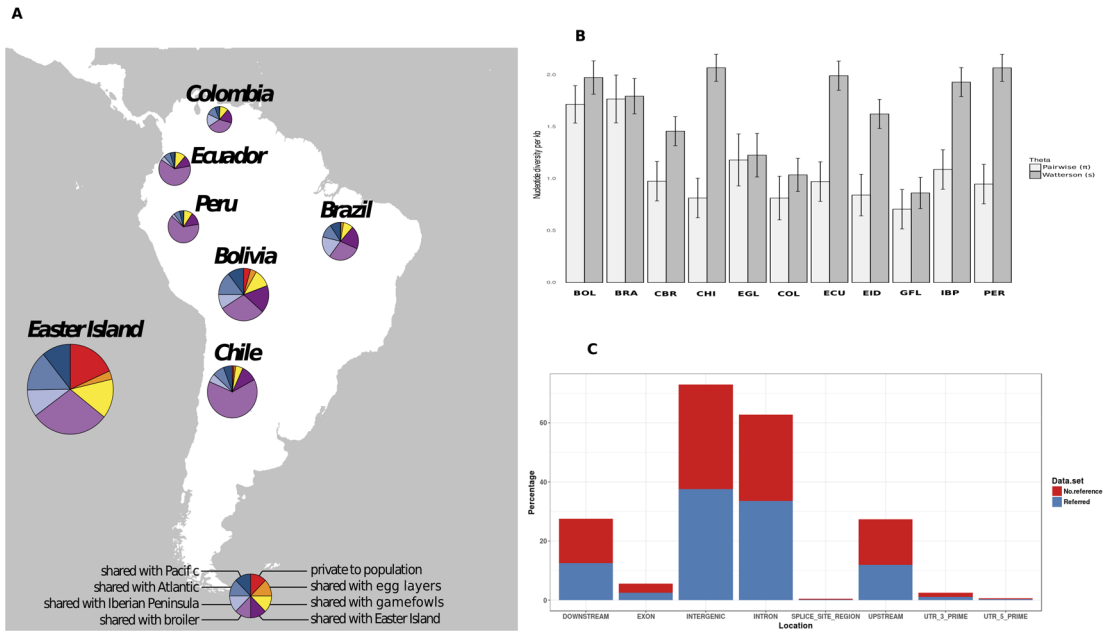


**FIGURE 1 | (A)** Shared polymorphic variants within the South American chicken populations. Pie charts are divided into eight slices as is described in the figure legend at the bottom. The size of the circle is proportional to the number of the variants. **(B)** Bar plots of the mean estimate values of nucleotide diversity with the line corresponding to the mean standard error. Bolivia (BOL), Brazil (BRA), Broiler (CBR), Chile (CHI), Colombia (COL), Ecuador (ECU), Easter Island (EID), Gamefowl (GFL), Iberian Peninsula (IBP), Egg-layer (EGL), Peru (PER). **(C)** Variant effect location, colored areas are proportional to the percentage of previously reported (blue) and first time reported (red) variants.

to avoid bias caused by outliers or linked loci. Several runs were done varying the number of K populations from 3 to 5; to extend this analysis, we have constructed a pie plot chart calculating the average contribution of all potential sources.

## The Origins of the South American Chicken Populations

Hypothetical ancestral admixture events among local South American chicken populations and the four possible population sources (Iberian Peninsula, egg layers, broiler, gamefowl) were assessed using TreeMix (Pickrell and Pritchard, 2012), which calculates a maximum likelihood population tree based on the allele frequencies. This method assigns an edge as a branch of the tree if it contributes with the majority of alleles to the descendant population; otherwise it is a migration edge. This process is performed in a stepwise likelihood mode to find the tree with the best fit for each admixture event (Pickrell and Pritchard, 2012). Here we used 117,962 autosomal phased SNPs, and the SNP dataset obtained from the genome resequencing of several red jungle fowls (Ulfah et al., 2016) as the outgroup.

The TreeMix results were also compared to those obtained using 3 Population Test (AdmixTools package; Patterson et al., 2012), which allows determining whether a population has inherited a mixture of ancestries (Reich et al., 2009). This method is similar to the *f3* (A, B, C), and when significantly negative values of the *f3* statistic are obtained, it implies that population A is admixed. Finally, ROLLOFF software (Patterson et al., 2012) was used to estimate the time of the admixture event. This method used the decay of the linkage admixture disequilibrium to approximate the time of admixture (Moorjani et al., 2011). In our case, the populations from the Iberian Peninsula and gamefowl were used as potential source populations and the South American populations as the admixed populations. The TreeMix results were used to select source populations to be tested in the 3 Population Test. As before, we divided the South American populations into two groups (Atlantic and Pacific).

# RESULTS

## Genetic Diversity

Around 91% of our set of 122,801 nuclear SNPs matched with others already reported at dbSNP NCBI database. The majority of the identified variants were located in intergenic or intronic regions (**Figure 1**), from which approximately 60% were located across the nine macro chromosomes. On average, we roughly observed one SNP for every 8,900 bases (0.122 SNPs per kb).

Regarding the South American continental populations, the lowest number of private variants was observed in the Chilean continental populations, while the highest value was obtained in the Bolivian population. When grouping populations according to their geographic locations in South America (Atlantic and Pacific), all the populations showed a higher number of variants shared with the Pacific group, ranging from 108 in Peru up to 750 in Chile. In the Pacific group, the lowest number of private variants was found in Peru (19) and was highest in the Bolivian population (105). The Atlantic façade populations had a higher number of unique variants

compared with the Pacific, with the maximum found in Brazil (113). The number of variants shared between the South American chicken and the egg layer was lower (between 1 and 46) than the number of variants shared with the broilers (between 17 and 124), the Iberian Peninsula (between 29 and 169), and the gamefowl (between 32 and 130). Individually, the Easter Island population displayed the highest values in terms of private and shared variants. A deeper analysis showed that 643 SNPs were exclusively found in the Easter Island population; 106 were shared only with egg layers, 367 only with broilers, 487 shared with gamefowl, 504 shared with the Iberian Peninsula, and 1,024 and 345 shared with the Pacific and the Atlantic South American groups, respectively (**Figure 1A**). The population diversity theta parameters ($\theta_S$ and $\theta\pi$) estimated per 1,000 bp window attained the lowest values ($\theta_S$ and $\theta\pi$) in the gamefowl population, and the Chile local chicken population showed the highest values for $\theta_S$ and the Brazilian and Bolivian population the highest values for $\theta\pi$ (**Figure 1B**).

## Population Structure and Genetic Relationships

Regarding the population structure and genetic relationships, the most remarkable finding revealed in the PCA plot (**Figure 2**) was the separation between the gamefowl and all the other South American chicken obtained in PC1, whereas PC2 separates Easter Island individuals from all the others. Another separation, although less evident, was the formation of two groups of populations, one containing all countries located in the SA Pacific façade (Ecuador, Peru, and Chile) and the other constituted by Brazil, Colombia, Bolivia, and Iberian Peninsula chicken. We have noticed a slightly higher tendency of the commercial breeds and Iberian population to cluster closer to the South America Atlantic group, whereas the Pacific group is genetically closer to the Easter Island than it is from the Iberian population.

Regarding the pairwise differentiation between the all analyzed populations (**Figure S1**), the gamefowl was the most differentiated population, with $F_{ST}$ values ranging from 11% (Colombia) to 28% (egg layer). All the remaining populations showed lower differentiation levels ranging between 1% between Brazil and Bolivia and 17% between egg layer and Easter Island populations. A one-way ANOVA and Tukey's *post hoc* analysis of the weighted $F_{ST}$ estimates (50 kb sliding windows) showed that the differentiation between South America and the hypothesized population sources (Iberian Peninsula, broiler, egg-layer, gamefowl) is highly significant ($P < 0.001$). When ranking the potential source populations

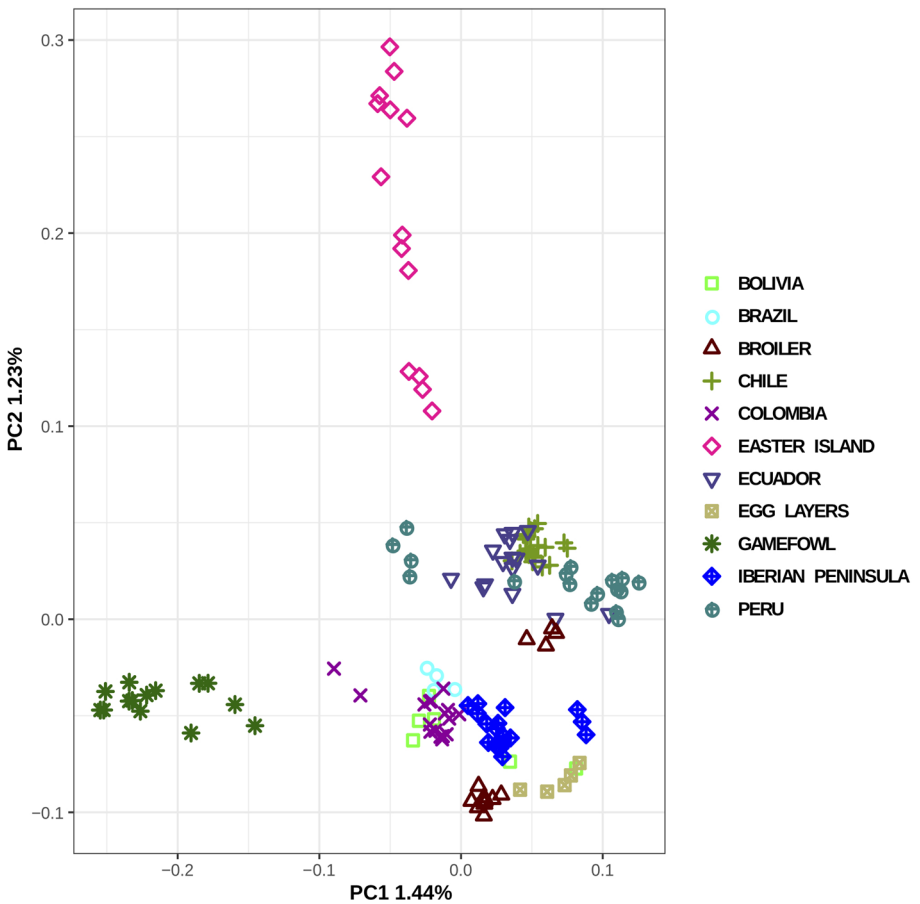


**FIGURE 2 |** Principal component analysis of the local South American populations and putative genetic material sources.

according to their degree of differentiation from the SA, the Iberian Peninsula showed the lowest differentiation ($F_{ST}$ = 0.014), followed by the egg layer ($F_{ST}$ = 0.039) and the broiler ($F_{ST}$ = 0.056), and the gamefowl displayed the highest value ($F_{ST}$ = 0.1) (**Figure S1**).

The Bayesian clustering analysis performed with NGSadmix was consistent with the PCA results. The relatively closely related group formed by all South American populations depicted by the PCA is also confirmed by plotting the admixture analysis results (**Figure 3**). Here, we observe a certain degree of admixture between all the South American chickens and the influence of the commercial egg layers and broiler lineages as well as the gamefowl in the contemporary South American chicken. The Easter Island population displays a different admixture pattern in which a specific (non-shared) genetic background is very pronounced. Moreover, the admixture plot shows that in the Easter Island population, the most frequent genetic component is represented, although at a very small frequency, at the continental South American populations.

## The Origins of the South American Populations

As the previous analysis pointed to a large influence of commercial breeds in the South American chicken, we have quantified this influence using TreeMix analyses. The obtained phylogenetic tree reflects the divergence patterns among the different chicken populations (**Figure 3**) and depicts the large influence of the egg layer in the South American chicken (**Figures S2, S3**).

The $f_3$-statistics analysis, through *3-population test*, to confirm the introgression events identified using the TreeMix method, returned significant values for the combinations $f_3$ (Pacific; egg layer, gamefowl) and $f_3$ (Atlantic; egg layers, gamefowl). For the Pacific–egg layer–gamefowl combination, the calculated values were $f_{3\ =}$ -0.0017, Z = -12.44 and for the Atlantic–egg layers–gamefowl combination, calculated values were $f_{3\ =}$ -0.0017, Z = -16.599 (**Table S1**).

Finally, to quantify the contribution of each potential source, we have calculated the average values based on the NGSadmix results (**Figure 3C**). With the exception of Chile, the local

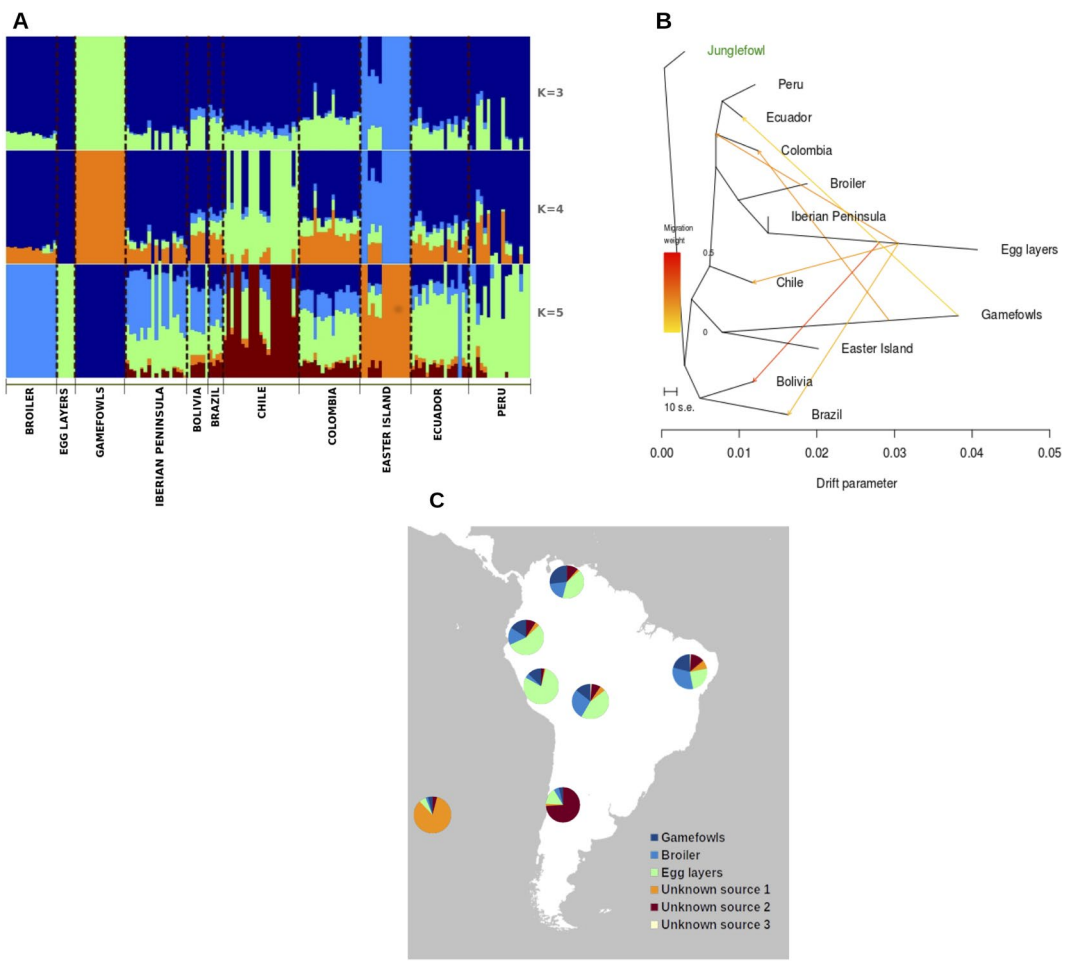


**FIGURE 3 | (A)** Individual ancestry proportion of each of the South American chicken samples and putative genetic material sources conditional on the number of genetic clusters (k = 3-5). **(B)** TreeMix analyses of the genetic relationships between South American chicken and potential source populations. **(C)** Putative genetic material sources conditional on the number of genetic clusters.

South American chicken populations were highly influenced by commercial chicken breeds, in which the egg-layers average admixture level ranges between 44% and 79%, while the broiler had a lower influence with an average admixture level ranging between 16% and 32%, and the gamefowl contribution ranges from 4% to 27% with the lowest in Chile and the highest to Colombia. Chile and Easter Island populations show different admixture patterns relative to the other populations with a high percentage of contributions from unknown sources. Interestingly, the results from the roll-off analyzes that are dated to be the most influential migratory events from around 70 ± 10 generations ago, which represents between 70 and 35 years considering a 1-year or 2-year generation interval, respectively.

# DISCUSSION

## Genetic Diversity

The commercial and relatively accessible high-density SNP array for the chicken became the most common tool used in genomic studies recently. However, the use of this pre-ascertained SNP panel distorts population genetic inferences on local livestock populations, as the sample sizes and the highly selected populations in which SNPs were discovered pose significant biases (Albrechtsen et al., 2010; Lachance and Tishkoff, 2013). Here, we used reduced representation library sequencing, in this case, RADseq, to interrogate a medium-high number of SNPs (122,801). The comparison of this set with those SNPs identified in the NCBI dbSNP database revealed that 91% of our SNPs match with others previously identified and 97% of them are located in intergenic or intronic regions, showing great potential to be used in genetic diversity studies.

The summary statistics of genetic variation using two theta estimators ($\theta\pi$ and $\theta s$) showed similar diversity per population (**Figure 1B**). The gamefowl proved to be an exception to this as they showed the lowest values and can be explained as the result of the inbreeding practices used to swiftly fix desired traits (García, 1997). The very similar values obtained for the two parameters ($\theta s$, $\theta\pi$) in Brazil and Bolivia populations are better explained by the sample size effect (Korneliussen et al., 2013), as the sampling for both populations was substantially smaller than for the other South American populations. On the other hand, the different values displayed between the two theta parameters, with the $\theta s$ showing higher values than $\theta\pi$, at the remaining populations (e.g., CHI, PER, ECU, and IBP), can be explained by differences in the proportion of alleles segregating at intermediate frequencies. It is known that the $\theta\pi$ algorithm ascribes more weight to alleles segregating at intermediate frequencies, while $\theta s$ weights all categories equal (Korneliussen et al., 2013), and thus populations showing a lower number of alleles with intermediate frequencies will result in smaller $\theta\pi$ values.

The patterns of the genetic variants shared among the different populations also provide insights about the continental South American chicken population diversity. Interestingly, the Easter Island population is the one displaying the highest number of unique variants (643), and this can be interpreted as the result of its different demographic history and/or different population origins. The high number of unique alleles could be explained by the different origins of the chicken introduced on this island across time (Luzuriaga-Neira et al., 2017). Alternatively, the high number of shared variants between this population and the other continental South American chickens can be explained by a source-sink metapopulation process (e.g., Gaggiotti, 1996). The occurrence of this phenomenon can simultaneously explain the occurrence of a high number of private variants (sink) and shared variants (source) as the result of different migration events from SA continent that have arrived at this island since at least 1772 (Wilhelm, 1957).

## Population Structure and Genetic Relationships

The PCA plot (**Figure 2**) constructed with all individuals shows that the individuals belonging to the gamefowl and Easter Island populations are relatively well separated from the remaining populations. Curiously, despite the low differentiation between the remaining continental South American populations, the PCA divides them into two groups, which might be related with whether its geographic location is on the Atlantic *façade* (Brazil, Columbia) or the Pacific *façade* (Peru, Chile, Ecuador).

The large differentiation indicated by $F_{ST}$ estimates between the gamefowl and all the other South American populations (Table S2) is not very surprising. The different breeding objectives (i.e., behavior) and the observed low levels of diversity are the two most probable causes of this high differentiation regarding the other South American populations. Indeed, the admixture analysis shows the absence of influence from the other tested breeds in the gamefowl (**Figure 3A**) but shows some influence of this population in the other populations. This might indicate that the different breeding goal of this population, regarding the rest, has prevented its crossing with the commercial chicken breeds, particularly with the commercial egg-layer breed, as is evident in the other South American populations.

The Easter Island population is a very interesting example, as despite being the most divergent from the other populations, it is also the one in which its individuals are relatively more dispersed. The PCA grouping of the individuals (**Figure 2**) is a relatively good method to detect the coancestry relationship among individuals from the same population. It is expected that two individuals closely related would be closer to each other, but the Easter Island population has individuals that are considerably more distant from the others of their own population than relatively other individuals from other populations (e.g., Peru). In fact, this pattern is usually associated with different migration events (Fumagalli et al., 2013; Schraiber and Akey, 2015), and in this case, may indicate the influence of the chicken populations from the SA Pacific fringe in the Easter Island population. The higher differentiation is displayed by both Easter Island and the gamefowl populations, whereas the small differentiation amidst South American chicken populations and between these and the commercial breeds suggests differential gene-flow rates as the main driver of the extant South American chicken population structure.

The post-Columbian human migration events and the subsequent spread of people from the coastal areas to the

interior become particularly massive at the end of the nineteenth century and might have led to multiple introductions of chicken from different populations. The quantification of the admixture proportion for each of the studied populations and a large number of migration edges needed to add (13) to explain most of the variance (99.8%) depicted by the phylogram (**Figure S2**) demonstrates that those populations have had a constant flux of foreign genes.

### The Origins of the South American Chicken Populations

It has been hypothesized that European and Asian chickens were introduced in SA after 1500 (Storey et al., 2011); nevertheless, the modern introductions have been less described. However, we found that a single source population (Iberian Peninsula) could not explain the diversity displayed by the South American chicken suggesting a different demographic history for the South American chicken populations, opening the possibility of a multiple origin scenario. The poultry industrialization that started after World War II resulted in the globalization of massive industrial production and dispersal, leading to extensive crossbreeding between individuals from few highly selected and cosmopolitan chicken varieties (egg-layers, broilers) with local varieties, which have taken place in SA. Remarkably, the *roll off* admixture analysis detected signs of a strong introgression in SA population dating between 35 and 70 years ago, which is concordant with the worldwide expansion of poultry industry based on highly productive chicken lineages. If this is correct, then the current SA local chicken accumulates the legacy of the older chicken introduced with those modern highly selected varieties.

In Ecuador, Peru and, Colombia, cock-fighting is a popular part of their culture and local recreation activities (Finsterbusch, 1990). However, the origin of the SA gamefowl is poorly known, with many anecdotal reports linking their introduction with the arrival of Spanish and Portuguese colonizers who may have brought these birds from their colonies in South and Southeast Asia, where cock-fighting is a very ancient tradition (Lawler, 2014). Here, we could not identify the potential source population, but the TreeMix tree positions it at the same branch with the Easter Island population (**Figure 3B**), which might be indicative of a common origin of these two populations. Although the Easter Island chicken may have their roots linked to the Polynesian people expansion throughout the South Pacific (Wilhelm, 1957; Fitzpatrick and Callaghan, 2009), which have arrived at Easter Island around 1,200 A.D. (Hunt and Lipo, 2006), its genetic proximity with the SA continental gamefowl can be explained by the fact that both populations were not crossed with cosmopolitan breeds and therefore remain closer to the ancestral population that originated them. Moreover, if this is true, then these populations may represent the genomes of the first chicken that were introduced in this part of the world, which have been replaced in other populations by uncontrolled crosses between local and newly selected chicken cosmopolitan populations (broiler and egg-layers) that were developed during the intensification of poultry production. Indeed, the admixture levels obtained in this study point for a replacement of the local genomes of the older local chicken populations that were taken from the Iberian Peninsula to South America five centuries ago.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the BioProject database: https://www.ncbi.nlm.nih.gov/bioproject/573756 (Accession: PRJNA573756).

## ETHICS STATEMENT

Standard techniques were used to collect very small piece of tissue from each animal, by local veterinary trained personnel. The procedure was reviewed and approved by CIBIO-University of Porto Committee of ethics.

## AUTHOR CONTRIBUTIONS

AB-P, AL-N, GV-R, and FC-C conceived the study. AL-N, SO'R and AB-P drafted the manuscript. AL-N, LP-P, and MM participated in the data analysis. AL-N and SO'R did the laboratory work. AL-N, AU-N, GE-S, J A-P, MV and MR-Q did the sampling. AB-P and MRM supervised the study. All the authors read and approved the manuscript.

## FUNDING

This work was supported by funds from the project NORTE-01-0145-FEDER-000007, from the Norte Portugal Regional Operational Program (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). AL-N was supported by a doctoral grant from SENESCYT, LP-P was a recipient of a postdoctoral grant from the Portuguese Science Foundation (FCT) (SFRH/BPD/94518/2013), and AB-P was a recipient of an IF contract from the FCT.

## ACKNOWLEDGEMENTS

The authors wish to thank the logistical support from the Centro de Biotecnologia at the Universidad Nacional de Loja (Ecuador) as well as the South American farmers and the Portuguese local chicken breeds association (AMIBA) for providing access to sample their chickens.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01172/full#supplementary-material

## REFERENCES


Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi: 10.1093/molbev/msq148

Ali, O. A., O'rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., et al. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genet.* 202, 389–400. doi: 10.1534/genetics.115.183665

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Caminsky, N. G., Mucaki, E. J., Perri, A. M., Lu, R., Knoll, J. H. M., and Rogan, P. K. (2016). Prioritizing variants in complete hereditary breast and ovarian cancer genes in patients lacking known BRCA mutations. *Hum. Mutat.* 37, 640–652. doi: 10.1002/humu.22972

Carter, G. F. (1971). Precolumbian chickens in America in *Man across the sea: problems of precolumbian contacts*. Eds. Riley, J. C., Kelley, P. C., W, R. L., and Rands, C. L. (Austin: University of Texas Press), 178–218.

Crawford, R. D. (1990). "Origin and history of poultry species," in *poultry breeding and genetics,* ed. *R.D. Crawford. (Amsterdam - Oxford - New York - Tokyo: Elsevier)* 1, 42.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., De Los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genet.* 2, 347–365. doi: 10.1534/genetics.112.147983

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

Feichtinger, J., Hernández, I., Fischer, C., Hanscho, M., Auer, N., Hackl, M., et al. (2016). Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. *Biotechnol. Bioeng.* 113, 2241–2253. doi: 10.1002/bit.25990

Finsterbusch, C. A. (1990). *Cock fighting all over the world*. Alton, England: Nimrod Press.

Fitzpatrick, S. M., and Callaghan, R. (2009). Examining dispersal mechanisms for the translocation of chicken (*Gallus gallus*) from Polynesia to South America. *J. Archaeol. Sci.* 36, 214–223. doi: 10.1016/j.jas.2008.09.002

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genet.* 180, 977–993. doi: 10.1534/genetics.108.092221

Foote, A. D., Vijay, N., Avila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., et al. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.* 7, 11693. doi: 10.1038/ncomms11693

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sanchez, E., Albrechtsen, A., et al. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genet.* 195, 979–992. doi: 10.1534/genetics.113.154740

Fumagalli, M., Vieira, F. G., Linderoth, T., and Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinf.* 30, 1486–1487. doi: 10.1093/bioinformatics/btu041

Fumihito, A., Miyake, T., Sumi, S.-I., Takada, M., Ohno, S., and Kondo, N. (1994). One subspecies of the red junglefowl (*Gallus gallus gallus*) suffices as the matriarchic ancestor of all domestic breeds. *Proc. Natl. Acad. Sci. U.S.A.* 91, 12505–12509. doi: 10.1073/pnas.91.26.12505

Fumihito, A., Miyake, T., Takada, M., Shingu, R., Endo, T., Gojobori, T., et al. (1996). Monophyletic origin and unique dispersal patterns of domestic fowls. *Proc. Natl. Acad. Sci. U. S. A.* 93, 6792–6795. doi: 10.1073/pnas.93.13.6792

Gaggiotti, O. E. (1996). Population genetic models of source sink metapopulations. *Theor. Popul Biol.* 50, 178–208. doi: 10.1006/tpbi.1996.0028

García, A. J. (1997). Relationship between the degree of kinship and reproductive ability in game fowl breeders. *Rev. Cubana Ciencia Avícola* 21, 109–112.

Gongora, J., Rawlence, N. J., Mobegi, V. A., Jianlin, H., Alcalde, J. A., Matus, J. T., et al. (2008). Reply to storey et al.: more DNA and dating studies needed for ancient El Arenal-1 chickens. *Proc. Natl. Acad. Sci. U.S.A.* 105, E100. doi: 10.1073/pnas.0809681105

Hunt, T. L., and Lipo, C. P. (2006). Late colonization of Easter Island. *Sci.* 311, 1603–1606. doi: 10.1126/science.1121879

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nat.* 432, 695–716. doi: 10.1038/nature03154

Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinf.* 14, 289. doi: 10.1186/1471-2105-14-289

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinf.* 15, 356. doi: 10.1186/s12859-014-0356-4

Kristensen, T. N., Hoffmann, A. A., Pertoldi, C., and Stronen, A. V. (2015). What can livestock breeders learn from conservation genetics and vice versa? *Front. In Genet.* 6, 38. doi: 10.3389/fgene.2015.00038

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, 780–786. doi: 10.1002/bies.201300014

Lawler, A. (2014).*Why did the chicken cross the world: the epic saga of the birds that powers civilization*. (New York: Atria Books).

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinf.* 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. *Bioinf.* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

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

Liu, Y. P., Wu, G. S., Yao, Y. G., Miao, Y. W., Luikart, G., Baig, M., et al. (2006). Multiple maternal origins of chickens: out of the Asian jungles. *Mol. Phylogenet Evol.* 38, 12–19. doi: 10.1016/j.ympev.2005.09.014

Luzuriaga-Neira, A., Villacis-Rivas, G., Cueva-Castillo, F., Escudero-Sanchez, G., Ulloa-Nunez, A., Rubilar-Quezada, M., et al. (2017). On the origins and genetic diversity of South American chickens: one step closer. *Anim. Genet.* 48, 353–357. doi: 10.1111/age.12537

Miao, Y. W., Peng, M. S., Wu, G. S., Ouyang, Y. N., Yang, Z. Y., Yu, N., et al. (2013). Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity (Edinb)* 110, 277–282. doi: 10.1038/hdy.2012.83

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., et al. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PloS Genet.* 7, e1001373. doi: 10.1371/journal.pgen.1001373

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

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PloS Genet.* 8, e1002967. doi: 10.1371/journal.pgen.1002967

Purcell, S., and Chang, C. *PLINK v1.9*, 2015, 1.9 ed. (www.cog-genomics.org/plink/1.9/).

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing.

Ramírez-Aliaga, J.-M. (2010). The polynesian – mapuche connection: soft and hard evidence and new ideas. *Rapa Nui J.* 24, 29–33.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nat.* 461, 489–494. doi: 10.1038/nature08365

Schraiber, J. G., and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16, 727–740. doi: 10.1038/nrg4005

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genet.* 195, 693–702. doi: 10.1534/genetics.113.154138

Storey, A. A., Ramírez, J. M., Quiroz, D., Burley, D. V., Addison, D. J., Walter, R., et al. (2007). Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10335–10339. doi: 10.1073/pnas.0703993104

Storey, A. A., Quiroz, D., Beavan, N., and Matisoo-Smith, E. (2011). Pre-Columbian chickens of the Americas: a critical review of the hypotheses and evidence for their origins. *Rapa Nui J* 25, 5–19.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genet.* 123, 585–595.

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. *Phil. Trans. R. Soc B* 365, 2853–2867. doi: 10.1098/rstb.2010.0134

Ulfah, M., Kawahara-Miki, R., Farajalllah, A., Muladno, M., Dorshorst, B., Martin, A., et al. (2016). Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatra and Kedu Hitam. *BMC Genomics* 17, 320. doi: 10.1186/s12864-016-2652-z

Underhill, A. P. (1997). Current issues in chinese neolithic archaeology. *J. World Prehistory* 11, 103–160. doi: 10.1007/BF02221203

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul Biol.* 7, 256–276. doi: 10.1016/0040-5809(75)90020-9

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

West, B., and Zhou, B. X. (1988). Did chickens go North? new evidence for domestication. *J. Archaeol. Sci.* 15, 515–533. doi: 10.1016/0305-4403(88)90080-5

Wilhelm, O. G. (1957). Las gallinas de la Isla de Pascua in *Boletín de la Sociedad de Biología de Concepcion*, vol. 32, 133–137.

Xiang, H., Gao, J., Yu, B., Zhou, H., Cai, D., Zhang, Y., et al. (2014). Early holocene chicken domestication in northern China. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17564–17569. doi: 10.1073/pnas.1411882111

Zhang, Q., Gou, W., Wang, X., Zhang, Y., Ma, J., Zhang, H., et al. (2016). Genome resequencing identifies unique adaptations of tibetan chickens to hypoxia and high-dose ultraviolet radiation in high-altitude environments. *Genome Biol. Evol.* 8, 765–776. doi: 10.1093/gbe/evw032

**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

REVIEW
published: 17 January 2020
doi: 10.3389/fgene.2019.01304

# On the Extent of Linkage Disequilibrium in the Genome of Farm Animals

*Saber Qanbari*[1,2]*

[1] Leibniz Institute for Farm Animal Biology (FBN), Institute of Genetics and Biometry, Dummerstorf, Germany, [2] Animal Breeding and Genetics Group, Department of Animal Sciences, Center for Integrated Breeding Research, University of Göttingen, Göttingen, Germany

Given the importance of linkage disequilibrium (LD) in gene mapping and evolutionary inferences, I characterize in this review the pattern of LD and discuss the influence of human intervention during domestication, breed establishment, and subsequent genetic improvement on shaping the genome of livestock species. To this end, I summarize data on the profile of LD based on array genotypes vs. sequencing data in cattle and chicken, two major livestock species, and compare to the human case. This comparison provides insights into the real dimension of the pairwise allelic correlation and haplo-block structuring. The dependency of LD on allelic frequency is pictured and a recently introduced metric for moderating it is outlined. In the context of the contact farm animals had with human, the impact of genetic forces including admixture, mutation, recombination rate, selection, and effective population size on LD is discussed. The review further highlights the interplay of LD with runs of homozygosity and concludes with the operational implications of the widely used association and selection mapping studies in relation to LD.

**Keywords: association mapping, selection mapping, runs of homozygosity, allele frequency spectrum (AFS), haplotype block**

OPEN ACCESS

***Edited by:***
*Gábor Mészáros, University of Natural Resources and Life Sciences Vienna, Austria*

***Reviewed by:***
*Farai Catherine Muchadeyi, Agricultural Research Council of South Africa (ARC-SA), South Africa*
*Maja Ferenčaković, University of Zagreb, Croatia*

***Correspondence:**
*Saber Qanbari*
*qanbari@fbn-dummerstorf.de*

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

***Received:*** *31 July 2019*
***Accepted:*** *26 November 2019*
***Published:*** *17 January 2020*

***Citation:***
*Qanbari S (2020) On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. Front. Genet. 10:1304. doi: 10.3389/fgene.2019.01304*

## INTRODUCTION

Linkage disequilibrium (LD) is the non-random assortment of alleles at different loci. The terms linkage and LD are often confused. As highlighted by Slatkin (2008), LD is one of those unfortunate terms that do not reveal its meaning. Indeed, LD means simply a correlation between alleles, and detecting LD does not ensure either linkage or a lack of equilibrium. This stems from the fact that mechanisms other than just physical proximity on a chromosome (linkage) such as mutation, genetic drift, and epistatic combinations might also cause (gametic phase) disequilibrium between unlinked markers. For example, admixing genetically distinct populations creates association between two loci with different allele frequencies even if they are unlinked. LD can also arise due to population stratification and cryptic relationships within a population that results in correlated allelic frequencies (reviewed in Hellwege et al., 2017).

The pattern of LD is a powerful indicator of the genetic forces shaping a population. For example, knowledge of LD helps inferring a population's effective size (*Ne*) and past demography. Populations with smaller *Ne* experience more genetic drift than larger populations. This genetic drift causes LD

between alleles at independently-segregating loci, at a rate inversely proportional to *Ne* (Waples et al., 2016). This way, an estimate of contemporary *Ne* can be concluded from LD information (Sved, 1971; Hill, 1981). On the contrary, past *Ne* is a function of LD between physically-linked loci, given that the inter-loci recombination fractions are available (Sved, 1971). Accordingly, the closely-linked loci indicate population sizes over historical past, while loosely-linked loci signify *Ne* in the immediate past (Hill, 1981, Hayes et al., 2003). Unlike the non-model species, these methods can be applied in the populations of farm animals for which the high resolution genetic maps are becoming available (Tortereau et al., 2012; Ma et al., 2015a; Petit et al., 2017).

LD between linked markers also determines the power and precision of association mapping studies,directly influencing our ability to localize genes and or loci responsible for economic traits in agriculture or inherited diseases in human (reviewed in Goddard and Hayes, 2009). Given the economic impact of domestic animals, understanding the dimension of LD enables planning and performing successful genomic breeding programs, when working towards global food security. This review aims to outline the definition of LD, summarize data on patterns of LD in the genome of farm animals, and discuss the various properties and implications that LD causes for gene mapping and evolutionary studies of livestock species.

## A HISTORICAL GLANCE

The concept of LD was first introduced in Jennings (1917), and its quantification (*D*) was developed by Lewontin and Kojima (1960). LD became a hot topic in the last two decades once the usefulness of LD for gene mapping became evident and genotyping of large numbers of linked single-nucleotide polymorphism (SNP) became feasible through high-throughput technologies.

The simple formulation of the commonly used LD measure *D* is the differencebetween the observed and the expected gametic haplotype frequencies comprising two loci A and Bunder linkage equilibrium ($D=P_{AB}-P_AP_B=P_{AB}P_{ab}-P_{Ab}P_{aB}$). Besides *D*, several measures of LD (for example, $D'$, $\lambda$, $\delta$, $r^2$, $\chi^2$ $\rho^2$, among others) have been suggested (Lewontin, 1964; Bengtsson and Thomson, 1981; Hill and Weir, 1994; Terwilliger, 1995; Zhao et al., 2005; Gianola et al., 2013). The merits, comparison, and methodologies of these metrics with the utilization of biallelic or multi-allelic loci have been extensively described in the literature (e.g., Jorde, 2000; Pritchard and Przeworski, 2001; Mueller, 2004; Sved, 2009). Choosing the appropriate LD measure depends on the objective of the study, and one may perform better than another in particular situations. The two widely used measures of LD are $r^2$ and $D'$. $r^2$ is indicative of the correlation that a marker might have with the gene of interest and is often preferred for association studies.

## LD-BASED MAPPING OF GENES

Identifying the genetics underlying phenotypic variation is the ultimate goal of most mapping studies. In general, there are two different, but to some extent, complementary methodologies to localize genes controlling traits. Both methodologies, outlined below, benefit from the properties of LD to accomplish the mapping task.

*Association mapping:* is the most common approach of mapping quantitative trait loci (QTLs) that takes advantage of the historic LD to connect phenotypes to genotypes. This approach detects inherited markers in the vicinity of the genetic causatives or loci controlling the complex quantitative traits. It is often performed by scanning the entire genome for significant associations between a panel of SNPs and a particular phenotype (e.g., Hayes et al., 2010). Subsequent analyses will then be required to verify the realized association independently in order to confirm that it either directly controls the trait of interest, or is linked to (in LD with) a QTL that contributes to the trait of interest.

Association analysis is based on the principle that an unbeknownst causative variant is located on a haplotype, and a marker allele in LD with the causative variant should signify (by proxy) an association with the trait of interest. Given the fact that SNPs are in LD with one another, if a common SNP affects a trait, one can probably genotype a SNP in LD with it (a "marker" SNP) and that marker will be correlated with the trait of interest.

Quantifying the extent of LD is the essential first step to determine the number of markers required to cover the entire genome in an association study with succinct power and precision. Theoretically, extensive LD reduces the number of markers required to localize an association between marker and trait but in lower resolution. In contrast, when LD promptly decays within a short distance, many markers are needed to map a gene of interest.

Although the LD-based association analysis is a powerful tool routinely applied for gene mapping, it has not been very successful for targeting genes of complex traits, especially where the causative variants are low in frequency. This is due to the fact that commercial genotyping arrays largely under-represent infrequent alleles (reviewed in Lee et al., 2014). For a detailed discussion, refer to the article by Goddard and Hayes (2009) reviewing the pros and cons of association analysis in farm animals. Here I stress the importance of LD in exploring the genetic variability underlying phenotype-genotype relationship. It is noteworthy that with the advancement of bioinformatics tools and high throughput sequencing technologies that provides the full profile of an individual's genetic variation, it is now possible to test for the effects of every single DNA polymorphism on phenotypic variation, without requiring LD information. However, given the presence of confounding factors such as cryptic correlations in interpreting the GWAS results, LD remains useful as evidence for validation of a detected association (Bulik-Sullivan et al., 2015).

*Mapping selection:* Selection generates LD between distant loci through a "hitch-hiking" effect (Smith and Haigh, 1974), which happens when a haplotype carrying the favored allele rises in frequency so fast and drags neighboring loci to higher frequencies. Scanning the genome for long unbroken haplotypes accompanied by extensive LD can reveal past

selection responding to an adaptive quality (e.g., Sabeti et al., 2002). Domestic species have been intensively selected during the recent past through domestication, breed establishment and genetic improvement and as such, have achieved tremendous phenotypic changes. Consequently, genomic regions controlling traits of economic importance are expected to exhibit footprints of selective breeding (reviewed in Qanbari and Simianer, 2014a).

## DEPENDENCY ON ALLELIC FREQUENCY

The widely used measure of LD in animal breeding and genome-wide association mapping is $r^2$. This metric has an allele frequency-dependent character (see **Figure 1**), as is quoted in Lewontin (1988) "*there are generally no gene frequency independent measures of association between loci*". The dependence of $r^2$ on allele frequencies affects the outcomes and interpretations of population genetics studies in several ways. For example, there are population characteristics that are related to the estimated value of LD, such as effective population size and pattern of recombination landscapes. This implies that the estimates of effective size or recombination maps developed based on expected values of $r^2$ are frequency-dependent as well (e.g., Ober et al., 2013). Furthermore, in gene mapping studies, power to detect a causative variant using SNP markers is a function of $r^2$ between the causative variant and the marker. Thus, if a SNP marker and a causative variant have different minor allele frequencies, then the power to detect an effect at the marker can be small since high values of $r^2$ are not realized. This property of $r^2$ becomes especially more significant in human models, where the most disease-causing variants are rare and genome-wide association studies should be adapted to target these variants.

Even if a frequency independent measure of LD may not exist, it would be desirable to develop one which is less affected by frequencies than $r^2$. In a recent study (Gianola et al., 2013), we developed a new estimator of LD parameter ($\rho^2$) based on a metric proposed by Plackett (1965) that is a tetra-choric correlation (Pearson, 1901). Plackett (1965) introduced bivariate distributions indexed by a single parameter $\psi$ that, in the case of the 2 x 2 table, takes the form $\psi = \frac{P_{AA}P_{BB}}{P_{AB}P_{BA}}$. The relationship between the tetra-choric correlation and $\psi$ is given by

$$\rho = -cos\left[\pi \frac{\sqrt{\psi}}{1+\sqrt{\psi}}\right],$$

where, $\rho$ is easy to compute and much less dependent on allele frequency than $r^2$ (see **Figure 2**).

We argue that $\rho^2$ is a useful metric and potent to the further research and developments for applications in population and quantitative genetics. For instance, $\rho^2$ can facilitate comparison of levels of LD among populations that are subjected to different allelic frequencies, whereas such comparisons are distorted by the frequency-dependent nature of $r^2$. Likewise, in the quantitative genetics context, the power analyses are

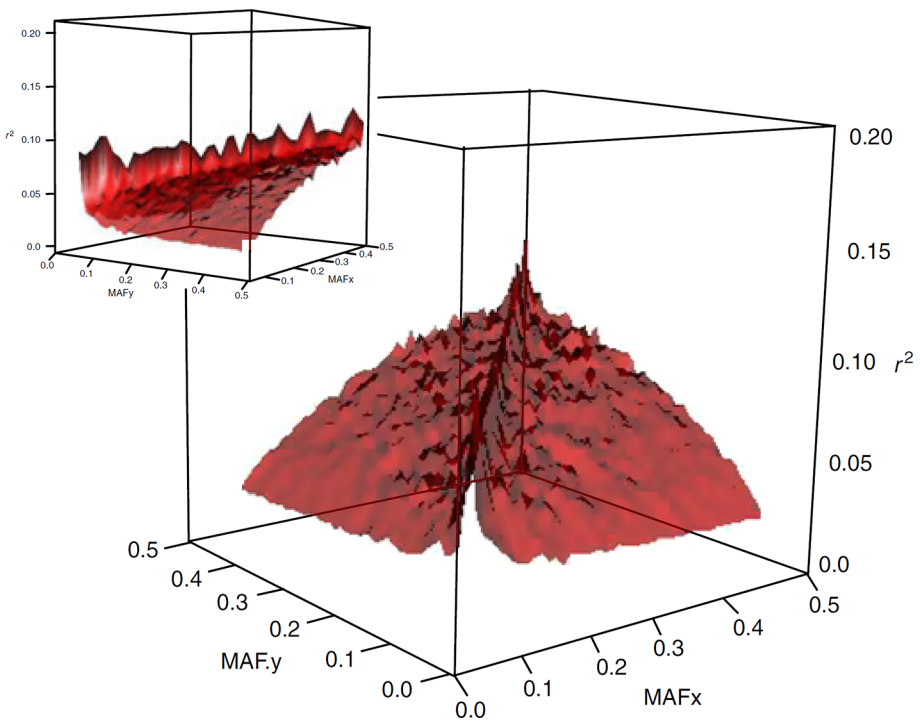


**FIGURE 1 |** Surface plot of the dependency of LD on allelic frequency of SNP pairs. The means of $r^2$ are plotted for 45 bins of 0.01 allele frequency each (from Qanbari et al., 2010a).

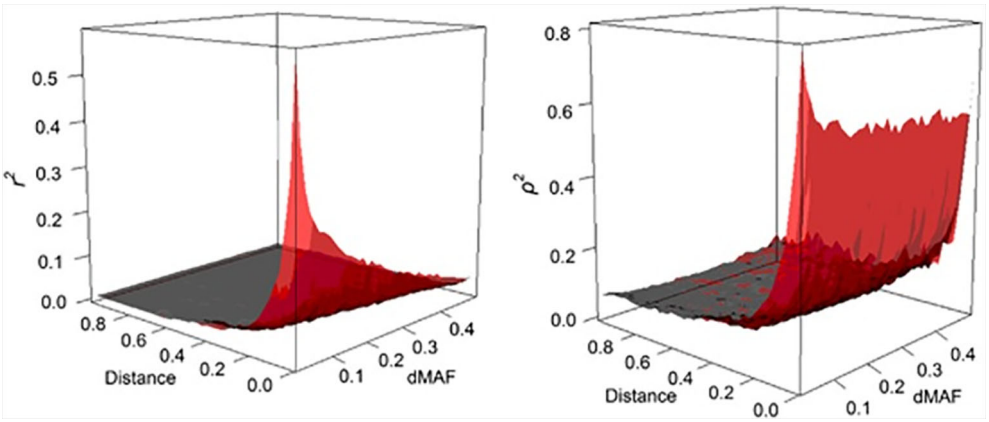


**FIGURE 2 |** The behavior of LD as a function of inter-marker distance (Mb) and MAF interval (dMAF). The estimates of $r^2$ (left panel) and $\rho^2$ (right panel) are depicted as surface plots for SNP loci on chromosome 3 of the Italian Tuscan population in HapMap III (from Gianola et al., 2013).

formulated based on $r^2$ in association studies or genomic selection programs. For example, the sample size in indirect association studies must be increased by roughly $1/r^2$ for detecting the causal mutation directly (Kruglyak, 1999; Pritchard and Przeworski, 2001). Similarly it is suggested that the required level of LD ($r^2$) for genomic selection to achieve an accuracy of 0.85 for genomic breeding values has to be 0.2 (Meuwissen et al., 2001). Perhaps, similar relationships can also be developed for $\rho^2$, which is a subject for future research.

## THE EXTENT OF LD: GENOTYPE VS. SEQUENCE DATA

The strength of LD is of crucial importance for the genome-based analysis of evolutionary history, fine-tuning of applications like association mapping, genomic selection and selection mapping. Most of the previous studies on LD in farm animals have used panels of ascertained genotypes of different densities available by SNP genotyping arrays. The availability of population sequencing for livestock species nowadays has provided the opportunity to figure patterns of LD in unprecedented resolution. With advances in high-throughput sequencing technologies, read lengths are becoming longer, an ideal situation for estimating LD, as longer reads allow direct phasing of double heterozygotes (Maruki and Lynch, 2014).

The extent to which LD decays in the genome of farm animals has been extensively studied on the basis of genotypes from SNP arrays (Porto-Neto et al., 2014; Khanyile et al., 2015; Prieur et al., 2017; Marchiori et al., 2019; Mokhber et al., 2019; Muñoz et al., 2019, among others). While genotyping arrays exhibited LD extending at several hundreds of kilobases, a denser catalog of SNPs generated from genome re-sequencing reveals LD decaying at much shorter distances (see **Figure 3**). This is attributed to the SNP profile used to measure LD. As shown in **Figure 4**, the distribution of allele frequency drawn from sequence data is a decreasing function that involves a sizable fraction of infrequent alleles. In contrast, frequency distribution in genotyping arrays is rather an increasing function, as SNPs were mainly ascertained aiming at frequent alleles and coverage of the genome during the establishment of the array (also see Fu et al., 2015 and Makina et al., 2015). Given that LD, as measured by $r^2$ depends on allele frequencies, the difference between the studies is partially due to the biased SNPs selection on the genotyping arrays. Other factors such as the influence of population sub-structuring in the sample composition or sequencing errors may also affect the allelic correlations. However, LD measures in this experiment were

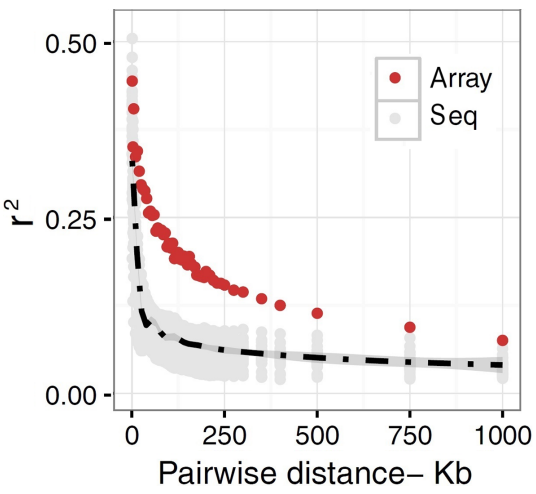


**FIGURE 3 |** A schematic representation of decay of LD in domestic chicken. $r^2$ values are plotted as a function of pair-wise inter-marker distances based on sequence (Seq) versus SNP50K (Array) data in a population of Lohmann brown layer line. The gray dots represent sequence-based $r^2$ plotted for each chromosome separately, whereas LD based on array data was simply averaged genome-wide due to the lack of enough LD estimates in shorter distance bins. The black dashed line is fitted as mean LD in each distance bin across chromosomes. The $r^2$ values representing sequence data are estimated for sub-samples of all pairwise estimates in macrochromosomes, but include all SNP by SNP relationships in microchromosomes.

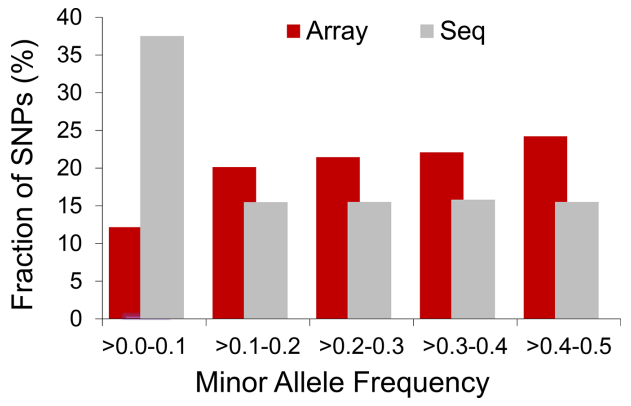


**FIGURE 4 |** Distribution of allelic frequency in domestic chicken. Histogram compares profile of minor allele frequency between 50K array and sequence data in a population of Lohmann brown layer.

drawn from the identical set of samples for both array and sequence resolution and the differences between the two marker sets are too significant to be caused by sequencing errors. For further validation of this observation based on possible scenarios I refer to the experiments described in Qanbari et al. (2014b).

## LD HAPLO-BLOCKS: GENOTYPE VS. SEQUENCE DATA

A haplotype block is a set of closely linked markers on a chromosome with a strong LD between each other that tend to inherit together (Gabriel et al., 2002). The haplo-blocks could have been produced by interplay of several possible mechanisms, including domestication, population subdivision, founding events, selection, and recombination hotspots. These structures, when discovered, were of great practical importance for the gene mapping studies; as such, testing one SNP within each block for significant association with a trait might be sufficient to indicate association with every SNP in that block (Carlson et al., 2004). This could reduce the number of SNPs required to be tested in association studies.

Haplotype blocks have been studied in human and other farm animals. Previous studies in farm animals based on array data have reported haplo-blocks extending to several hundreds of kilobasepairs (e.g., Qanbari et al., 2010a; Qanbari et al., 2010c; Al-Mamun et al., 2015, among others). The assembly of large LD blocks appearing in array-based analyses, however, breaks into series of shorter tracts when LD is assessed by sequence data in the cattle genome (**Figure 5**). Consistent with the reduced LD profile presented in **Figure 4**, resolving large haplo-blocks in sequence resolution is a consequence of shift in allele frequency spectrum towards infrequent alleles that are under-represented in the ascertained array genotypes. This way, a sizable number of pairwise LD estimates comprising infrequent alleles become smaller so that a reduced LD profile breaks stretched LD blocks formed in the array-based experiments.

## TO WHAT EXTENT IS LD IN FARM ANIMALS INFLUENCED BY HUMANS?

Addressing this question requires speculating about the possible influence of domestication,breed establishment and animal farming on genetic factors implicating LD. Principally, LD is influenced by several factors, including drift, admixture, mutation and recombination rates, selection, finite population size, population bottlenecks, or other genetic events which a population experiences (reviewed in Slatkin, 2008). For example, population admixture creates sizable LD, depending on the similarity of the allele frequency profiles in the admixed populations. LD due to crossbreeding of inbred lines is significant but, it could be small when crossing breeds have similar gene frequencies, and it erodes quickly and disappears after a limited number of generations. Mutation, due to its minor effect on changing gene frequencies, has a negligible impact on the LD in the time frame of domestication. Selection is probably

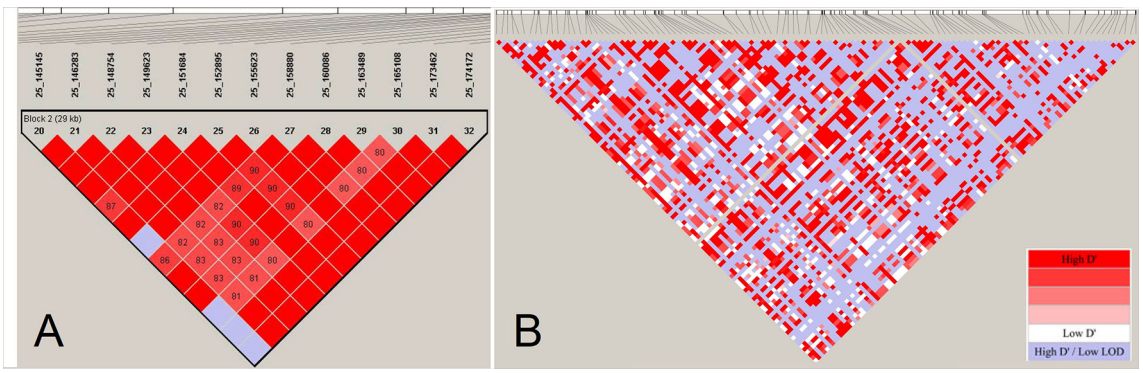


**FIGURE 5 |** The LD-block structuring as a function of SNP density. (Panel **A**) displays a LD block of length 29 Kb based on estimates of pair-wise *D'* among 13 SNPs located on BTA25 in Fleckvieh cattle. (Panel **B**) displays LD structure in the same region in sequencing resolution consisting of 115 markers. The LD blocks are obtained using "confidence intervals" algorithm (Gabriel et al., 2002) in Haploview (Barrett et al., 2005). LD analysis has been conducted with a constant number of individuals.

a significant cause of LD, however, its effect is likely localized around specific (major) genes, and so has relatively little effect on the amount of LD averaged across the genome.

While the buildup of LD can be a result of several population genetic forces, recombination isthe only primary mechanism to break it down. The absence of recombination between sites under selection can reduce the efficiency of selection in what is known as the 'Hill-Robertson effect' (Hill and Robertson, 1966). It is suggested that high rates of recombination during domestication have contributed to strong selection response (reviewed in Ross-Ibarra, 2004), but remains a debate since the evidences are ambiguous and inconclusive. The most recent study found no difference in the number and distribution of recombination breakpoints between dogs and wolves suggesting that both upper and lower bounds of crossover rates may be tightly regulated (Muñoz-Fuentes et al., 2015).

The finite population size is generally thought to be the leading cause of LD as effectivepopulation size has been severely eroded for most domestic species. For example, our experimentbased on sequence data suggests that chicken has experienced a drastic decline in*Ne*, evidencing a severe bottleneck most likely driven by domestication started inrecent past (see **Figure 6**). As shown, chicken hadthe largest effective population size 10,000 years ago which coincides with the generally accepted timing of chicken domestication (e.g., Xiang et al., 2014). The most recent *Ne* has dropped to a few hundred individuals and the Red Jungle Fowl (RJF) appears to have a larger population size present day in comparison to the commercial birds. A similar pattern of historical demography is observed in cattle (The Bovine HapMap Consortium, 2009). In human, the story is the opposite (The 1000 Genomes Project Consortium, 2015); improved agricultural productivity and industrialization have led to dramatic increases in population size. If LD is a result of the (current) finite population size, then the extent of LD should be many times more in livestock, as these species have *Ne* order of magnitude smaller (Leroy et al., 2013; Hall, 2016; Boitard et al., 2016) than the recent estimates reported for humans (Keinan and Clark, 2012; Browning and Browning, 2015). In reality, this is observed only for a portion of the marker pairs situated apart up to several hundreds of kilobases (Szyda et al., 2017). Instead, the observations based on full re-sequencing data revealed that the average genome-wide LD in chicken (see **Figure 4**) and cattle (Qanbari et al., 2014b) extends less than 40Kb, slightly greater than that in human populations. Since this is obtained from the full profile of polymorphisms, it represents the real strength of LD in these genomes, and far less than the extent previously reported.

Indeed, the observation of nearly comparable strength of LD in human and livestock is aconsequence of a sizable amount of polymorphism preserved in the genome of livestock. We observe millions of SNPs in the genome of cattle (e.g., Daetwyler et al., 2014) and chicken (Qanbari et al., 2019), in line with the latest updates of the genome sequencing projects in other livestock populations, including horse (Jagannathan et al., 2019), pig (Rubin et al., 2012), and sheep (Naval-Sanchez et al., 2018) that identified tens of millions SNP variants. This is comparable to the polymorphism content found in the human genome on the basis of sequencing several hundreds of individuals (The 1000 Genomes Project Consortium, 2015).

Hypothetically, the observed level of nucleotide diversity is much larger than a small population with *Ne* as low as several tens or hundreds is expected to generate or carry. This implies that chicken and cattle must have experienced much larger *Ne* in their history, which is indeed what exactly emerges from demographic inferences in these species. For example, analysis of sequence data suggests that chicken had a historical *Ne* around

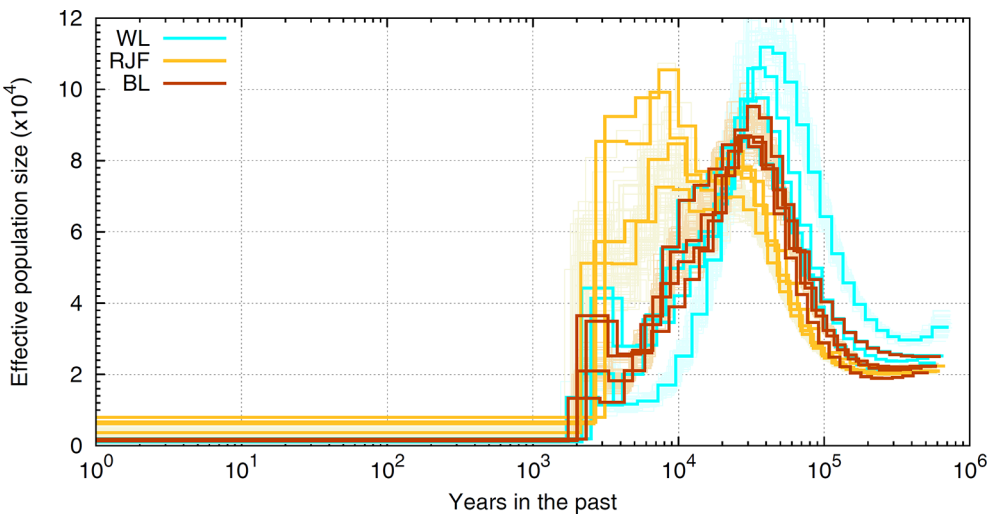


**FIGURE 6 |** A schematic illustration of historical *Ne* in chicken. The ancestral demography is inferred in sequence resolution for RJF and white (WL) and brown (BL) layers employing the Pairwise Sequentially Markovian Coalescent [PSMC, Li and Durbin (2011)] framework. The scale on the x-axis is years in the past and the scale on the y-axis represents the historical effective population numbers. Orange (RJF), brown (BL), and cyan (WL) lines represent inferred demography for different populations with bootstraps in lighter colors. Note that inferences of bootstraps are depicted only for one sample of each population.

25,000 at 1 million years ago that persisted for several hundreds of thousands years, before chicken population expanded starting from 50,000 to 100,000 years ago (see **Figure 6**). A somewhat similar picture of ancestral demography was also reported for the bovine genome (The Bovine HapMap Consortium 2009). Comparing the LD pattern across breeds of livestock species can reveal the influence of humans in shaping the genetic buildup. LD have been reported across breeds of cattle (Qanbari et al., 2011; Porto-Neto et al., 2014; Makina et al., 2015), sheep (Al-Mamun et al., 2015; Prieur et al., 2017), pig (Badke et al., 2012; Ai et al., 2013; Muñoz et al., 2019), buffalo (Deng et al., 2019; Mokhber et al., 2019), chicken (Khanyile et al., 2015; Hérault et al., 2018), and horse (Wade et al., 2009; McCue et al., 2012, Marchiori et al., 2019), among others. The general trend is that in local breeds or populations that experienced less intensive breeding programs, LD decays faster between distant markers than the commercial populations in which, LD extends for larger pairwise distances. For example, Holstein exhibits extensive LD than the other cattle breeds, despite having the largest contemporary population. In comparison, Indicine breeds have a lower LD than Taurine, suggestive of a larger ancestral population (e.g., Porto-Neto et al., 2014). The involvement of human in shaping genetic makeup of livestock is also evident in domestic chickens, where local breeds mostly exhibit shorter extent of LD (Khanyile et al., 2015) and among the commercials, the broilers presents faster decay of LD than layer populations (Pengelly et al., 2016; Seo et al., 2018 and Hérault et al., 2018). This is attributed to a more intensive selection scheme running over many generations during past several decades in layers resulting in a lower population haplotype diversity and a smaller *Ne*.

Further to the comparable polymorphism content, a somewhat similar pattern of allele frequency spectra (SFS) emerges in human and livestock genomes from sequence data (see Qanbari et al., 2014b and Qanbari et al., 2019). The SFS in livestock follows a decreasing trend consistent with many other organisms, including human (e.g., Nielsen et al., 2012). The distinction in livestock is that the spectra are skewed towards a larger fraction of intermediate frequencies (**Figure 4**). This is most likely stemming from an extremely small effective population size in present day livestock species and substantiates the significant under-representation of infrequent alleles in commercial breeds (e.g., see Muir et al., 2008 and Qanbari et al., 2019).

## GENOME-WIDE VARIATION IN LD

Across the genome, every chromosome behaves as a unique linkage group and may experience independent demography. This is similar to the inter-species or inter-population scenarios, where it generates different profiles of LD for each unit. LD levels are also higher for sex chromosomes than autosomes because recombination on the sex chromosomes only occurs in females. Previous studies of measuring LD revealed a substantial difference among chromosomes of farm animals (e.g., Sargolzaei et al., 2008). In human models, evidence also exists for significant variation in LD across genome, between sexes and among populations (Vega et al., 2005; Baudat et al., 2010; Kong et al., 2010, among others). Besides the recombination landscape which is the primary mechanism in shaping genome-wide LD, other factors such as genetic drift, demographic forces, mutation rate, and selection play a role as well. This depicts how challenging predicting LD between two sets of polymorphism based solely on physical distance could be. The design of LD mapping experiments and placement of SNPs will, therefore, require a thorough understanding of the local interplay of these factors for precisely localizing a target locus.

## THE DECAY OF LD IN HUMAN AND LIVESTOCK

LD persists for several hundreds of kilobases at least for a portion of marker pairs in the contemporary populations of chicken and cattle (Szyda et al., 2017; Hérault et al., 2018), which causes a slightly higher LD averaged over the genome compared to human. This is primarily stemming from the "family-based LD," a representation of the large chunks of chromosomes of founder animals segregating in the population. The consanguine parents transmit these identical-by-descent segments to the progenies and create uninterrupted stretches of homozygous genotypes, known as "run of homozygosity" (ROH), the hallmark of these autozygous segments inherited from a recent common ancestor (reviewed in Peripolli et al., 2017; Ceballos et al., 2018). The frequency, size, and distribution of ROH in the genome provide insights into the inbreeding, past demography, and selection in livestock populations (e.g., Bosse et al., 2012; Purfield et al., 2012, among others). In general, the extent of ROH islands is a function of the number of generations to the common ancestor, so that longer ROH indicate recent inbreeding, whereas ROH of older origin are generally shorter. The livestock populations involve more recent inbreeding loops through assortative mating, therefore, are expected to carry longer ROH than outbred populations like human that hold a much larger effective population size and diverse population (Gibson et al., 2006). Although a direct comparison of ROH between species in previous studies is impractical due to the lack of a gold standard in defining ROH islands, the extent to which the genome is covered by ROH tracts is expected to be higher in domestic animals relative to their wild counterparts. The long unbroken homozygosity hold in ROH islands, therefore, gives rise to an extended LD in livestock than that in human.

The unusually long ROH may also persist in outbred populations. These homozygosity islands may originate from the locally low mutation or recombination rates, or be a result of the positive selection for a favorable allele followed by the hitch-hiking of the polymorphism around the target locus (see section "Mapping selection").

## IMPLICATIONS FOR GENE MAPPING STUDIES

LD in sequencing resolution decays more rapidly than previously reported using array data. This enables higher resolution mapping of a trait of interest in outbred populations employing either association or selection mapping strategies. This also implies that selection mapping using haplotype-based metrics demands a panel of denser SNPs arrays to efficiently reveal patterns generated by unusually long haplotypes than medium-density arrays. The low reproducibility of the results reported in some of the first genome-wide selection studies in farm animal populations (e.g., Qanbari et al., 2010b) based on medium-density SNP arrays (~50 k SNPs) may be due to the lack of power prompted by overestimating the extent of LD demonstrated here. This is backed by our recent study in which extensive simulations were used to investigate the power of combining selection signatures detected with multiple methods under different scenarios of marker density, sample size, and selection intensity (Ma et al., 2015b). The authors showed that a reasonable power to detect selection signatures is achieved with high marker density (>1 SNP/Kb). Ultimately, uncovering older selective sweeps that carry shorter haplotypes will need sequencing resolution.

The extent of LD varies across the genomic regions, chromosomes, among populations and between species. In other words, genome-wide averaged estimates of the extent of LD may not adequately reflect LD patterns of specific regions or population groups. These observations have broader practical relevance in genomic studies of farm animals, as such the optimal number of samples and marker density in either genome-wide association or selection mapping studies may largely vary due to the extremely adverse pattern of LD within and among chromosomes. Finally, confounding population characteristics such as cryptic allelic correlations or stratification may have serious impact on pattern and structure of LD in livestock populations that need to be taken into consideration in conducting unbiased genome-wide association mapping (reviewed in Hellwege et al., 2017, also see Ma et al., 2012 and Bulik-Sullivan et al., 2015).

## LD ASSESSMENT SOFTWARE TOOLS

Estimating LD coefficients is computationally simple and can be performed using in-house scripts when the marker density is restricted to the genotypes of SNP arrays. $r^2$ is particularly straightforward to achieve based on built-in commands as it corresponds the spearman correlation between SNPs pairs. Moreover, the standard population genetics programs, among them are Haploview (Barrett et al., 2005) and Arlequin (Excoffier et al., 2005), along with several R packages provide tools to estimate LD statistics. In sequence resolution, however, estimation LD coefficients can be computationally burdensome specifically for the mega reference panels such as genome sequencing consortiums of different livestock species. For example, a panel of 1000 genomes of a mammalian species sequenced may include over 35M shared variants, which corresponds to over $4 \times 10^{11}$ pairwise LD coefficients within 1 Mbp windows genome-wide. A number of sophisticated programs to estimate LD statistics from sequencing data are freely available. PLINK is a widely used software toolkit for analyzing genetic data and is among the most computationally efficient tools for estimating LD (Purcell et al., 2007). VCFtools is another widely used software toolkit for manipulating and analyzing genetic data that provide utilities to estimate LD from the Variant Call Format (VCF) (Danecek et al., 2011). VCFtools works with compressed VCF files (VCF.gz) which require far less storage space than PLINK BED files; however, it can be computationally demanding for large data sets. M3VCFtools (Das et al., 2016), an extension of VCFtools uses a compact haplotype representation format called M3VCF, to estimate LD statistics. M3VCF requires far less storage than genotype formats. M3VCF toolkit provides more efficient querying and data processing and has option to convert a VCF file into M3VCf format.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## FUNDING

This research is financially supported by the grants from the German Research Foundation (DFG, project ChickenSeq ID. QA55/1-1) and the Federal Ministry of Education and Research (BMBF, project CLARITY, ID. 031L0166). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## ACKNOWLEDGMENTS

I sincerely thank Henner Simianer, Dörte Wittenburg, and Abdulraheem Arome Musa for reviewing the paper and valuable comments that significantly improved the manuscript. I acknowledge support by the Open Access Publication Fund of the Leibniz Institute for Farm Animal Biology (FBN).

## REFERENCES

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PloS One* 8, e56001. doi: 10.1371/journal.pone.0056001

Al-Mamun, H. A., Clark, S. A., Kwan, P., and Gondro, C. (2015). Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. *Genet. Selection Evol.* 47, 90. doi: 10.1186/s12711-015-0169-6

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., and Steibel, J. P. (2012). Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* 13, 24. doi: 10.1186/1471-2164-13-24

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

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., et al. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327, 836–840. doi: 10.1126/science.1183439

Bengtsson, B. O., and Thomson, G. (1981). Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* 18, 356–363.

Boitard, S., Rodríguez, W., Jay, F., Mona, S., and Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach. *PloS Genet.* 12, e1005877. doi: 10.1371/journal.pgen.1005877

Bosse, M., Megens, H.-J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., et al. (2012). Regions of Homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PloS Genet.* 8, e1003100. doi: 10.1371/journal.pgen.1003100

Browning, S. R., and Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* 97, 404–418. doi: 10.1016/j.ajhg.2015.07.012

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J. Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 74, 106–120. doi: 10.1086/381000

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., and Wilson, J. F. (2018). Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* 19, 220–234. doi: 10.1038/nrg.2017.109

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

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

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

Deng, T., Liang, A., Liu, J., Hua, G., Ye, T., Liu, S., et al. (2019). Genome-wide snp data revealed the extent of linkage disequilibrium, persistence of phase and effective population size in purebred and crossbred buffalo populations. front. *Genet* 9. doi: 10.3389/fgene.2018.00688

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1, 47–50. doi: 10.1177/117693430500100003

Fu, W., Dekkers, J. C., Lee, W. R., and Abasht, B. (2015). Linkage disequilibrium in crossbred and pure line chickens. *Genet. Sel. Evol.* 47, 11. doi: 10.1186/s12711-015-0098-4

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424

Gianola, D., Qanbari, S., and Simianer, H. (2013). An evaluation of a novel estimator of linkage disequilibrium. *Heredity (Edinb)* 111, 275–285. doi: 10.1038/hdy.2013.46

Gibson, J., Morton, N. E., and Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* 15, 789–795. doi: 10.1093/hmg/ddi493

Goddard, M. E., and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10, 381–391. doi: 10.1038/nrg2575

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

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13, 635–643. doi: 10.1101/gr.387103

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

Hellwege, J., Keaton, J., Giri, A., Gao, X., Velez Edwards, D. R., and Edwards, T. L. (2017). Population stratification in genetic association studies. *Curr. Protoc. Hum. Genet.* 95, 1.22.1–1.22.23. doi: 10.1002/cphg.48

Hérault, F., Herry, F., Varenne, A., Burlot, T., Picard–Druet, D., Recoquillay, J., et al. (2018). "A linkage disequilibrium study in layers and broiler commercial chicken populations," in *Proceedings of the World Congress on Genetics Applied to Livestock Production (WCGALP)*(Auckland, NZL). (2018-02-11 - 2018-02-16).

Hill, W. G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* 8, 269–294. doi: 10.1017/S0016672300010156

Hill, W. G., and Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* 54, 705–714.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium1. *Genet. Res.* 38, 209–216. doi: 10.1017/S0016672300020553

Jagannathan, V., Gerber, V., Rieder, S., Tetens, J., Thaller, G., Drögemüller, C., et al. (2019). Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Anim. Genet.* 50, 74–77. doi: 10.1111/age.12753

Jennings, H. S. (1917). The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. *Genetics* 2, 97–154.

Jorde, L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res.* 10, 1435–1444. doi: 10.1101/gr.144500

Keinan, A., and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743. doi: 10.1126/science.1217283

Khanyile, K. S., Dzomba, E. F., and Muchadeyi, F. C. (2015). Population genetic structure, linkage disequilibrium and effective population size of conserved and extensively raised village chicken populations of Southern Africa. *Front. Genet.* 6, 13. doi: 10.3389/fgene.2015.00013

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467, 1099–1103. doi: 10.1038/nature09525

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144. doi: 10.1038/9642

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23. doi: 10.1016/j.ajhg.2014.06.009

Leroy, G., Mary-Huard, T., Verrier, E., Danvy, S., Charvolin, E., and Danchin-Burge, C. (2013). Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. *Genet. Selection Evol.* 45, 1. doi: 10.1186/1297-9686-45-1

Lewontin, R. C., and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* 14, 458–472. doi: 10.1111/j.1558-5646.1960.tb03113.x

Lewontin, R. C. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics* 49, 49–67.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genet.* 120 (3), 849–852.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231

Ma, L., Wiggans, G. R., Wang, S., Sonstegard, T. S., Yang, J., Crooker, B. A., et al. (2012). Effect of sample stratification on dairy GWAS results. *BMC Genomics* 13, 536. doi: 10.1186/1471-2164-13-536

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., et al. (2015a). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PloS Genet.* 11, e1005387. doi: 10.1371/journal.pgen.1005387

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015b). Properties of different selection signature statistics and a new strategy for combining them. *Heredity* 115, 426–436. doi: 10.1038/hdy.2015.42

Makina, S. O., Taylor, J. F., van Marle-Köster, E., Muchadeyi, F. C., Makgahlela, M. L., MacNeil, M. D., et al. (2015). Extent of linkage disequilibrium and effective population size in four South African Sanga Cattle breeds. *Front. Genet.* 6, 337. doi: 10.3389/fgene.2015.00337

Marchiori, C. M., Pereira, G. L., Maiorano, A. M., Rogatto, G. M., Assoni, A. D., Augusto, I. I. V., et al. (2019). Linkage disequilibrium and population structure characterization in the cutting and racing lines of Quarter Horses bred in Brazil. *Livestock Sci.* 219, 45–51. doi: 10.1016/j.livsci.2018.11.013

Maruki, T., and Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics* 197, 1303–1313. doi: 10.1534/genetics.114.165514

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., et al. (2012). A high density SNP array for the domestic horse and extant perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PloS Genet.* 8, e1002451. doi: 10.1371/journal.pgen.1002451

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.

Mokhber, M., Shahrbabak, M. M., Sadeghi, M., Shahrbabak, H. M., Stella, A., Nicolzzi, E., et al. (2019). Study of whole genome linkage disequilibrium patterns of Iranian water buffalo breeds using the Axiom Buffalo Genotyping 90K Array. *PloS One* 14, e0217687. doi: 10.1371/journal.pone.0217687

Muñoz, M., Bozzi, R., García-Casco, J., Núñez, Y., Ribani, A., Franci, O., et al. (2019). Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high density SNP chip. *Sci. Rep.* 9, 1–14. doi: 10.1038/s41598-019-49830-6

Muñoz-Fuentes, V., Marcet-Ortega, M., Alkorta-Aranburu, G., Linde Forsberg, C., Morrell, J. M., Manzano-Piedras, E., et al. (2015). Strong artificial selection in domestic mammals did not result in an increased recombination rate. *Mol. Biol. Evol.* 32, 510–523. doi: 10.1093/molbev/msu322

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Brief Bioinform.* 5, 355–364. doi: 10.1093/bib/5.4.355

Muir, W. M., Wong, G. K.-S., Zhang, Y., Wang, J., Groenen, M. A. M., Crooijmans, R. P. M. A., et al. (2008). Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *PNAS* 105, 17312–17317. doi: 10.1073/pnas.0806569105

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9, 859. doi: 10.1038/s41467-017-02809-1

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS One* 7, e37558. doi: 10.1371/journal.pone.0037558

Ober, U., Malinowski, A., Schlather, M., and Simianer, H. (2013). *The expected linkage disequilibrium in finite populations revisited*, Mannheim Available at: http://arxiv.org/pdf/1304.4856v2.pdf [Accessed June 13, 2019].

Pearson, K. (1901). I. Mathematical contributions to the theory of evolution. — VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. London. Ser. A Containing Papers Math. Phys. Character* 195, 1–47. doi: 10.1098/rsta.1900.0022

Pengelly, R. J., Gheyas, A. A., Kuo, R., Mossotto, E., Seaby, E. G., Burt, D. W., et al. (2016). Commercial chicken breeds exhibit highly divergent patterns of linkage disequilibrium. *Heredity (Edinb)* 117, 375–382. doi: 10.1038/hdy.2016.47

Peripolli, E., Munari, D. P., Silva, M. V. G. B., Lima, A. L. F., Irgang, R., and Baldi, F. (2017). Runs of homozygosity: current knowledge and applications in livestock. *Anim. Genet.* 48, 255–271. doi: 10.1111/age.12526

Petit, M., Astruc, J.-M., Sarry, J., Drouilhet, L., Fabre, S., Moreno, C. R., et al. (2017). Variation in recombination rate and its genetic determinism in sheep populations. *Genetics* 207, 767–784. doi: 10.1534/genetics.117.300123

Plackett, R. L. (1965). A class of bivariate distributions. *J. Am. Stat. Assoc.* 60, 516–522. doi: 10.1080/01621459.1965.10480807

Porto-Neto, L. R., Kijas, J. W., and Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. *Genet. Selection Evol.* 46, 22. doi: 10.1186/1297-9686-46-22

Prieur, V., Clarke, S. M., Brito, L. F., McEwan, J. C., Lee, M. A., Brauning, R., et al. (2017). Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. *BMC Genet.* 18, 68. doi: 10.1186/s12863-017-0534-2

Pritchard, J. K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69, 1–14. doi: 10.1086/321275

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Purfield, D. C., Berry, D. P., McParland, S., and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genet.* 13, 70. doi: 10.1186/1471-2156-13-70

Qanbari, S., and Simianer, H. (2014a). Mapping signatures of positive selection in the genome of livestock. *Livestock Sci.* 166, 133–143. doi: 10.1016/j.livsci.2014.05.003

Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R., et al. (2010a). The pattern of linkage disequilibrium in German Holstein cattle. *Anim. Genet.* 41, 346–356. doi: 10.1111/j.1365-2052.2009.02011.x

Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R., et al. (2010b). A genome-wide scan for signatures of recent selection in Holstein cattle. *Anim. Genet.* 41, 377–389. doi: 10.1111/j.1365-2052.2009.02016.x

Qanbari, S., Hansen, M., Weigend, S., Preisinger, R., and Simianer, H. (2010c). Linkage disequilibrium reveals different demographic history in egg laying chickens. *BMC Genet.* 11, 103. doi: 10.1186/1471-2156-11-103

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., et al. (2011). Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* 12, 318. doi: 10.1186/1471-2164-12-318

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

Qanbari, S., Rubin, C.-J., Maqbool, K., Weigend, S., Weigend, A., Geibel, J., et al. (2019). Genetics of adaptation in modern chicken. *PloS Genet.* 15, e1007989. doi: 10.1371/journal.pgen.1007989

Ross-Ibarra, J. (2004). The evolution of recombination under domestication: a test of two hypotheses. *Am. Nat.* 163, 105–112. doi: 10.1086/380606

Rubin, C.-J., Megens, H.-J., Martinez Barrio, A., Maqbool, K., Sayyab, S., Schwochow, D., et al. (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19529–19536. doi: 10.1073/pnas.1217149109

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

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., and Schaeffer, L. R. (2008). Extent of linkage disequilibrium in Holstein cattle in North America. *J. Dairy Sci.* 91, 2106–2117. doi: 10.3168/jds.2007-0553

Seo, D., Lee, D. H., Choi, N., Sudrajad, P., Lee, S.-H., and Lee, J.-H. (2018). Estimation of linkage disequilibrium and analysis of genetic diversity in Korean chicken lines. *PloS One* 13, e0192063. doi: 10.1371/journal.pone.0192063

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. doi: 10.1038/nrg2361

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672300014634

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2, 125–141. doi: 10.1016/0040-5809(71)90011-6

Sved, J. A. (2009). Linkage disequilibrium and its expectation in human populations. *Twin Res. Hum. Genet.* 12, 35–43. doi: 10.1375/twin.12.1.35

Szyda, J., Suchocki, T., Qanbari, S., Liu, Z., and Simianer, H. (2017). Assessing the degree of stratification between closely related Holstein-Friesian populations. *J. Appl. Genet.* 58, 521–526. doi: 10.1007/s13353-017-0409-2

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56, 777–787.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

The Bovine HapMap Consortium, Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A.. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528–532. doi: 10.1126/science.1167936

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., et al. (2012). A high density recombination map of the pig reveals a correlation

between sex-specific recombination and GC content. *BMC Genomics* 13, 586. doi: 10.1186/1471-2164-13-586

Vega, F. M. D. L., Isaac, H., Collins, A., Scafe, C. R., Halldórsson, B. V., Su, X., et al. (2005). The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res.* 15, 454–462. doi: 10.1101/gr.3241705

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326, 865–867. doi: 10.1126/science.1178158

Waples, R. K., Larson, W. A., and Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity (Edinb)* 117, 233–240. doi: 10.1038/hdy.2016.60

Xiang, H., Gao, J., Yu, B., Zhou, H., Cai, D., Zhang, Y., et al. (2014). Early Holocene chicken domestication in northern China. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17564–17569. doi: 10.1073/pnas.1411882111

Zhao, H., Nettleton, D., Soller, M., and Dekkers, J. C. M.(2005). Evaluation of linkage disequilibrium measures betweenmulti-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet. Research* 86, 77–87. doi: 10.1017/S001667230500769X

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:** The 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fgene.2020.00134

# Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium

*Sohyoung Won[1†], Jong-Eun Park[2†], Ju-Hwan Son[2], Seung-Hwan Lee[3], Byeong Ho Park[2], Mina Park[2], Won-Chul Park[2], Han-Ha Chai[2], Heebal Kim[1,4,5], Jungjae Lee[6*] and Dajeong Lim[2*]*

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [2] National Institute of Animal Science, RDA, Wanju, South Korea, [3] Department of Animal Science and Biotechnology, Chungnam National University, Daejeon, South Korea, [4] Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, South Korea, [5] eGnome, Inc, Seoul, South Korea, [6] Jung P&C Institute, Inc., Yongin-si, South Korea

OPEN ACCESS

**Edited by:**
Marco Milanesi,
São Paulo State University,
Brazil

**Reviewed by:**
Zhe Zhang,
South China Agricultural University,
China
Gregor Gorjanc,
University of Edinburgh,
United Kingdom

***Correspondence:**
Jungjae Lee
jungjae.ansc@gmail.com
Dajeong Lim
lim.dj@korea.kr

[†]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:** 26 July 2019
**Accepted:** 04 February 2020
**Published:** 06 March 2020

**Citation:**
Won S, Park J-E, Son J-H, Lee S-H, Park BH, Park M, Park W-C, Chai H-H, Kim H, Lee J and Lim D (2020) Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium. Front. Genet. 11:134. doi: 10.3389/fgene.2020.00134

Genomic prediction is an effective way to estimate the genomic breeding values from genetic information based on statistical methods such as best linear unbiased prediction (BLUP). The used of haplotype, clusters of linked single nucleotide polymorphism (SNP) as markers instead of individual SNPs can improve the accuracy of genomic prediction. Since the probability of a quantitative trait loci to be in strong linkage disequilibrium (LD) with a cluster of markers is higher compared to an individual marker. To make haplotypes efficient in genomic prediction, finding optimal ways to define haplotypes is essential. In this study, 770K or 50K SNP chip data was collected from Hanwoo (Korean cattle) population consisted of 3,498 cattle. Using SNP chip data, haplotype was defined in three different ways based on 1) the number of SNPs included, 2) length of haplotypes (bp), and 3) agglomerative hierarchical clustering based on LD. To compare the methods in parallel, haplotypes defined by all methods were set to have comparable sizes; 5, 10, 20 or 50 SNPs on average per haplotype. A linear mixed model using haplotype to calculated the covariance matrix was applied for testing the prediction accuracy of each haplotype size. Also, conventional SNP-based linear mixed model was tested to evaluate the performance of the haplotype sets on genomic prediction. Carcass weight (CWT), eye muscle area (EMA) and backfat thickness (BFT) were used as the phenotypes. This study reveals that using haplotypes generally showed increased accuracy compared to conventional SNP-based model for CWT and EMA, but found to be small or no increase in accuracy for BFT. LD clustering-based haplotypes specifically the five SNPs size showed the highest prediction accuracy for CWT and EMA. Meanwhile, the highest accuracy was obtained when length-based haplotypes with five SNPs were used for BFT. The maximum gain in accuracy was 1.3% from cross-validation and 4.6% from forward validation for EMA, suggesting that genomic prediction accuracy can be increased by using haplotypes. However, the improvement from using haplotypes may depend on the trait of interest. In addition, when the number of alleles generated by each haplotype

defining methods was compared, clustering by LD generated the least number of alleles, thereby reducing computational costs. Therefore, finding optimal ways to define haplotypes and using the haplotype alleles as markers can improve the accuracy of genomic prediction.

**Keywords: genomic prediction, haplotype, hierarchical clustering, linkage disequilibrium, best linear unbiased prediction, accuracy, Hanwoo**

## INTRODUCTION

Genomic prediction is an effective way to measure the genetic merit and breeding values of livestock based on their genetic information. Practically, genotype data of the animals particularly the single nucleotide polymorphisms (SNP) and statistical prediction methods such as the best linear unbiased prediction (BLUP) are required to calculate the genomic estimated breeding values (GEBV). The accuracy of genomic prediction depends on the degree of linkage disequilibrium (LD) between the SNP markers and real quantitative trait loci (QTL) (Goddard, 2009). Fundamentally, linkage disequilibrium is a nonrandom association between different loci in a certain population, which can be calculated by measuring the frequencies of alleles and the haplotype frequencies of the pair of alleles at the loci (Slatkin, 2008).

By using clusters of related SNPs as markers instead of individual SNPs, the probability that a QTL is in strong LD with a marker becomes higher (Goddard and Hayes, 2007). Thus, the accuracy of genomic prediction can be improved by using clusters of SNPs, which are referred to as haplotypes. With the higher LD with QTLs, haplotypes better detect identity-by-descent structure while making the genomic relationship matrix, resulting in increased genomic prediction accuracy (Hess et al., 2017). To make efficient use of haplotypes in genomic predictions, numerous studies have focused on finding optimal ways to define a cluster of SNPs as a haplotype. The simplest way is to consider equal sizes of segments in the genome as haplotypes (Villumsen and Janss, 2009; Sun et al., 2015; Ferdosi et al., 2016; Hess et al., 2017). By this method, equal size can be determined through physical length in base pairs (Ferdosi et al., 2016; Hess et al., 2017), the length in centimorgans (Sun et al., 2015), or the number of SNPs (Villumsen et al., 2009). In addition, methods to define haplotypes such as combining information about identity by descent (IBD) with clusters of adjacent SNPs (Calus et al., 2008; Calus et al., 2009), and using predicted genealogy (Edriss et al., 2013) were studied. Also, setting minimum pairwise LD cutoffs to grouped SNPs into haplotypes was considered (Cuyabano et al., 2014).

Some of the methods to define haplotypes for genomic prediction attempts to incorporate the LD structure of the genome (Calus et al., 2008; Cuyabano et al., 2014; Cuyabano et al., 2015). Lesser number of haplotype alleles brings an advantage in LD based haplotypes since the number of explanatory variables used for computation is reduce compared to other methods (Cuyabano et al., 2014). Recently, the application of some clustering methods originated in the data mining field represent a more precise LD structure when defining haplotypes (Dehman, 2015). Among these methods is hierarchical clustering, which produces a tree that has nodes representing clusters in a hierarchical order from, where each element being each cluster is the leaf the all the elements being one cluster is the root. Applying hierarchical clustering to make SNP clusters based on LD was implemented to genome-wide association study (Dehman, 2015).

In this study, agglomerative hierarchical clustering was used to construct haplotypes based on LD from phased genotypes of 770K SNP chips. In addition, haplotypes were alternatively defined as segments with given sizes. The length of a haplotype in base pairs and the number of SNPs within a haplotype were respectively used as criteria of sizes. Differently define haplotypes were tested and compared with the accuracy of using individual SNPs to find out whether which method can bring improvement in genomic prediction. Also, to find out the optimal size of haplotypes, various sizes of haplotypes defined by each method were tested. To compare the methods in parallel, haplotypes defined by all methods were set to have comparable sizes.

## MATERIALS AND METHODS

### Genotypic and Phenotypic Data

The genotypic and phenotypic information were collected from the 3,498 Hanwoo (Korean cattle) population. Animal health and welfare issues were followed according to the appropriate guidelines approved by the Animal Care and Use Committee of the National Institute of Animal Science, Rural Development Administration, Korea. Available information such as sex and slaughter age was used for analysis. The traits analyzed in this study were carcass weight (CWT), eye muscle area (EMA) and backfat thickness (BFT), measured after slaughter. Genotyping was performed using Illumina BovineHD 770K Genotyping BeadChip for 1,166 samples and Illumina BovineSNP50 Genotyping BeadChip for 2,332 samples. The 50K genotypes were imputed to 770K using Eagle (https://data.broadinstitute.org/alkesgroup/Eagle/) and Minimac3 (http://genome.sph.umich.edu/wiki/Minimac3) pipeline.

For further analyses, SNPs having low minor allele frequency (<0.01), low genotyping rate (<0.95), significant deviation from Hardy–Weinberg equilibrium ($p < 0.001$) were discarded, while only one SNP was kept if multiple SNPs were located on the same site. Individuals with low genotyping call rate (<0.95) were excluded from the study. From the data collecting stage, phenotypes including sex and slaughter age of some animals were not fully recorded and were removed from the study. Moreover, two-sided Grubb's test with alpha = 0.05 was performed to check whether

there were outliers in phenotypic data. Test results revealed that one sample of BFT and two samples of EMA were considered outlier. After the removal of identified outliers, none of the tests were significant ($p < 0.05$) with $p = 0.80$ for CWT, $p = 0.14$ for EMA, and $p = 0.10$ for BFT. Similarly, nine significant outliers from the covariate age were also removed.

Thus, the total number of SNPs used for genomic prediction was 555,678 from 2,494 animals (821 males and 1,673 females) The summary statistics of the phenotype data are presented in **Table 1**, while the distributions of the phenotypes used in this study are presented in **Supplementary Figure 1**. The total genotyping rate was 0.9971. Genotypes were phased and imputed using SHAPEIT2 with 200 states and a window size of 0.5 Mb for haplotyping (Delaneau et al., 2012).

## Defining Haplotypes

Three methods to define haplotypes were considered respectively in this study. First, segments of the genome containing constant number of SNPs were treated as haplotypes (method 1). Second, segments of the genome with equal sizes in basepairs were regarded as haplotypes (method 2). Third, hierarchical clustering based on LD was used to construct haplotypes (method 3). In these three methods, the start and end points of haplotypes were designated accordingly and the SNPs within the point formed haplotypes.

In each method, we varied the sizes of haplotypes to find out the optimal size of haplotypes for accurate genomic prediction. To compare the three methods in a comparable way, the average number of SNPs per block were balanced to be approximately 5, 10, 20, or 50. Briefly, three haplotype defining methods with four average size criteria, making twelve kinds of haplotype were tested. The lengths of haplotypes in method 1 was calculated by dividing the total length of the genome by the total number of SNPs, then multiplying 5, 10, 20, or 50. In method 3, the number of clusters (number of haplotype regions) were set as the total number of SNPs divided by 5, 10, 20, or 50. The lengths of haplotypes in method 1 and number of clusters in method 3 are later shown in **Table 2**.

## Hierarchical Clustering Based on LD

In hierarchical clustering based on LD, the pairwise LD between SNPs were calculated as D', based on the following equation (Lewontin, 1964).

$$D_{AB} = p_{AB} - p_A p_B$$

$$D_{max} = \begin{cases} \max(-p_A p_B,\ -(1-p_A)(1-p_B)) & \text{when } D < 0 \\ \min(p_A(1-p_B),\ (1-p_A)p_B) & \text{when } D > 0 \end{cases}$$

$$D' = D_{AB}/D_{max}$$

**TABLE 1** | Summary statistics of the phenotypes used for the study.

| | Minimum | 1st Qt. | Median | Mean | 3rd Qt. | Maximum |
|---|---|---|---|---|---|---|
| **CWT** | 197 | 335 | 374 | 377.5789 | 415 | 623 |
| **EMA** | 42 | 77 | 84 | 84.85138 | 92 | 126 |
| **BFT** | 1 | 7 | 10 | 11.02117 | 14 | 39 |

*CWT, carcass weight (kg); EMA, eye muscle area ($cm^2$); BFT, backfat thickness (mm).*

**TABLE 2** | Haplotype and allele statistics of each haplotype defining method at different sizes.

| **SNP count-based haplotypes** | **5 SNPs** | **10 SNPs** | **20 SNPs** | **50 SNPs** |
|---|---|---|---|---|
| Number of haplotype alleles | 1,303,861 | 1,877,160 | 2,713,296 | 3,710,659 |
| Number of haplotypes | 111,123 | 55,554 | 27,768 | 11,099 |
| Average number of SNPs per haplotypes | 5 | 10 | 20 | 50 |
| Average number of alleles per haplotypes | 11.73349 | 33.78983 | 97.71305 | 334.3237 |
| Minimum SNPs in haplotypes | 5 | 10 | 20 | 50 |
| Maximum SNPs in haplotypes | 5 | 10 | 20 | 50 |
| **Length-based haplotypes** | **22.25 kb** | **44.5 kb** | **89 kb** | **222.5 kb** |
| Number of haplotype allele markers | 1,364,861 | 1,867,261 | 2,621,574 | 3,581,059 |
| Number of haplotypes | 97,061 | 54,163 | 27,797 | 11,196 |
| Average number of SNPs per haplotypes | 5.725038 | 10.25936 | 19.99057 | 49.63183 |
| Average number of alleles per haplotypes | 14.06188 | 34.47484 | 94.31140 | 319.8516 |
| Minimum SNPs in haplotypes | 2 | 2 | 2 | 2 |
| Maximum SNPs in haplotypes | 29 | 47 | 71 | 136 |
| **LD clustering-based haplotypes** | **K = N/5** | **K = N/10** | **K = N/20** | **K = N/50** |
| Number of haplotype alleles | 1,277,525 | 1,764,074 | 2,472,637 | 3,358,562 |
| Number of haplotypes | 111,123 | 55,554 | 27,768 | 11,099 |
| Average number of SNPs per haplotypes | 5.000567 | 10.00248 | 20.01145 | 50.06559 |
| Average number of alleles per haplotypes | 11.49649 | 31.75422 | 89.04628 | 302.6004 |
| Minimum SNPs in haplotypes | 1 | 1 | 1 | 1 |
| Maximum SNPs in haplotypes | 114 | 131 | 141 | 213 |

*K is the number of clusters and N is the number of total SNPs.*

Clustering groups similar objects together. Here, SNPs with high LD were regarded as similar SNPs and were assigned to the same clusters. In other words, the measure of LD, D' was set as the proximity measure of two SNPs and (1 – D') was defined as the distance between two SNPs in the clustering algorithm. To define the distance between two clusters, complete linkage was used. In complete linkage clustering, the link between two clusters contains all element pairs, and the distance between two clusters is measured as the maximum pairwise distance among all elements in the clusters. Here, the distance between clusters was defined as the maximum of 1 – D' between all pairwise SNPs in two clusters. Agglomerative hierarchical clustering is an iterative process of merging clusters starting from each element being a cluster of its own (Rokach and Maimon, 2005). First, two clusters with the closest distance are found and are merged to form a new cluster. After two clusters were merged, the distance between clusters is updated by calculating the distances between the new clusters and the others. This is repeated until the number of clusters reaches the threshold, which was the total number of SNPs divided by 5, 10, 20, or 50.

In this study, to make non-overlapping and linear clusters using all the SNPs for defining haplotype, only physically adjacent SNPs or clusters were merged by keeping a linear distance list of adjacent clusters instead of a distance matrix.

For example, when the ith and the (i + 1)th clusters were merged as the I ∗ th, the distances between the (i − 1)th and ith cluster, ith and (i + 1)th cluster, (i+1)th and the (I + 2)th cluster are removed from the list and the distance of the (i − 1)th and the i ∗ th cluster, the i ∗ th cluster and the (i + 2)th cluster are added to the list for updating. In this way, when finding the closest two clusters from the list, only the distances between adjacent clusters are being considered.

## Haplotype Alleles and Diplotypes

After defining the start and endpoints of haplotypes throughout the genome, the phased genotype was re-coded according to the haplotype alleles. The individual diplotypes were then coded as 0, 1 or 2 for each haplotype allele in a haplotype region. This results in an N × H matrix, where and N is the number of animals and H is the total number of haplotype alleles. R package 'GHap' was used for this procedure (Utsunomiya et al., 2016).

## Genomic Prediction

A linear mixed model was used to perform genomic predictions using the haplotype markers defined in the previous stage. The model was described as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \epsilon,$$

where y is the vector of observations (CWT, BFT and EMA), **b** is the vector of fixed effects including sex and slaughter age, **g** is the vector of additive genetic effects, **ϵ** is the vector of residual errors, and **X** is the design matrix for fixed effects. The additive genetic effects **g** and residual errors **ϵ** were assumed as random effects assuming that it follows the distributions specified bellow:

$$\mathbf{g} \sim \mathbf{N}\left(0,\ \mathbf{G}\sigma_g^2\right)$$
$$\epsilon \sim \mathbf{N}\left(0,\ \mathbf{I}\sigma_e^2\right)$$

Here, **G** is the genetic relatedness matrix and **I** is an identity matrix. **G** was calculated from the following equation.

$$\mathbf{G} = \frac{\mathbf{MM\prime}}{2\sum p_i(1-p_i)}$$

**M** was the haplotype matrix obtained from the haplotyping step (*Haplotype Alleles and Diplotypes*) adjusted for allele frequencies. The *ij*th element of **M** is calculated as $m_{ij} = (x_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$, where $x_{ij}$ is the number of *j*th haplotype allele carried by the *i*th animal and $p_j$ is the minor allele frequency of the *j*th haplotype allele. For the SNP-based model, **M** was the matrix of genotype adjusted for minor allele frequency.

The BLUP solution of the linear mixed model, $\hat{\mathbf{u}}$ was computed using the equation $\hat{\mathbf{u}} = \mathbf{M}'\mathbf{G}^{-1}\hat{\mathbf{g}}/\mathrm{N}$, from restricted expectation maximization (REML). GCTA software was used for computation (Yang et al., 2011). Heritability was also estimated from REML by estimating the variance components $\sigma_g^2$ and $\sigma_e^2$ with GCTA.

Then, the GEBVs were obtained as the following equation:

$$\mathrm{GEBV} = \mathbf{M}\hat{\mathbf{u}}$$

Finally, the performances of different haplotype definitions were compared based on the accuracy of the models, which was calculated as the correlation of the GEBVs and pre-corrected phenotypes. Sex and slaughter age were used for pre-correction. Five times of 5-fold cross-validation (5 × 5 cross-validation) were performed to access the accuracies of different methods.

In addition, forward validation was done to access the performance of predicting breeding values of younger animals from the data of older animals. Animals born from January 2012 were assigned to test set and the remaining animals were assigned as a training set. Training set and test set consisted of 2,015 animals and 479 animals respectively. The accuracy was calculated as the correlation between predicted GEBVs and pre-corrected phenotypes as in cross-validation.

# RESULTS

## Haplotype Construction

The statistics of haplotypes constructed by different haplotype defining methods and the different average SNP number criteria in each method are presented in **Table 2** and **Supplementary Figure 2**. The actual average numbers of SNPs per haplotype were also obtained and evaluated to check whether the haplotypes were constructed with intended sizes. The average numbers of SNPs were consistent with the intended numbers in LD clustering-based haplotypes and length-based haplotypes with sizes of 44.5kb, 89kb and 222.5kb, while larger than intended in length-based haplotypes of 22.25kb.

The total number of haplotype alleles were computed to compare the number of explanatory variables used for genomic prediction (**Table 2**). The number of alleles increased as the average number of SNPs per haplotype increased. However, the numbers of alleles from haplotypes of similar sizes were where found to be smaller when LD clustering was used to define haplotypes. The average number of alleles per haplotypes showed similar tendencies with the total number of alleles.

## Genomic Prediction Accuracy

The genomic prediction accuracies from 5 × 5-fold cross-validation of haplotypes defined by three methods were higher compared to the SNP-based model except for haplotypes with 50 SNPs in CWT and EMA (**Figure 1**). For both CWT and EMA, LD clustering based-haplotypes with an average of 5 SNPs showed the highest gain in terms of accuracy. Prediction accuracy increased from 0.435 to 0.448 for CWT and 0.319 to 0.331 for EMA, which were 1.2% and 1.3%, respectively. Conversely, there was no observed improvement in prediction accuracy in BFT.

Meanwhile, when forward validation was used for testing prediction accuracy, the tendency of accuracies was similar, however, the overall accuracy was lower while the gain in

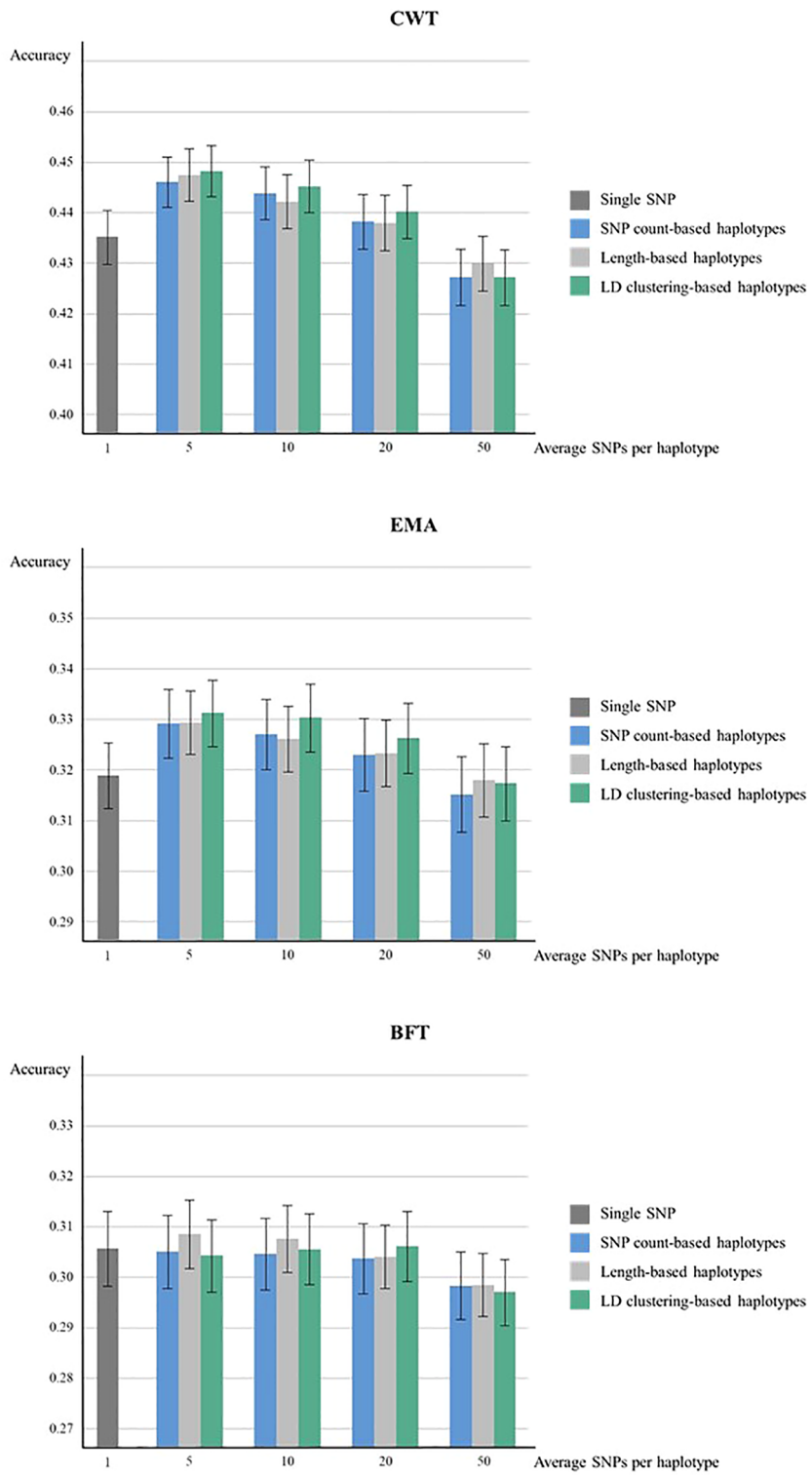


**FIGURE 1 |** Genomic prediction accuracies from five time five-fold cross validation. Prediction accuracies of using various sizes of haplotypes defined by different methods and using individual SNPs were compared for CWT, BFT and EMA respectively. The black lines on the bars show standard errors of the prediction accuracies. Accuracies were calculated as the correlation coefficients of GEBVs and pre-corrected phenotypes.

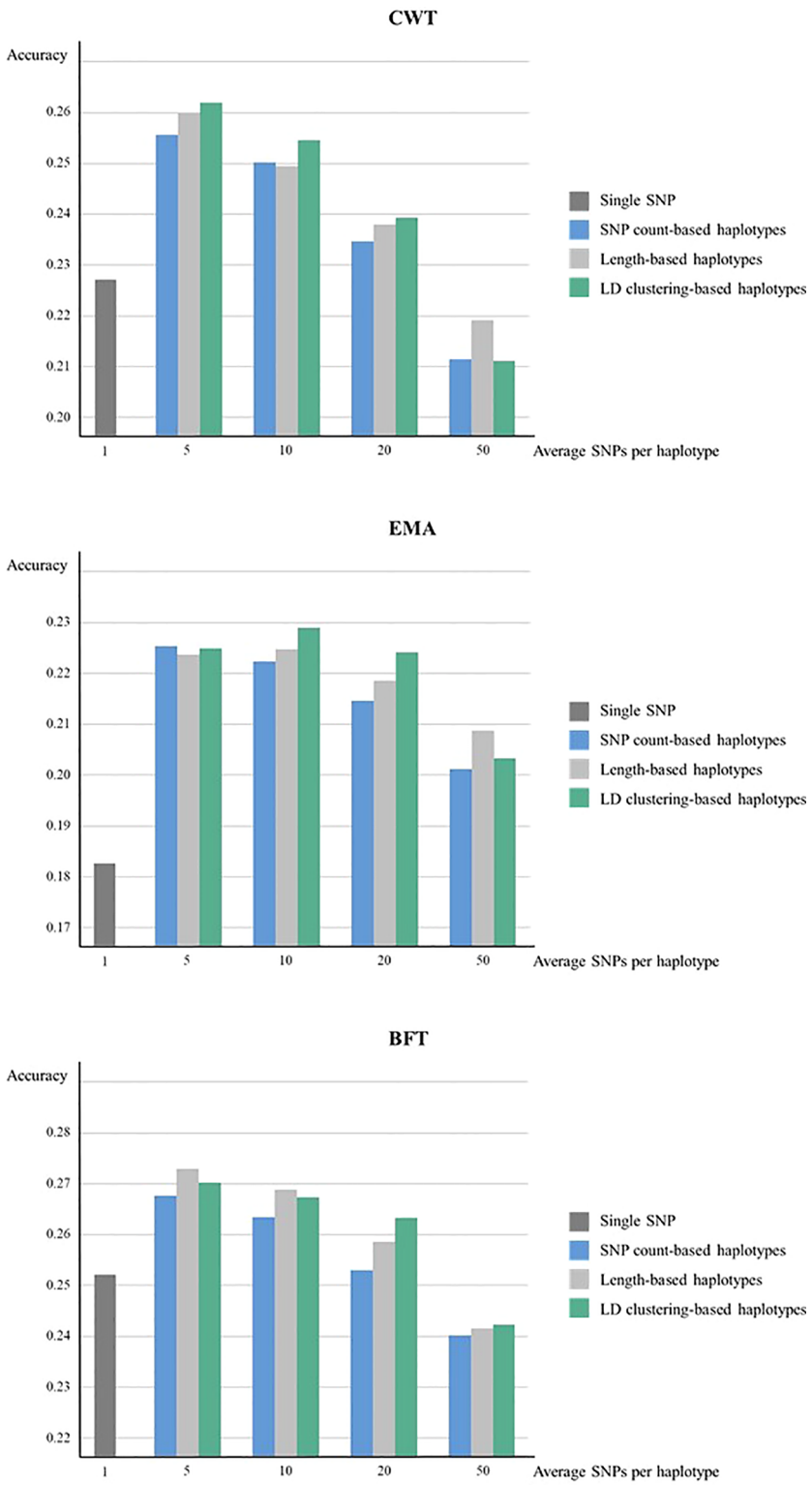


**FIGURE 2 |** Genomic prediction accuracies from forward validation. Prediction accuracies of using various sizes of haplotypes defined by different methods and using individual SNPs were compared for CWT, BFT and EMA respectively. Accuracies were calculated as the correlation coefficients of GEBVs and pre-corrected phenotypes.

accuracy by using haplotypes was larger (**Figure 2**). LD clustering-based haplotypes with 5 and 10 SNPs showed the highest accuracy for both CWT and EMA, respectively. Moreover, length-based haplotypes with five SNPs showed the highest accuracy for BFT. Numerically, the maximum increase in prediction accuracy was 3.5% for CWT, 4.6% for EMA, and 2.1% for BFT.

The prediction accuracy of haplotype-based model tended to decrease as the size of haplotypes became larger in all haplotype defining methods. Overall, LD clustering-based haplotypes showed the highest accuracy for all sizes except for 50 SNPs.

Paired t-tests were performed to determine whether the increases in prediction accuracies by using haplotypes compared to individual SNPs were statistically significant (**Table 3**). Statistical tests were also performed for different haplotype defining methods with different sizes for three traits. Results revealed that an observed increase in prediction accuracy in haplotypes with 5 or 10 SNPs defined by three methods were found to be statistically significant in both CWT and EMA.

Also, the heritability of the three traits were estimated using haplotypes and individual SNPs (**Table 4**). Estimated heritability for each trait using individual SNPs was 0.36, 0.43, 0.31 for CWT, BFT and EMA respectively. Interestingly, estimated heritability estimate using haplotypes was higher in all traits with values ranging from 0.38 to 0.43 for CWT, 0.44 to 0.52 for BFT and 0.33 to 0.38 for EMA.

**TABLE 3 |** P-values of paired t-tests comparing prediction accuracies using individual SNPs and haplotypes defined by different methods and sizes.

| | Average number of SNPs per haplotype | | | |
|---|---|---|---|---|
| **CWT** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.002** | 0.01* | 0.21 | 0.98 |
| **Length-based haplotypes** | 0.0008** | 0.03* | 0.23 | 0.92 |
| **LD clustering-based haplotypes** | 0.0005** | 0.005** | 0.09 | 0.98 |
| **EMA** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.00004** | 0.004** | 0.12 | 0.81 |
| **Length-based haplotypes** | 0.00007** | 0.007** | 0.09 | 0.58 |
| **LD clustering-based haplotypes** | 0.0002** | 0.002** | 0.07 | 0.86 |
| **BFT** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.64 | 0.67 | 0.77 | 0.99 |
| **Length-based haplotypes** | 0.07 | 0.20 | 0.74 | 0.99 |
| **LD clustering-based haplotypes** | 0.77 | 0.52 | 0.43 | 1.00 |

** and ** indicates significant at $\alpha = 0.05, 0.01$ respectively.*

**TABLE 4 |** Estimated heritabilities using haplotypes defined by different methods and sizes and using individual SNPs.

| | Average number of SNPs per haplotype | | | |
|---|---|---|---|---|
| **CWT** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.39 | 0.39 | 0.41 | 0.43 |
| **Length-based haplotypes** | 0.38 | 0.39 | 0.40 | 0.42 |
| **LD clustering-based haplotypes** | 0.39 | 0.39 | 0.41 | 0.43 |
| **Individual SNPs** | 0.36 | | | |
| **EMA** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.33 | 0.34 | 0.35 | 0.38 |
| **Length-based haplotypes** | 0.33 | 0.34 | 0.35 | 0.38 |
| **LD clustering-based haplotypes** | 0.33 | 0.34 | 0.36 | 0.38 |
| **Individual SNPs** | 0.43 | | | |
| **BFT** | **5** | **10** | **20** | **50** |
| **SNP count-based haplotypes** | 0.45 | 0.46 | 0.48 | 0.52 |
| **Length-based haplotypes** | 0.44 | 0.45 | 0.47 | 0.50 |
| **LD clustering-based haplotypes** | 0.44 | 0.45 | 0.46 | 0.50 |
| **Individual SNPs** | 0.43 | | | |

## DISCUSSION

Genomic prediction accuracy using haplotypes designed in this study was mostly higher than using individual SNPs and was statistically significant in the best performing haplotypes for CWT and EMA. The increased accuracy by using haplotypes may be due to higher LD between alleles and QTLs, better detection of ancestral relationships (identity-by-descent), and capturing of short-range epistatic effects (Hess et al., 2017). Haplotyping and constructing genomic prediction models using haplotype alleles can improve prediction accuracy without any additional cost for data production though it may cause some more computational cost. The maximum gain in accuracy was more than 1% in 5 × 5 cross-validation and more than 4% in forward validation, suggesting that genomic prediction accuracy can be improved by using haplotypes. However, improvement depends on traits of interest, some traits may elicit the same results with the use of haplotypes for the genomic prediction but other traits may also result contrariwise.

In addition, although overall prediction accuracy was low in forward validation, the used of haplotypes still brought higher prediction accuracy. Only length-based haplotypes with 5 or 10 SNPs showed higher accuracy than SNP-based model in EMA when 5 × 5 cross validation was used while all haplotypes with 5, 10 or 20 SNPs showed increased accuracy in forward validation. Also, prediction accuracy increased using haplotypes with 50 SNPs for EMA in forward validation but not in 5 × 5 cross-validation. This shows that haplotypes can be more effectively used for predicting the breeding values of younger animals from older animals, thereby making it more useful for animal breeding purposes.

Haplotype defining method with highest accuracy were found to differ in each trait, specifically LD clustering for CWT and EMA, while length-based haplotypes for BFT. Explicitly, LD clustering-based haplotypes showed the highest accuracies at all sizes except 50 SNPs for both CWT and EMA, and 20 SNPs for BFT. Generally, using LD clustering-based haplotypes resulted in high prediction accuracies. However, the effect of haplotype size was greater than the effect of haplotype defining method on prediction accuracy. In terms of haplotype size, the average five SNPs for all three traits preformed best. In general, the prediction accuracy was higher when smaller haplotypes were used. In larger haplotypes, some redundant markers may be present, for

instance, haplotype alleles carried by only few animals which will result in low prediction accuracy.

The optimal size to define haplotypes for genomic prediction depends on the distance between SNPs and the LD structure of the population (Calus et al., 2009). The mean distance between SNPs was 4,118.24 bp and the mean LD ($r^2$) was 0.43 in the Hanwoo population used for the study. In this study, the haplotype size of best performance was 5 SNPs, while in other studies the optimal numbers of SNPs per haplotype were 4–10, while genotype sizes ranged from 5,000 to 50,000 SNPs (Calus et al., 2009; Villumsen and Janss, 2009; Hess et al., 2017). Further study testing the haplotypes sizes ranging from 2 to 10 may be proceeded to find the optimal haplotype size in Hanwoo.

The number of haplotype alleles indicates the number of explanatory variables used for genomic prediction. As the number of explanatory variables increases, the dimension of the design matrix in the equation becomes larger, taking more time and memory to solve the mixed model equation. Thereby, reducing the number of haplotype alleles enables more efficient calculation of GEBVs. In this study, two methods are possible to reduce the number of haplotype alleles. The first is LD clustering to define haplotypes and the second is using smaller sizes of haplotypes. However, the effect of haplotype size was larger than the effect of haplotype defining method on number of alleles. Considering both prediction accuracy and the number of haplotype alleles, LD clustering was the optimal method for CWT and EMA.

Higher heritability estimate values were obtained using haplotypes compared to individual SNPs. Estimated heritability tended to increase as the number of haplotype alleles increased. As the number of alleles increases, more markers are used to explain the phenotypic variance, thus a higher proportion of total variance can be explained, resulting in higher heritability. However, caution is needed to interpret genomic heritability since there may be bias in the likelihood estimate of the variance components caused by linkage equilibrium between some markers and QTLs (de los Campos et al., 2015). In this study, the estimated heritabilities did not differ much with the results of other studies regarding Hanwoo where the estimated heritability of CWT, BFT and EMA were 0.30–0.33, 0.27–0.41 and 0.35–0.50, respectively (Yoon et al., 2002; Park et al., 2013; Lee et al., 2014).

The estimation of GEBV from haplotype alleles depends on the imputation and phasing results from genotypes. Errors from imputation or phasing may produce wrong alleles that are not actually carried by the sample. Especially in haplotypes defined by LD clustering, inaccurate phasing may cause haplotype boundaries to be differently defined resulting in lower accuracy. Therefore, finding more accurate phasing methods can further improve the prediction accuracy by using haplotypes. Besides, methods modeling the genetic relatedness from haplotype similarity can be considered to resolve such inaccuracies occurring from phasing errors (Hickey et al., 2013). In addition, discarding haplotype alleles of low frequencies by regarding them to have zero effects can be considered, since the generation of alleles having an extremely low frequency (e.g. only one in the population) can be a cause of overfitting, potentially lowering the prediction accuracy. Also, this can reduce the computational cost by lessening explanatory variables.

In this study, the advantage of using haplotypes in genomic prediction was testes in the Hanwoo population. Some studies that tested other livestock populations reported that haplotypes can be advantageous for genomic prediction. Applying haplotype to genomic prediction has been studied in Montbeliarde bulls (Jónás et al., 2016), New Zealand dairy cattle (Hess et al., 2017), Nordic Holstein (Cuyabano et al., 2014; Cuyabano et al., 2015), and Danish Holstein bulls (Edriss et al., 2013). Although different haplotypes were used in these studies and the design of the studies may differ, their study still shows the benefits of using haplotype for genomic prediction. Therefore, we expect that applying the haplotypes defined in this study can bring improvement to prediction performance not only in Hanwoo but also in other livestock populations. However, the optimal size of haplotype may vary from population to population and most of the studies about haplotype and genomic prediction were tested in dairy cattle or beef cattle. Thus, care should be taken when applying to other species.

In conclusion, genomic prediction using haplotypes in the Hanwoo population showed increase accuracy for three carcass traits, CWT, BFT and EMA. Haplotypes used for genomic prediction were defined by three methods, length, SNP count and hierarchical clustering based on LD with four different sizes. The haplotype defining method showing the highest prediction accuracy was LD clustering-based haplotypes with five SNPs for CWT and EMA and length-based haplotypes with 5 SNPs for BFT. LD clustering-based haplotypes had the least number of alleles, being favorable in terms of computation time. However, haplotype optimization methods for various traits need to be continuously.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the National Agricultural Biotechnology Information Center (NABIC) http://nabic.rda.go.kr/ostd/basic/snpVcfView.do?selectedId=NV-0618-000001.

## ETHICS STATEMENT

The animal study was reviewed and approved by National Institute of Animal Science.

## AUTHOR CONTRIBUTIONS

DL and SW conceived and designed the study. JL and J-HS were responsible for imputation of 50K and 777K genotype data to sequence level. BP and MP responsible for phenotypic data

collection. W-CP, H-HC and HK contributed in quality control of genotype data. All authors read and agreed on the contents of manuscript.

## FUNDING

This work was carried out with the support of "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ01251902)" Rural Development Administration, Republic of Korea.

## ACKNOWLEDGMENTS

We acknowledge to National Agricultural Cooperative Federation, Seosan, Korea for providing semen samples of Korean native cattle (Hanwoo) steers.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00134/full#supplementary-material

## REFERENCES

Calus, M., De Roos, A., and Veerkamp, R. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553–561. doi: 10.1534/genetics.107.080838

Calus, M. P., Meuwissen, T. H., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L., et al. (2009). Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.* 41, 11. doi: 10.1186/1297-9686-41-11

Cuyabano, B. C., Su, G., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15, 1171. doi: 10.1186/1471-2164-15-1171

Cuyabano, B. C., Su, G., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47 (1), 61 doi: 10.1186/s12711-015-0143-3

de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? *PloS Genet.* 11 (5), e1005048. doi: 10.1371/journal.pgen.1005048

Dehman, A. (2015). *Spatial clustering of linkage disequilibrium blocks for genome-wide association studies* (Île-de-France: Université d'Evry Val d'Essonne; Université Paris-Saclay; Laboratoire de).

Delaneau, O., Zagury, J.-F., and Marchini, J. J. N. M. (2012). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5. doi: 10.1038/nmeth.2307

Edriss, V., Fernando, R. L., Su, G., Lund, M. S., and Guldbrandtsen, B. (2013). The effect of using genealogy-based haplotypes for genomic prediction. *Genet. Sel. Evol.* 45, 5. doi: 10.1186/1297-9686-45-5

Ferdosi, M. H., Henshall, J., and Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. *Genet. Sel. Evol.* 48, 75. doi: 10.1186/s12711-016-0253-6

Goddard, M., and Hayes, B. (2007). Genomic selection. *J. Anim. Breed. Genet.* 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* 49, 54. doi: 10.1186/s12711-017-0329-y

Hickey, J. M., Kinghorn, B. P., Tier, B., Clark, S. A., van der Werf, J. H. J., and Gorjanc, G. (2013). Genomic evaluations using similarity between haplotypes. *J. Anim. Breed. Genet.* 130 (4), 259–269. doi: 10.1111/jbg.12020

Jónás, D., Ducrocq, V., Fouilloux, M. N., and Croiseau, P. (2016). Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.* 99 (6), 4537–4546. doi: 10.3168/jds.2015-10433

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

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49 (1), 49.

Park, B., Choi, T., Kim, S., and Oh, S.-H. (2013). National genetic evaluation (system) of Hanwoo (Korean native cattle). *Asian Australas. J. Anim. Sci.* 26, 151. doi: 10.5713/ajas.2012.12439

Rokach, L., and Maimon, O. (2005). "Clustering methods," in Data mining and knowledge discovery handbook. *Springer*, 321–352. doi: 10.1007/0-387-25465-X_15

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477. doi: 10.1038/nrg2361

Sun, X., Fernando, R. L., Garrick, D. J., and Dekkers, J. (2015). Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes. *Anim. Ind. Rep.* 661, 86. doi: 10.31274/ans_air-180814-1339

Utsunomiya, Y. T., Milanesi, M., Utsunomiya, A. T., Ajmone-Marsan, P., and Garcia, J. F. (2016). GHap: an R package for genome-wide haplotyping. *Bioinformatics* 32, 2861–2862. doi: 10.1093/bioinformatics/btw356

Villumsen, T. M., and Janss, L. (2009). Bayesian genomic selection: the effect of haplotype length and priors. *BMC Proc.*, S11 BioMed Central. doi: 10.1186/1753-6561-3-S1-S11

Villumsen, T., Janss, L., and Lund, M. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126, 3–13. doi: 10.1111/j.1439-0388.2008.00747.x

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yoon, H. B., Kim, S. D., Na, S. H., Chang, U. M., Lee, H. K., Jeon, G. J., et al. (2002). Estimation of genetic parameters for carcass traits in Hanwoo steer. *J. Anim. Sci. Technol.* 44 (4), 383–390. doi: 10.5187/JAST.2002.44.4.383

**Conflict of Interest:** JL was employed by company Jung P&C Institute, Inc and HK was employed by company eGnome, Inc.

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: 09 June 2020
doi: 10.3389/fgene.2020.00544

# Ancestral Haplotype Mapping for GWAS and Detection of Signatures of Selection in Admixed Dairy Cattle of Kenya

*Hassan Aliloo*[1]*, *Raphael Mrode*[2,3], *A. M. Okeyo*[2] and *John P. Gibson*[1]

[1] School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia, [2] Animal Biosciences, International Livestock Research Institute, Nairobi, Kenya, [3] Animal and Veterinary Science, Scotland's Rural College, Edinburgh, United Kingdom

Understanding the genetic structure of adaptation and productivity in challenging environments is necessary for designing breeding programs that suit such conditions. Crossbred dairy cattle in East Africa resulting from over 60 years of crossing exotic dairy breeds with indigenous cattle plus inter se matings form a highly variable admixed population. This population has been subject to natural selection in response to environmental stresses, such as harsh climate, low-quality feeds, poor management, and strong disease challenge. Here, we combine two complementary sets of analyses, genome-wide association (GWA) and signatures of selection (SoS), to identify genomic regions that contribute to variation in milk yield and/or contribute to adaptation in admixed dairy cattle of Kenya. Our GWA separates SNP effects due to ancestral origin of alleles from effects due to within-population linkage disequilibrium. The results indicate that many genomic regions contributed to the high milk production potential of modern dairy breeds with no region having an exceptional effect. For SoS, we used two haplotype-based tests to compare haplotype length variation within admixed and between admixed and East African Shorthorn Zebu cattle populations. The integrated haplotype score (iHS) analysis identified 16 candidate regions for positive selection in the admixed cattle while the between population Rsb test detected 24 divergently selected regions in the admixed cattle compared to East African Shorthorn Zebu. We compare the results from GWA and SoS in an attempt to validate the most significant SoS results. Only four candidate regions for SoS intersect with GWA regions using a low stringency test. The identified SoS candidate regions harbored genes in several enriched annotation clusters and overlapped with previously found QTLs and associations for different traits in cattle. If validated, the GWA and SoS results indicate potential for SNP-based genomic selection for genetic improvement of smallholder crossbred cattle.

**Keywords: local ancestry inference, admixed cattle, GWAS, signatures of selection, haplotype**

OPEN ACCESS

***Edited by:***
*Marco Milanesi,*
*São Paulo State University, Brazil*

***Reviewed by:***
*Gerson Oliveira,*
*University of Guelph, Canada*
*Hussain Mahdi Bahbahani,*
*Kuwait University, Kuwait*

***Correspondence:**
*Hassan Aliloo*
*haliloo@une.edu.au*

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

***Received:*** *24 October 2019*
***Accepted:*** *06 May 2020*
***Published:*** *09 June 2020*

***Citation:***
*Aliloo H, Mrode R, Okeyo AM and Gibson JP (2020) Ancestral Haplotype Mapping for GWAS and Detection of Signatures of Selection in Admixed Dairy Cattle of Kenya. Front. Genet. 11:544. doi: 10.3389/fgene.2020.00544*

## INTRODUCTION

Exotic dairy breeds have been extensively imported to Kenya since the 1950s for use in crossbreeding aimed to improve the productivity of indigenous cattle by combining the environmental adaptation features of the latter with the high milk yield potential of the former. This has resulted in a large population of admixed cattle that, for several decades, have been subject to natural selection in response to different environmental stresses, such as harsh climate, low-quality feeds, poor management, and strong disease challenge. Smallholder farmers retain the admixed cattle typically in herds of size one to five cows and breed them mainly through natural mating to local bulls. A small proportion of matings (∼10%) are made by AI to imported and locally produced purebred dairy bulls. There is no record of pedigree or performance of smallholder cattle and no current genetic improvement program for crossbred cattle. Genomic technologies can aid smallholder dairy farmers to develop genetically improved animals when the genetic improvement by traditional breeding schemes is impossible (Mrode et al., 2018; Marshall et al., 2019; Ojango et al., 2019).

With high-density SNP markers, it is possible to identify genomic regions that may be useful in future selection. This can be done through genome-wide association (GWA) analysis, which relies on linkage disequilibrium (LD) between SNPs and causal variants and requires phenotype plus genotype data, and by detection of signatures of selection, which only requires genotypic information. In admixed populations, the LD between SNP markers and causal variants can arise from the LD that existed in the parental populations that contributed to the admixed population and from *de novo* LD that was created when crossing populations (Cole and Silva, 2016). Performing a standard GWA in an admixed population doesn't have the same power as that in a purebred population. This is because the within-population LD is not expected to be the same in all the ancestral populations, and the ancestral within-population LD differs from the *de novo* LD that is created by the crossing process. However, it is possible to separately map the within-breed LD with causal variants from the between-breed LD with causal variants that are fixed or are at very high frequencies for different alleles in different ancestral populations (the variants that contribute to the phenotype differences between ancestral breeds) if alleles in the admixed population can be correctly assigned to their ancestral origin. The latter can be done through methods that infer the ancestry of haplotypes, such as LAMP-LD (Baran et al., 2012). Detecting the presence of causative loci that differentiate ancestral populations is of particular interest in crosses between *Bos taurus* dairy breeds and African indigenous breeds given their huge (up to 10-fold) difference in milk production potential.

When a beneficial allele increases in frequency by natural or artificial selection, the allele frequencies of neighboring loci in LD are also altered, and this creates extended blocks of haplotypes with increased LD and reduced variation. The changes in allele frequencies, LD, and genetic variation accumulate over time and generate unique patterns at specific regions of genome, which are referred to as signatures of selection (Walsh and Lynch, 2018). The identification of signatures of selection in modern livestock populations can help to uncover genes and biological mechanisms involved in the domestication process, breed formation, and artificial selection for economically important traits as well as local adaptation to new environments. Several genome scans aimed to detect recent and past selection have been implemented for purebred (e.g., Qanbari et al., 2014) and composite (e.g., Goszczynski et al., 2018) breeds as well as admixed livestock populations (Gautier and Naves, 2011; Bahbahani et al., 2018; Cheruiyot et al., 2018).

In admixed populations generated by crossing genetically differentiated ancestral breeds, the first generation of crosses retains intact haplotypes from parental breeds. Recombination in subsequent generations of within-population matings breaks down the parental haplotypes and forms mosaicism that expands as the admixed population ages. The fragmentation of ancestral haplotypes across generations can be assessed through the ancestry mapping of closely linked markers to obtain information about the history of the admixed population (Freeman et al., 2006). Since a recent admixture can mimic the patterns of variation left by selection around a selected site and introduce noise in detection of selection signatures (Lohmueller et al., 2010), it is necessary to take the admixture process into account before searching for any post-admixture selection signal in admixed populations.

Several statistical methods have been developed for detection of genomic footprints of selection that essentially compare the patterns of genetic variation within or between populations and decide on whether one should accept or reject the null hypothesis of "no selection" and interpret the test statistics as evidence for selection or not (see review by Vitti et al., 2013). Among the different approaches designed to identify positive selection, the haplotype-based methods are more powerful because they combine information from patterns of allele frequencies and persistence of LD. The extended haplotype homozygosity (EHH) statistic developed by Sabeti et al. (2002) measures the probability of being identical by descent for any two randomly chosen chromosomes within a population carrying a core genomic region surrounding a presumably selected allele. Voight et al. (2006) proposed a within-population variation of EHH based on the contrast between the integral of the EHH for derived (selected) and ancestral (control) alleles called integrated haplotype score (iHS). The iHS test is especially powerful in detection of recent selection that has swept the selected allele to moderate frequencies, but the selected allele has not yet been fixed. A complementary method for iHS to detect sweeps near fixation is the between-population Rsb test proposed by Tang et al. (2007). The Rsb statistic compares the integrated EHH profiles between pairs of populations and searches for alleles that have been targeted by selection and swept toward fixation in one population but not in the other. There are several examples of application of iHS and Rsb statistics for detecting both recent and ancient positive selections in different livestock population (Bahbahani et al., 2015; Cheruiyot et al., 2018).

Here, we use 521,362 autosomal SNPs and scan the genome of 1,475 admixed cattle from Kenya in (1) a GWA analysis

that separates breed origin SNP effects from effects due to within-population LD to find SNPs associated with milk yield and (2) a signature of selection (SoS) analysis to detect signals of post-admixture selection. The GWA and SoS analyses are complementary because in relatively young populations, SoS are not expected to have led to fixation of alleles, and therefore, the results from one can be used as partial validation of the other.

# MATERIALS AND METHODS

## Genotypes

The genotypic data included 1,475 admixed and 19 East African Shorthorn Zebu (EASZ) cattle sampled in Kenya between 2010 and 2014 and genotyped for 777,962 SNP markers using Illumina BovineHD BeadChip (Illumina, San Diego, CA). More information on collection of samples can be found in Aliloo et al. (2018). We retained the autosomal SNPs for analysis. The genotype calls with a GC score $< 0.6$ were set as missing, and then, SNPs with a call rate $> 0.95$ were kept. A reference set of high-density genotypes of 105 pure *Bos indicus* animals (IND) from 12 Indian breeds were obtained by stratified sampling of the larger data set analyzed by Strucken et al. (2019). Reference genotypes were also obtained for six different cattle populations representing the two other major ancestral groups in East Africa, i.e., (i) African taurine (AFT) ancestors of indigenous cattle: NDama (ND, $n = 24$) and (ii) European taurine (EUT) ancestors of admixed cattle: Holstein (HO, $n = 71$), Jersey (JE, $n = 46$), Guernsey (GU, $n = 21$), British Friesian (BF, $n = 26$), and Ayrshire (AY, $n = 519$). All genotypes except BF and AY, which were provided by the Scottish Rural University College (SRUC) and Canadian Dairy Network (CDN), respectively, were obtained from the Bovine HapMap Consortium (http://bovinegenome.org). These genotypes were obtained post-quality control, so only the common SNPs between them and African and Indian genotypes were extracted. We sampled an equal number of 21 animals from each EUT breed and considered the five EUT breeds as recent ancestors of Kenyan admixed dairy cattle. SNPs with a MAF less than 0.01 across the whole sample were excluded. Animals were also required to have genotypes for more than 90% of SNPs. These controls resulted to 521,362 SNPs on 1,475 admixed, 19 EASZ, 105 IND, 24 AFT, and 105 EUT animals distributed over 29 autosomes based on the UMD3.1 bovine reference genome. Details of the cattle populations in this study are presented in **Table 1**.

**TABLE 1 |** Details of the different cattle populations used in this study.

| Breed group | Source | Original population size | Sample size | Ancestral group* |
|---|---|---|---|---|
| Kenyan crossbred | Kenya | 1,475 | 1,475 | – |
| East African Shorthorn Zebu | Kenya | 19 | 19 | – |
| Dangi | India | 65 | 13 | IND |
| Gavlao | India | 19 | 4 | IND |
| Gir | India | 118 | 24 | IND |
| Hallikar | India | 27 | 5 | IND |
| Haryana | India | 11 | 2 | IND |
| Khilar | India | 24 | 5 | IND |
| Krishnavalley | India | 17 | 3 | IND |
| Lalkandhari | India | 35 | 7 | IND |
| Malinar Gidda | India | 14 | 3 | IND |
| Ongole | India | 46 | 9 | IND |
| Sahiwal | India | 104 | 21 | IND |
| Tharparkar | India | 45 | 9 | IND |
| NDama | HapMap | 24 | 24 | AFT |
| Holstein | HapMap | 71 | 21 | EUT |
| Jersey | HapMap | 46 | 21 | EUT |
| Guernsey | HapMap | 21 | 21 | EUT |
| British Friesian | UK | 26 | 21 | EUT |
| Ayrshire | Canada | 519 | 21 | EUT |

**IND, Bos indicus; AFT, African Bos taurus; and EUT, European Bos taurus.*

## Phenotypes

Milk yield deviations (MYD) were obtained for the individual test-days of 1,034 (out of 1,475) Kenyan admixed cows in smallholder farms from the analyses of Brown et al. (2016). In their analyses, test-day milk yields (TDMY) were analyzed using a model that included fixed effects for parity and Legendre polynomial of order 4 fitted for each of five dairy breed classes. The dairy breed classes were assigned based on admixture (Alexander et al., 2009) estimates of total dairy breed proportion for each animal using SNP genotypes (Ojango et al., 2019). Random effects were included for contemporary management group-year-season, animal permanent environment, and animal additive genetic effects, using a genomic relationship matrix based on VanRaden (2008). The MYD were obtained by correcting the TDMY for fixed effects plus the random management group effect (Brown et al., 2016).

## Population Structure Analysis

To investigate the population structure of admixed cattle in relation to the ancestral breeds, a principal component analysis (PCA) based on all SNP genotypes after quality control (521,362) was implemented. The PCA was applied to a (co)variance matrix between all animals' genotypes (**G**) constructed using the VanRaden (2008) method. The first and second principal component were plotted to visualize the distribution of admixed cattle across the different ancestral breeds.

## Local Ancestry Estimation of Admixed Sample

To infer the local ancestry of admixed cattle at individual SNPs, we used LAMP-LD software (Baran et al., 2012) with three groups of ancestral haplotypes, i.e., IND, AFT, and combined EUT. The admixed population being analyzed results from crosses between local indigenous cattle, i.e., the EASZ and EUT breeds. The indigenous cattle are known to be old, probably ancient, admixtures of *Bos indicus* and African *Bos taurus* cattle (Strucken et al., 2017). Thus, in the absence of a large sample of the

indigenous EASZ population, we used IND and AFT as proxies to track the indigenous haplotypes. The genotypes of all individuals, i.e., ancestors and admixed animals, were phased together using Eagle v2.4 (Loh et al., 2016) to provide haplotypes for local ancestry inference and also for calculation of test statistics for detection of selection signatures across the admixed genome. LAMP-LD uses hidden Markov models of haplotype diversity of ancestral populations within a window-based framework to trace the origin of alleles in the admixed population (Baran et al., 2012). We used the default input parameters, i.e., a 300-SNP window size and 15 as the number of states, to run LAMP-LD and obtained the local ancestries of admixed animals.

## Crossover Events Across the Admixed Genome

The local ancestry inferences obtained above were used to calculate the average number of crossover events across the admixed genome. We defined a recent crossover as the transition from either IND or AFT ancestry to EUT ancestry and vice versa. For each haplotype of a given admixed individual, we counted the number of recent crossovers and standardized it by chromosome length to obtain the number of crossover events per Morgan. For this calculation we assumed a recombination rate of 1 Morgan = 100 Mbp. Then we ranked the two haplotypes of each admixed individual within each chromosome from lowest to highest number of crossovers. Finally, the average (across all chromosomes) frequency of crossovers in haplotypes with lowest number of crossovers was used to rank the admixed animals.

## Genome-Wide Association Mapping

A mixed linear model was used to test for associations between genome-wide SNPs and MYD of the admixed cattle. A single SNP regression model (fitting one SNP at a time) simultaneously estimated the effect of the ancestral origin (exotic vs. indigenous) of the SNP and the residual effects of SNP alleles after accounting for the ancestral origin. The local ancestry inferences obtained above were used to assign the ancestral origin of SNP alleles with ancestral origin coded as 0, 1, and 2 for no copies, one copy, or two copies coming from the EUT ancestor, respectively. The GWA model was as follows:

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\beta + \mathbf{Wu} + \mathbf{Wpe} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of MYD of size $n$, $\mathbf{1}_\mathrm{n}$ is a vector of ones, $\mu$ is the population mean term, $\boldsymbol{\beta}$ is a $2 \times 1$ vector containing the ancestral origin of allele effect and residual SNP effect, $\mathbf{u}$ contains polygenic effects assumed to be distributed as $u \sim N(0, \mathbf{G}\sigma_g^2)$ with $\mathbf{G}$ being the genomic relationship matrix based on all SNP genotypes except the SNPs on the chromosome of the marker for which the association is tested (VanRaden, 2008), $\mathbf{pe}$ is the vector of random permanent environment effects with $pe \sim N(0, \mathbf{I}\sigma_{pe}^2)$, and $\mathbf{e}$ is the vector of random residual deviates assumed to be distributed as $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. $\mathbf{X}$ is an $n \times 2$ design matrix allocating genotypes to ancestral origin of allele effect and residual marker effect, and $\mathbf{W}$ is the incidence matrix for the random animal and permanent environmental effects. $\sigma_g^2$, $\sigma_{pe}^2$, and $\sigma_e^2$ are polygenic additive, permanent environment, and residual variances, respectively.

The above model was fitted by WOMBAT (Meyer, 2007). The SNP effects obtained from WOMBAT were tested using a Wald test and then the associated $p$-values were supplied to the *q-value* package (Storey and Tibshirani, 2003) in R to account for multiple testing and to generate the corresponding $q$ values (i.e., the SNP false discovery rate, FDR) and FDR thresholds.

## Detection of Selection Signatures

We used two complementary haplotype-based methods to scan the genome of the admixed cattle for candidate regions under selection. The integrated haplotype score (iHS) is an intra-population measure of the extent of haplotype homozygosity (Voight et al., 2006), and the Rsb test compares haplotype homozygosity length between populations (Tang et al., 2007).

### iHS

The iHS values were calculated within each chromosome of admixed genome according to Voight et al. (2006) using the *rehh* package (Gautier et al., 2017) for R software. At each locus with an MAF >0.05, we calculated the integrated extended haplotype homozygosity for the ancestral ($\mathrm{iHH_a}$) and the derived ($\mathrm{iHH_d}$) alleles, and then, the iHS was calculated as $iHS = \ln(\frac{iHH_a}{iHH_d})$. The iHH was defined as the area under the extended haplotype homozygosity (EHH) curve at a core allele within a chromosome using a homozygosity decay threshold of 0.05. The EHH for each core allele was calculated based on Sabeti et al. (2002) as

$$EHH_{a_s,t} = \frac{1}{n_{a_s}(1-n_{a_s})}\sum_{k=1}^{K_{a_s,t}} n_k(n_k-1),$$

where $K_{a_s,t}$ is the number of distinct haplotypes from the core SNP $s$ to SNP $t$ carrying the core allele $a_s$, $n_k$ is the number of times the $k$th haplotype is observed, and $n_{a_s}$ is the total number of haplotypes carrying $a_s$ and is calculated as $\sum_{k=1}^{K_{a_s,t}} n_k$. The iHS values were standardized to have a mean of 0 and a standard deviation of 1 according to the allele frequency bins to which they belonged. The frequency bins were determined by varying the frequency of the derived allele with a step of size 0.025. Then, the iHS values were transformed into $p$-values of "no selection" hypothesis according to Gautier and Naves (2011):

$$p_{iHS} = -\log[1 - 2|\Phi(iHS) - 0.5|],$$

where $\Phi$(iHS) represents the Gaussian cumulative distribution function of iHS values. To define the ancestral allele for each locus, we calculated allele frequencies in the entire data set and assigned the most common allele as the ancestral allele.

### Rsb

The Rsb values between admixed and EASZ cattle populations were calculated within each chromosome according to Tang et al. (2007) using the R software *rehh* package (Gautier et al.,

2017). The site-specific extended haplotype homozygosity was calculated for admixed and EASZ cattle populations separately:

$$EHHS_{s,t} = \frac{1-h_{s,t}}{1-h_s},$$
$$h_{s,t} = \frac{n_s}{n_s-1}\left(1-\frac{1}{n_s^2}\sum_{k=1}^{K_{a_s,t}} n_{a_s}^2\right), \text{ and}$$
$$h_s = \frac{n_s}{n_s-1}\left(1-\frac{1}{n_s^2}\sum_{a_s=1}^{2} n_{a_s}^2\right),$$

where $n_s$ is the total number of haplotypes carrying $a_s$ and is calculated as $\sum_{a_s=1}^{2} n_{a_s}$ for ancestral ($a_s = 1$) and derived ($a_s = 2$) alleles, and $K_{a_s,t}$ is the number of distinct haplotypes from the core SNP $s$ to SNP $t$ carrying the core allele $a_s$. The iES was defined as the area under the EHHS curve at a core allele within a chromosome using a homozygosity decay threshold of 0.05. The Rsb score between admixed and EASZ cattle populations was defined as $Rsb = \ln\left(\frac{iES_{admixed}}{iES_{EASZ}}\right)$ for each focal SNP and then standardized as

$$Rsb(s) = \frac{Rsb - med_{Rsb}}{\sigma_{Rsb}},$$

where $med_{Rsb}$ and $\sigma_{Rsb}$ are the median and standard deviation of Rsb across all SNPs within genome. The $p$-values for $\text{Rsb}_{(s)}$ were calculated according to Gautier and Naves (2011):

$$p_{Rsb_{(s)}} = -\log\left[1 - 2\left|\Phi\left(Rsb\right) - 0.5\right|\right],$$

where $\Phi(Rsb_{(s)})$ represents the Gaussian cumulative distribution function of $Rsb_{(s)}$ values. The *qvalue* package (Storey and Tibshirani, 2003) in R software was used to correct $p$-values for multiple testing in iHS and $Rsb_{(s)}$ by generating the corresponding $q$-values and FDR thresholds.

We calculated measures of selection signatures in two scenarios. In the first scenario, all the admixed samples were used to obtain estimates of iHS. In the second scenario, admixed cattle with less than three crossovers were removed prior to iHS and Rsb analyses because they were deemed to be recently admixed individuals in which selection has not had enough time to leave a signature on their genome.

## Annotation and Tracking of Candidate Regions

A candidate region detected by the SoS analyses was defined by first identifying SNPs with a $q$ value $< 0.1$ and then searching within the 500-Kbp interval downstream and upstream (1 Mbp window) of the identified SNP for SNPs with $q$ value $< 0.5$ and $q$ value $< 0.25$ for iHS and Rsb analyses, respectively. We extended the detected region (with a 500-Kbp step size) until there was no SNP with a $q$ value less than the suggestive thresholds within the 500-Kbp interval from the last identified SNP. The boundaries of the candidate region were determined based on the base pair positions of the last-identified SNP in each direction. The same procedure was used for iHS and Rsb analyses. Where GWA results were used for partial validation of SoS analyses (see below), we used a suggestive $p$-value threshold of $10^{-3}$ to define the candidate regions from GWA, and to define the boundaries of each candidate region, we searched the 500-Kbp upstream and downstream intervals for SNPs whose $p$-values were smaller than $10^{-3}$ and extended the region until there was no SNP $p$-value less than our suggestive threshold. The candidate regions designated by iHS and Rsb analyses were then annotated using the Ensemble Biomart 94 based on the UMD v3.1 bovine genome assembly for the underlying genes, and the biological functions of the discovered genes were evaluated and compared to the existing literature. We also calculated the ancestral allele dosages for the identified candidate regions in order to track the candidate regions under selection to each of the ancestral populations described above. In an attempt to validate SoS regions in the admixed cattle, we looked for overlap between the candidate regions identified in each of the SoS analyses, i.e., iHS or Rsb, and those identified by GWA.

The QTL and SNP association data mapped on the UMD3.1 bovine reference genome were obtained from the cattle QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index) on July 8, 2019, and was used to compare the results of the present study with the reported QTL regions in the literature. We compared the genes within our identified candidate regions for selection from iHS and Rsb analyses to the whole bovine genome background using functional annotation clustering by DAVID online bioinformatics resource v 6.8 (Huang et al., 2009) to find the pathways that are significantly overrepresented.

# RESULTS

## Genetic Structure of Admixed and Ancestral Cattle Populations

The PCA revealed a complex population structure for the admixed cattle in relation to their ancestral breeds (**Figure 1**). The Kenyan admixed cattle was found to be an unstabilized population with very high genetic diversity. Samples ranged along the axis from pure exotic dairy breeds through to pure indigenous EASZ. The EASZ animals formed a tight cluster on the axis between IND and AFT reference samples consistent with EASZ being an old or ancient admixture of IND and AFT ancestors that has a higher proportion of IND than AFT. The three ancestral breeds, i.e., EUT, AFT, and IND, were separated by the first PC explaining around 90% of the total variation between all genotypes. The second PC only explained around 1.6% of the variation and separated AFT from EUT. The locations of crossbred animals in **Figure 1** suggest that most animals were of Ayrshire, Holstein, and/or British Friesian ancestry with little contribution from Jersey and Guernsey, consistent with the previous findings of Strucken et al. (2017).

## Local Ancestry of Admixed Cattle

The ancestral haplotypes from the three groups (i.e., IND, AFT, and EUT) were used to infer the local ancestries of the admixed cattle at the individual loci level. The majority of haplotypes in the admixed cattle were found to be originated from EUT ancestor (≈0.73), and IND and AFT ancestral populations

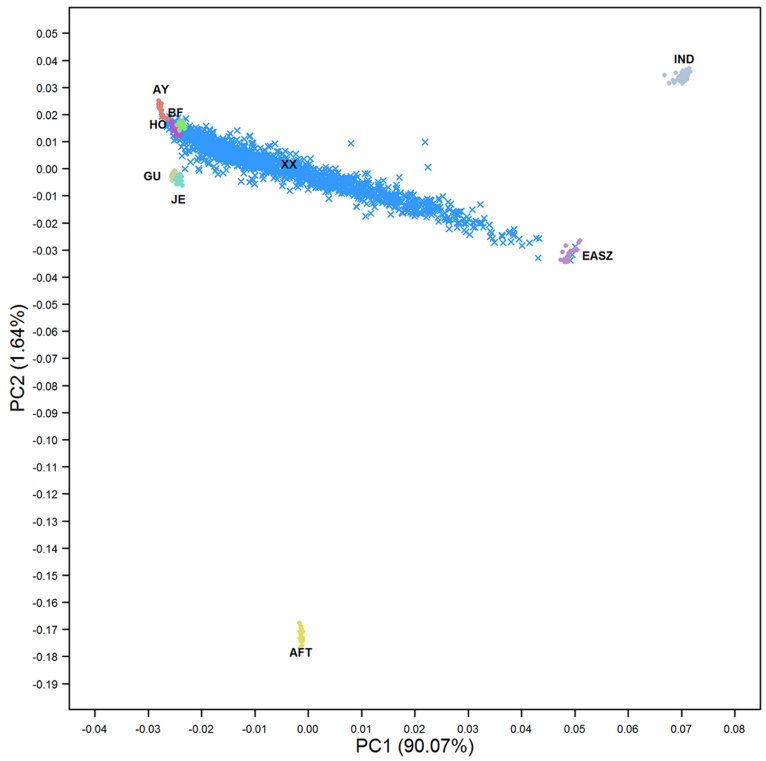


**FIGURE 1 |** The first two principal components showing the distribution of admixed cattle in relation to their ancestral breeds. IND, *Bos indicus*; AFT, African *Bos taurus*; AY, Ayrshire; HO, Holstein; BF, British Friesian; GU, Guernsey; JE, Jersey; EASZ, East African Shorthorn Zebu; and XX, Admixed cattle.

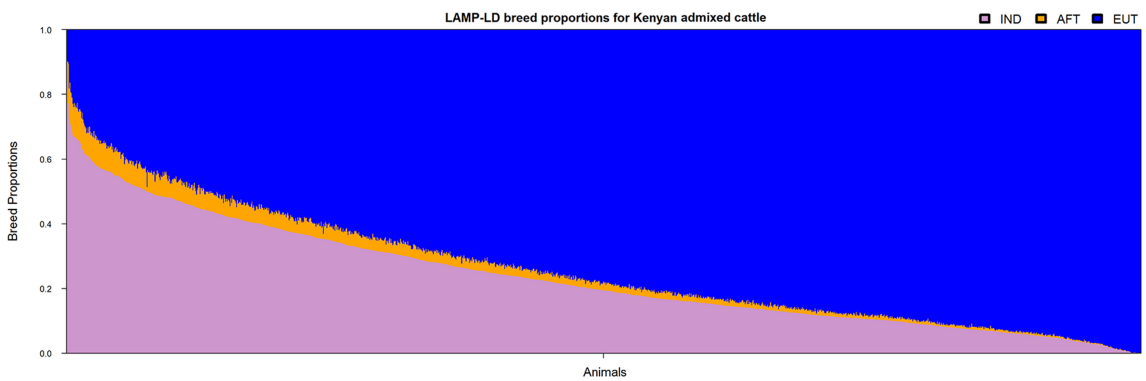


**FIGURE 2 |** The genome-wide average ancestries of the admixed cattle contributed by the three ancestral groups. IND, *Bos indicus*; AFT, African *Bos taurus*; and EUT, European *Bos taurus*.

contributed smaller proportions of admixed haplotypes (≈0.24 and ≈0.03, respectively). The genome-wide average ancestries of the three ancestral populations for each crossbred animal are shown in **Figure 2**. This confirms the wide range of admixture inferred from **Figure 1**. The distribution of local ancestries across different chromosomes of the admixed cattle (**Figure S1**) were, in general, agreement with genome-wide average ancestries showing that the admixture was relatively uniform across all chromosomes.

The distribution of number of recent crossovers on haplotypes with the lowest number of crossovers in different chromosomes is shown in **Figure S2**, and the corresponding distribution of genome average number of crossovers is shown in **Figure 3A**. For the majority of the admixed cattle, the number of recent crossovers was calculated to be small (<2 per Morgan) on almost all chromosomes. Only 55 animals passed a threshold of three or more crossovers per Morgan. The distribution of the number of recent crossovers on haplotypes carrying the highest number

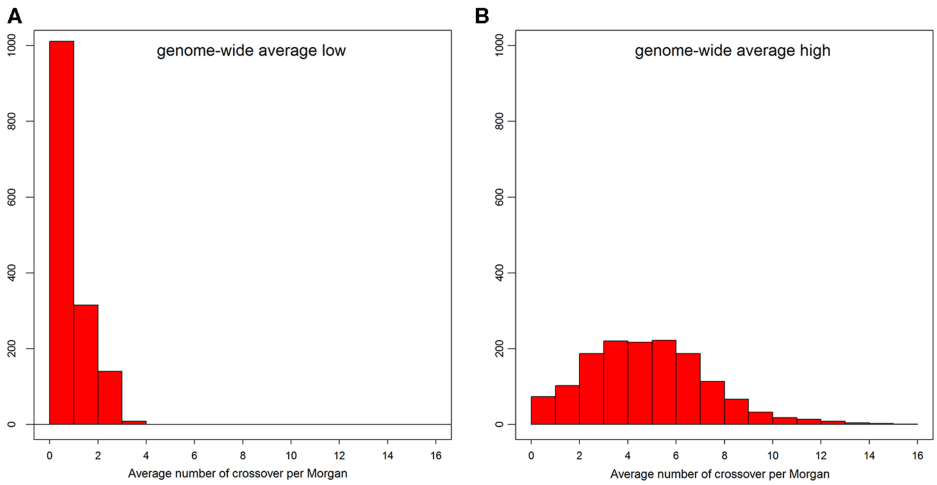


**FIGURE 3 |** The distribution of genome-wide average number of crossovers per Morgan on the admixed cattle haplotypes carrying the lowest **(A)** and highest **(B)** number of crossovers.

of crossovers across different chromosomes and the distribution of corresponding genome average number of crossovers are shown in **Figure S3** and **Figure 3B**, respectively. The animals with very low numbers of crossovers (<2 per Morgan) in **Figure 3B** are predominantly animals with high EUT ancestral proportion, in which most of the genome is homozygous EUT. However, most haplotypes presented a high number of recent crossovers (**Figure 3B**) with some individual chromosome haplotypes showing more than 20 crossovers (**Figure S3**).

The distribution patterns for the average local ancestries of admixed cattle with three or more recent crossovers per Morgan in haplotypes carrying the lowest number of crossovers are shown in **Figure 4**. The average contributions (calculated as average breed proportions) from IND + AFT (i.e., indigenous) vs. EUT ancestors were 0.52 and 0.48, respectively. This reflects that the ability to detect recombination events is highest in animals with ∼50% EUT vs. indigenous ancestry because, in animals with a high proportion of either indigenous or EUT ancestry, most historical crossover events occur within the dominant ancestral genome and, thus, are not detectable.

## Genome-Wide Associations for SNP Allele and Ancestral Origin of SNP Allele

The Manhattan plots of SNP allele effects and ancestral origin effects for MYD are presented in **Figures 5A,B**, respectively. No SNP passed an FDR threshold of <0.1 for these effects. For SNP allele effects, six SNP had the minimum observed FDR of 0.112 although, for ancestral origin effects, 518 SNP had the minimum observed FDR of 0.229. With an FDR threshold of <0.35, a total of 35 and 918 SNP passed the threshold for SNP allele effects and ancestral origin effects, respectively. The distribution of the estimated effects of SNP alleles and ancestral origin with a FDR < 0.35 are shown in **Figures S4A,B**, respectively. The estimated effects of SNP alleles on milk yield (**Figure S4A**) were approximately equally distributed on either side of zero as expected in GWA when the allele assignment is random. The estimated effects of ancestral origin in **Figure S4B** were predominantly positive, indicating that the alleles coming from the EUT ancestor had a positive effect on milk yield.

## Detection of Signals of Positive Selection

### Within Population

The Manhattan plots of *p*-values for genome-wide iHS scores calculated using all samples of admixed cattle as well as when using only the admixed cattle with three or more crossovers per Morgan on the chromosomes with lowest frequency of crossovers are given in **Figures 6A,B**, respectively. Although including all admixed cattle for calculation of iHS scores was not successful in detection of any candidate region at an FDR threshold of 0.1 (**Figure 6A**), removing admixed cattle with a genomic average crossover of less than three per Morgan identified 16 candidate regions across seven autosomes (**Figure 6B**). The size of these candidate regions ranged from only 112.25 Kbp on BTA 12 up to 0.68 Mbp on BTA 7 and together encompassed 106 genes. The details of the identified candidate regions from the iHS analysis of the filtered admixed cattle are in **Table 2**. BTA 7 had the highest number of candidate regions for selection (five regions), and BTA 3 contained 43 genes, which was the highest among all BTAs. Across all candidate regions, 10 genes were deemed as candidate genes for selection because there was at least 1 SNP with a FDR < 0.1 located within them. The ancestry of all candidate regions in BTA 3 was dominated by EUT, and for other chromosomes with more than one candidate region, the dominant ancestry was either IND or EUT.

### Between Populations

The distribution of *p*-values from Rsb analysis between the admixed cattle with a minimum number of three crossovers on the haplotype carrying the lowest number of crossovers across their genome and the EASZ population is shown in **Figure 7**. At FDR < 0.1, we identified 24 candidate regions for divergent selection between the admixed cattle and EASZ, indicating active

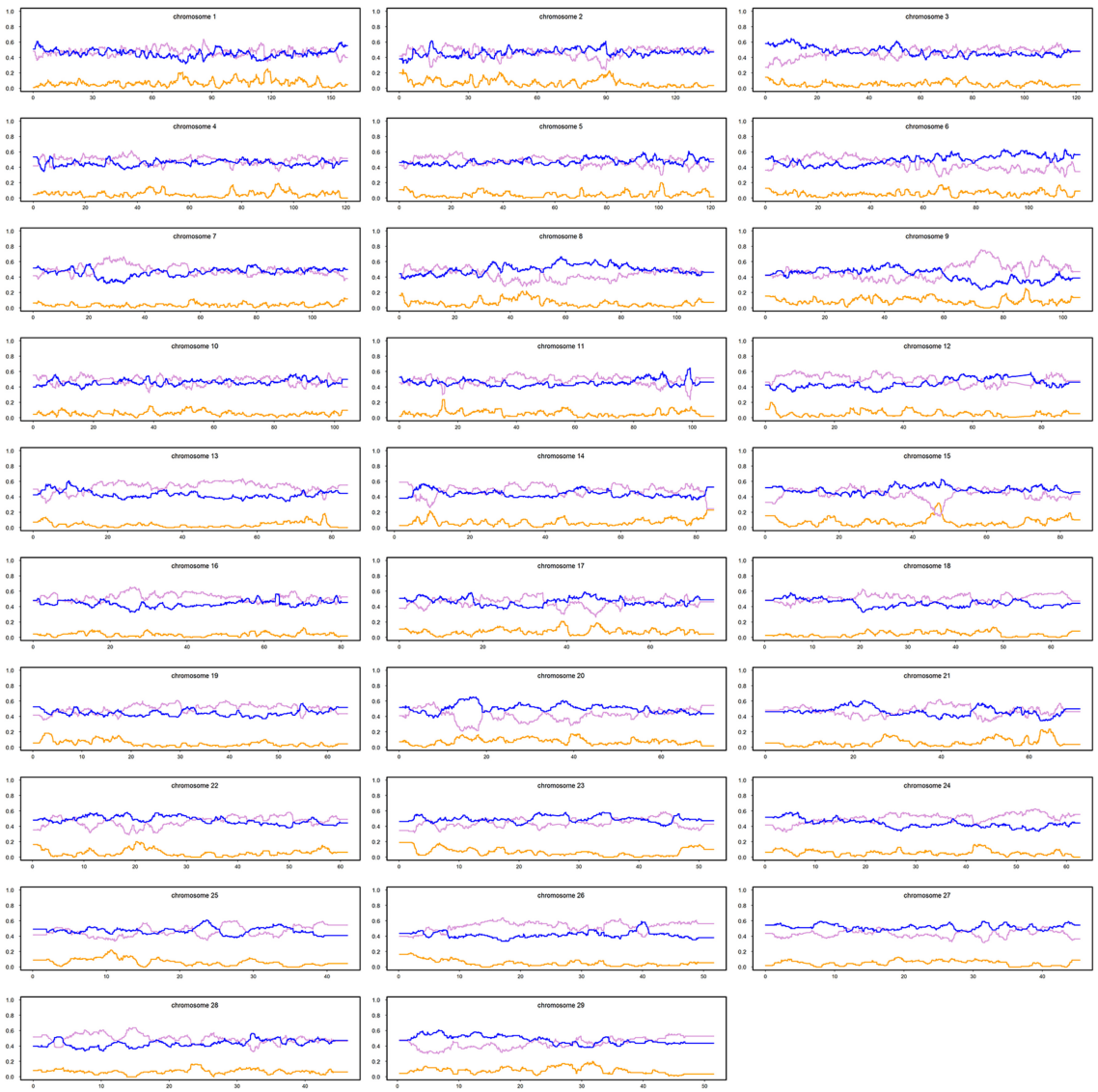


**FIGURE 4 |** Average local ancestries of the admixed cattle with three or more crossovers per Morgan in the haplotype carrying the lowest number of crossovers. The gray, yellow, and blue lines represent *Bos indicus*, African *Bos Taurus*, and European *Bos taurus* ancestry, respectively.

selection in the admixed population on 15 autosomes. These regions together harbored 15 candidate genes. BTA 23 contained the shortest candidate region with only 1 SNP, and the longest candidate region of 0.81 Mbp was found on BTA 6 (93 SNPs). The strongest selection signature with smallest SNP $q$ value and highest peak in the Manhattan plot of **Figure 7** was on BTA 3, followed closely by another candidate region on the same chromosome. The strongest candidate region for selection on BTA 3 also encompassed the highest number of candidate genes (a total of four) among all BTAs, whereas there was no candidate genes found in BTAs 2, 6, 9, 12, 13, 16, 21, and 29 (**Table 3**). The dominant ancestry of all candidate regions was EUT except four regions with IND ancestry on BTAs 3, 8, 21, and 29.

## Validation of Candidate Regions for Selection With GWA

Because the SoS showed lower FDR than the results from the GWA analysis, for the purpose of investigating possible candidate genes, we chose to cross-validate the SoS that passed FDR $<$ 0.1 with the GWA results. We used only the estimates of SNP allele effects because the confidence intervals for ancestral origin effects were very large. Four candidate regions from GWA, on BTAs 1, 7, and 20, overlapped with four candidate regions for selection obtained from iHS and Rsb analyses (shown in red boxes in **Figure 5A**). A candidate region for GWA on BTA 7 spanning from 44.12 to 44.96 Mbp covered around 0.04 Mbp of a selection signature discovered from iHS analysis (**Table 2**). In addition, two candidate regions for selection identified by

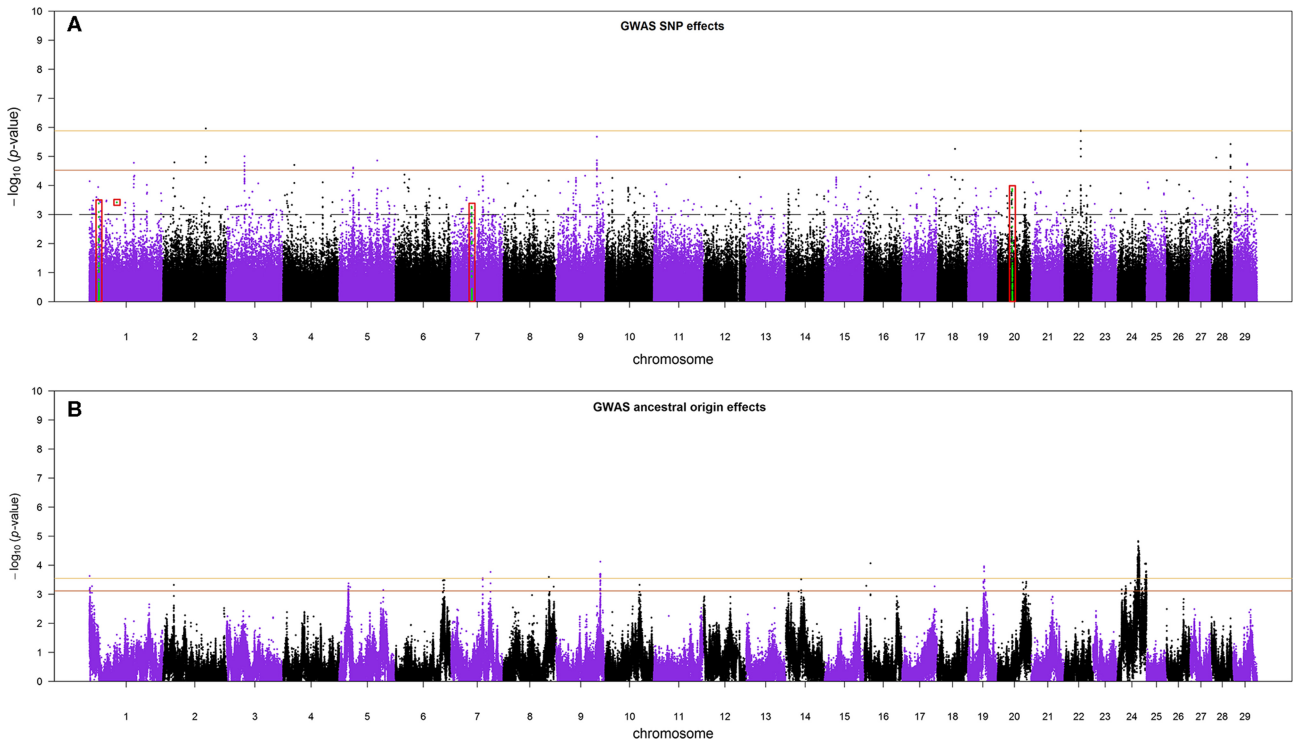


**FIGURE 5 | (A)** The Manhattan plot of *p*-values for SNP allele effects and **(B)** the Manhattan plot of *p*-values for ancestral origin effects. The red boxes in **(A)** are the candidate regions for selection signatures that overlap with GWA regions. The colored horizontal lines are false discovery rate thresholds at 0.112 [$-\log_{10}$(*p*-value) = 5.88] and 0.35 [$-\log_{10}$(*p*-value) = 4.53] in **(A)**, and at 0.229 [$-\log_{10}$(*p*-value) = 3.55] and 0.35 [$-\log_{10}$(*p*-value) = 3.11] in **(B)**, from top to bottom, respectively. The dashed line in **(A)** is the suggestive *p*-value threshold of $10^{-3}$.

Rsb on BTA 1 and distributed from 19.76 to 10.27 Mbp and from 58.74 to 59.22 Mbp overlapped with a candidate regions for GWA spanned from 20.09 to 20.60 Mbp and 1 SNP on 58.96 Mbp, respectively. Another candidate region identified by Rsb on BTA 20 also intersected with a candidate region from GWA that covered between 31.32 and 31.87 Mbp of the chromosome (**Table 3**).

## Functional Characterization of Candidate Regions for Selection

A total of 106 genes from iHS method (**Table S1**) are grouped into 13 annotation clusters, of which five are significantly enriched (enrichment score > 1.3 in **Table S2**). The enriched annotation terms from iHS analysis are associated with different biological functions, namely olfactory receptor activity, potassium ion transport, immunoglobulin molecules structure, SPRY domain, and innate immunity. The 119 genes within the candidate regions detected by Rsb analysis (**Table S1**) are categorized into 12 annotation clusters, of which two clusters are significantly enriched (**Table S2**). The significantly enriched annotation clusters from Rsb are involved in potassium ion transport and ephrin receptor signaling pathway.

The 16 and 24 candidate regions for selection identified by iHS and Rsb intersect with 208 and 373 QTLs or associations for different traits among which are reproduction, health, conformation, and meat and milk traits (**Table S3**).

# DISCUSSION

The distribution of admixed individuals in relation to the purebred ancestral breeds and the estimated ancestral breed proportions of the admixed cattle (**Figures 1**, **2**) confirms the previously reported findings that the Kenyan crossbred dairy cattle form an unstabilized and highly diverse admixture of local indigenous cattle and exotic dairy breeds (Strucken et al., 2017). It has been shown in the same population that it is important to take the variation in breed composition into account when undertaking genetic evaluations of admixed individuals (Ojango et al., 2019).

The method of assigning ancestry of admixed cattle at individual loci using haplotypes from three reference breed groups (i.e., IND, AFT, and EUT) appeared to work very well, yielding similar levels of indigenous vs. exotic admixture to previous Admixture analyses (Alexander et al., 2009) of the same population (Weerasinghe, 2014; Strucken et al., 2017). The number of available samples for AFT was limited, and fewer samples were used compared to the other two ancestral populations. This might have led to the observed underestimation of AFT relative to IND when compared to

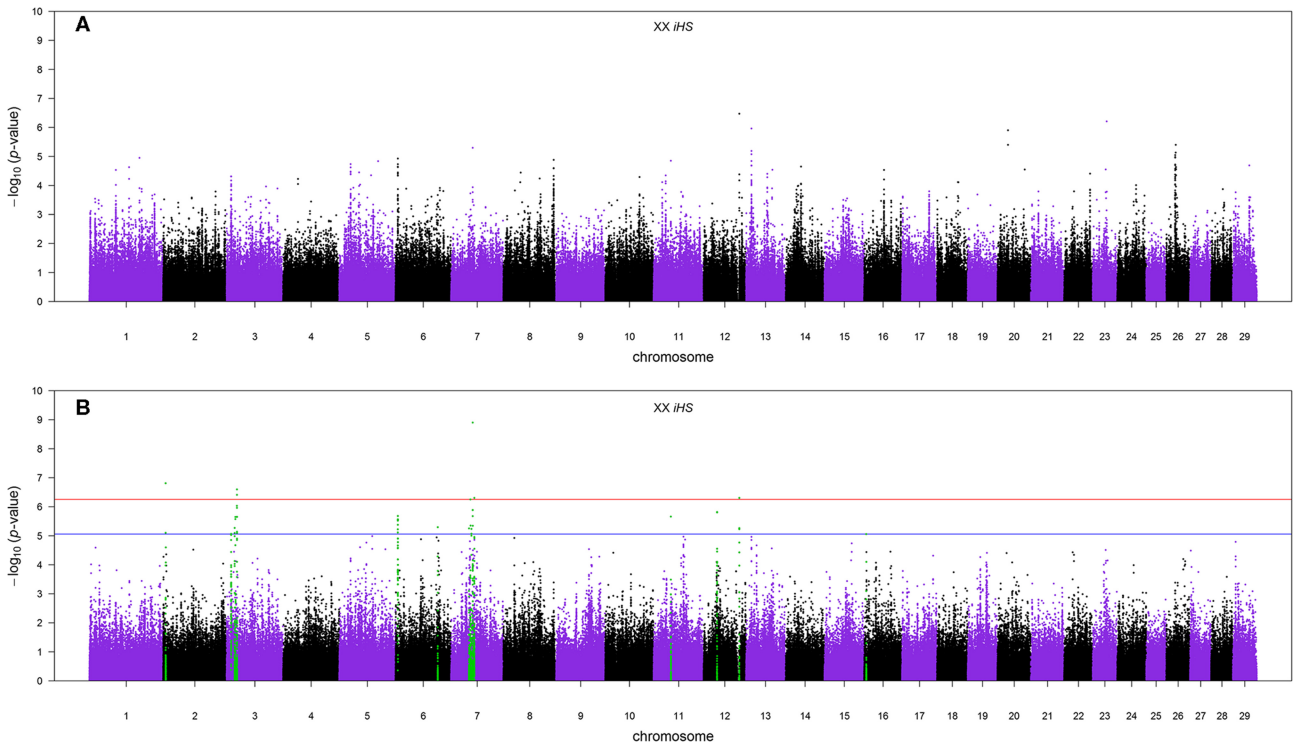


**FIGURE 6 | (A)** The Manhattan plot of *p*-values for genome-wide iHS scores calculated using all samples and **(B)** the Manhattan plot of *p*-values for genome-wide iHS scores calculated using only admixed cattle with three or more crossovers per Morgan. The red and blue horizontal lines are false discovery rate thresholds at 0.05 [$-\log_{10}$(*p*-value) = 6.25] and 0.10 [$-\log_{10}$(*p*-value) = 5.06], respectively. Green points are the SNPs within the candidate regions identified as being under selection.

whole-genome admixture analyses of the same population (Weerasinghe, 2014; Strucken et al., 2017). When we compared results of genome-wide admixture analyses (results not shown) to the genome-wide average of local ancestries from LAMP-LD, we observed a very high correlation (∼0.99) between all components of ancestry inference from the two analyses, notwithstanding that the estimated AFT ancestry was higher from admixture compared to LAMP-LD (0.08 vs. 0.03). Of the reduced AFT estimate, 0.03 appeared in the IND estimate and 0.02 in the EUT estimate. The small proportion of AFT ancestry that appears to have flowed into the EUT estimate may have caused a small error in the assignment of ancestral haplotypes and, hence, crossover events, in these analyses.

The Manhattan plot of the GWA analysis of SNP effects (**Figure 5A**) consisted of relatively sharp peaks that are typical of a within-population GWA. Thirteen peaks passed an FDR of 0.35, giving an expectation that 65% (i.e., approximately eight) of these peaks are real effects. The Manhattan plot of the GWA analysis of ancestral origin effects (**Figure 5B**) consisted of very broad peaks. This is expected because mapping ancestral origin effects is analogous to QTL mapping in crosses between inbred lines, where the confidence interval for location of a QTL effect is very large in early-generation crosses and reduces as the number of recombination events between ancestral haplotypes increases with increasing number of generations of inter se crossing (Lynch and Walsh, 1998). The situation in this crossbred cattle population is more complicated than inter se mating in populations created from inbred lines because the low frequency use of AI and the wide variation in breed compositions cause the number of recombination events on a given chromosome copy to vary from very few for recent crosses to purebred or high-grade animals to very many for chromosomes resulting from many generations of inter se matings.

Depending on what is deemed to be a single peak vs. multiple peaks, at FDR of 0.35, between 15 and 18 peaks for ancestral effects were detected with an expectation that 65% (i.e., 10 to 11) are real effects. The distribution of ancestral origin effects (**Figure S4B**) showed that the vast majority of positive effects on milk yield came from the exotic dairy breed ancestors. These estimates should be independent of effects of breed composition across the whole genome because the data had been pre-corrected for breed composition classes, and the statistical model used here included a GRM to account for whole genome relationships, which would also account for any residual additive effects on breed composition. The present results, therefore, indicate that there are many genomic regions that determine the high genetic milk potential of modern dairy breeds and that no one region carries an exceptionally large effect. The estimates of ancestral origin effects are allele-substitution effects so that the estimates of homozygous exotic

**TABLE 2 |** Candidate regions for selection obtained from iHS analyses in admixed cattle.

| BTA | Region (Mbp) | Top SNP q-value | Dominant ancestry* | Candidate genes |
|---|---|---|---|---|
| 2 | 5.46–6.00 | 0.0378 | IND | – |
| 3 | 9.58–9.80 | 0.0995 | EUT | – |
| 3 | 17.18–17.70 | 0.0861 | EUT | – |
| 3 | 18.80–19.29 | 0.0578 | EUT | *S100A10* |
| 3 | 22.07–22.71 | 0.0390 | EUT | *ACP6, RF00100* |
| 6 | 4.91–5.29 | 0.0578 | IND | – |
| 6 | 90.70–91.12 | 0.0861 | EUT | *MTHFD2L* |
| 7 | 38.55–38.92 | 0.0861 | IND | – |
| 7 | 41.40–42.00 | 0.0390 | IND | *BTNL9, NLRP3* |
| **7** | **43.84–44.16** | **0.0861** | **EUT** | ***LYPD8*** |
| 7 | 46.56–46.99 | 0.0006 | EUT | – |
| 7 | 49.91–50.25 | 0.0390 | IND | – |
| 11 | 36.81–37.13 | 0.0578 | IND | *ACYP2, ENSBTAG00000046563* |
| 12 | 28.64–29.05 | 0.0578 | IND | – |
| 12 | 76.82–76.93 | 0.0390 | EUT | *CLDN10* |
| 16 | 4.52–4.89 | 0.0995 | IND | – |

**IND, Bos indicus; AFT, African Bos taurus; and EUT, European Bos taurus. Bold regions overlap with regions identified in the genome-wide association analysis.*

dairy vs. indigenous effects are mostly between 0.44 and 0.56 kg milk per day. The average yield in this crossbred population, which has a breed composition average of about 70% exotic dairy, has been estimated around 5 kg milk per day (Ojango et al., 2019). The milk yield of indigenous cattle is not known but can reasonably be expected to be about 2 kg per day. Although the estimates of ancestral genomic effects are subject to ascertainment bias and need to be independently validated, it is possible that, collectively, they could explain much of the difference between exotic vs. indigenous cows in the smallholder production environment.

The distribution of estimated SNP effects (**Figure S4A**) shows the expected equal allocation of positive and negative SNP effects but has a proportion of estimates substantially higher in magnitude than those for ancestral origin effects. This likely reflects that the power of estimating ancestral origin effects is essentially uniform across the genome while that of estimating SNP effects is highly dependent on the allele frequency of each SNP so that some SNPs will be subject to substantially higher ascertainment bias than other SNPs (Lynch and Walsh, 1998).

Work is currently underway to phenotype and genotype much larger populations of crossbred dairy cows in East Africa. This will provide a future opportunity to validate the GWA results presented here. If some of these results are validated, it will be possible to identify groups of SNPs that track genomic region effects due either to within-population LD with causal variants or to ancestral genomic effects. In the latter case, SNPs can be chosen to provide a high accuracy of assigning ancestral haplotype allocation to the relatively large confidence regions encompassed by the ancestral QTL effects.

When mapping signatures of selection, the null hypothesis of "no selection" typically implies a lack of statistical significance in situations where there is no disturbance from common demographic factors. Therefore, the ability to clearly distinguish positive selection from neutral effects is a challenge given the high sensitivity of the test statistics for detection of selection (Tang et al., 2007). In this study, we employed a method based on the decay of ancestral haplotypes to remove the impact of recent admixture and continuous gene flow on detection of selection signatures in Kenyan admixed cattle. Our method relies on the fact that first-generation admixed individuals inherit two intact haplotypes, one from each inputting founder populations, and as mating happens within the admixed population, recombination mixes these haplotypes and creates a mosaic genome in subsequent generations. We measure the degree of fragmentation of ancestral haplotypes according to the distribution of crossover events across the admixed genome. We consider a shift in local ancestry of haplotypes carried by an admixed individual as a recombination event where individuals from later generations are expected to express higher numbers of recombination events generating a more fragmented genome. Since the ancestral populations of admixed cattle are highly diverged and show significantly different allele frequencies, it is possible to assign the ancestry to each allele of an admixed individual with high confidence. This was tested in a cross-validation approach for the local ancestry mapping of only ancestral breeds, and it was found that LAMP-LD was able to assign the ancestry origin of haplotypes with very high accuracies (results are not shown).

Continuous admixture and gene flow can leave different patterns of ancestry in the two haplotypes carried by an admixed individual. Backcrossing to pure parents will produce progenies with one copy of the mosaic genome and a copy of intact chromosomes inherited from pure parents. We found evidence for such patterns in our results when we observed very different distributions for the number of recent crossovers across the two haplotypes of admixed cattle (**Figures S2, S3**). One of the admixed haplotypes showed less than one crossover for the majority of individuals (**Figure S2**). This suggested that the majority of admixed cattle in Kenya have at least one ancestor that resulted from a recent cross with either an indigenous or an exotic breed. The other copy of the admixed haplotype showed higher number of crossovers (**Figure S3**) with an average of around five (**Figure 3B**). This provided additional evidence for the high rate of recent introgression of an exotic breed genotype in the region and recurrent admixture between them and the existing admixed cattle. Given this, we rank the two haplotypes of admixed cattle across different chromosomes based on the number of recent crossovers they incur and use the haplotype carrying the lowest number of recombination to quantify the degree of fragmentation of ancestral segments in the sampled genome and to measure the age of admixture in our samples.

Our results showed that the iHS analysis didn't detect any candidate region for positive selection at an FDR threshold of 0.1 when all admixed samples were included (**Figure 6A**). Using an empirical threshold of at least three for the genome-wide average

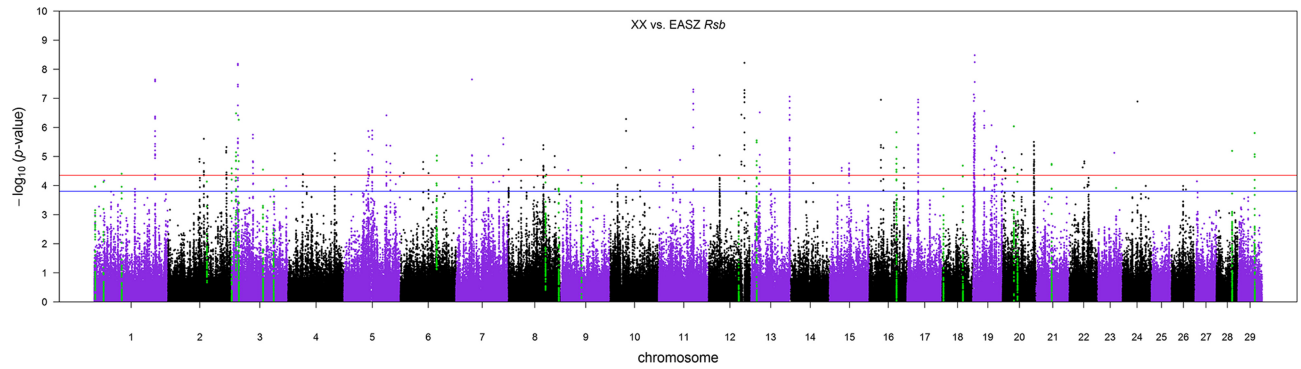


**FIGURE 7 |** The Manhattan plot of *p*-values for Rsb analysis between the admixed cattle with a minimum number of three crossovers per Morgan and the East African Shorthorn Zebu population. The red and blue horizontal lines are false discovery rate thresholds at 0.05 [$-\log_{10}$(*p*-value) = 4.35] and 0.10 [$-\log_{10}$(*p*-value) = 3.80], respectively. Green points are the SNPs within the candidate regions identified as being under selection in the crossbred population.

number of recent crossovers per Morgan in haplotypes carrying the lowest number of recombinations improved the detection of signatures of selection by making the signals stronger. When we excluded samples with less than three crossovers per Morgan, the iHS method was successful in detecting 16 candidate regions at the same FDR threshold (**Figure 6B**). Excluding individuals with some recent admixture from the analysis ensures that the sample analyzed has had sufficient time for selection to act to produce detectable signatures, thus increasing the power of the analysis. However, imposing more stringent thresholds greatly reduces the number of animals available, leading to a subsequent decrease in power. In such studies, there will be a threshold for data selection that optimizes power, and that threshold will be dependent on the size, structure, and history of the population.

## Cross-Validation of SoS With GWA

We employed a low-stringency criterion to define regions from GWA that might overlap with SoS, and this resulted in four overlapping regions. There was no overlap between the GWA and SoS that are deemed to be significant, and the low-stringency threshold we used for GWA regions in the cross-validation would implicate a substantial proportion of the genome being involved in genetic variation in milk yield. Thus, having just four regions overlapping between GWA and SoS provides no more than suggestive evidence that the same regions are involved.

The SoS and GWA regions are expected to overlap where regions controlling genetic variation in milk yield have been under selection and already yielded SoS while still segregating in the population and, hence, detectable in GWA analysis. In relatively young populations, it is likely that regions under selection are still segregating and, hence, detectable as SoS and GWA, but SoS are expected to result from selection on many traits other than milk yield, and so even with large data sets and very high power, only a proportion of SoS and GWA regions are expected to overlap. Given the modest statistical power of the current data set there could be many regions that do overlap but are not detected in either or both of the SoS and GWAS analyses.

## Functional Characterization of Candidate Regions for Selection

In the context of localizing the identified candidate regions under selection in Kenyan admixed cattle, we classify them into two groups with related functions in (1) productivity and (2) adaptation, recognizing that some regions might have pleiotropic effects in both categories. In the following, we characterize the functions of our identified regions in more detail.

### Productivity

Several candidate regions from iHS and Rsb analyses intersected with previously reported QTLs and associations for milk and meat production traits in the literature. Milk and milk composition encompassed the highest number of overlaps among all traits for both methods. Given that there is no genetic improvement program for milk yield in the population of smallholder cows analyzed here, this might be due to phenotypic selection by farmers who preferentially keep progeny from their best yielding cows. However, it should be noted that the milk production under these poor-quality environments relies on other factors, such as the ability of cows to achieve acceptable growth and reproductive performance with restricted feed and in the presence of disease pathogens.

Bovine chromosome 20 has been associated with several milk traits in dairy cattle (e.g., Nayeri et al., 2016). Our Rsb analysis identified two regions of selection signature on this chromosome by contrasting haplotype diversity between admixed and EASZ cattle. The region spanning from 31.68 to 32.17 Mbp overlaps with the growth hormone receptor (*GHR*) gene that has been proved to play a central role in variation of milk production in dairy cattle (Georges et al., 1995; Blott et al., 2003; Viitala et al., 2006). The findings of several genome-wide association studies (e.g., Pryce et al., 2010; Iso-Touru et al., 2016) as well as a genome scan for selection signatures in dairy cattle (Flori et al., 2009) strongly support the important function of *GHR* gene for milk traits. Both selection signatures on BTA 20 show an EUT ancestry, which supports the role of selection in favoring the EUT haplotypes.

**TABLE 3 |** Candidate regions for selection obtained from RSB analyses between the admixed and East African Shorthorn Zebu cattle populations.

| BTA | Region (Mbp) | Top SNP *q-value* | Dominant ancestry* | Candidate genes |
|---|---|---|---|---|
| 1 | 1.67–2.16 | 0.0805 | EUT | *ENSBTAG00000047288* |
| **1** | **19.76–20.27** | **0.0665** | **EUT** | **–** |
| **1** | **58.74–59.22** | **0.0463** | **EUT** | ***SIDT1*** |
| 2 | 83.56–84.07 | 0.0657 | EUT | – |
| 3 | 0.26–0.74 | 0.0369 | EUT | *TBX19* |
| 3 | 9.45–9.76 | 0.0033 | EUT | *COPA, PEX19, ATP1A2, KCNJ10* |
| 3 | 15.38–15.96 | 0.0037 | EUT | *GBA, MTX1* |
| 3 | 67.54–68.09 | 0.0387 | EUT | *AK5* |
| 3 | 90.34–90.83 | 0.0921 | IND | – |
| 6 | 77.36–78.17 | 0.0204 | EUT | – |
| 8 | 79.64–79.99 | 0.0514 | EUT | *NTRK2* |
| 8 | 108.09–108.64 | 0.0880 | IND | – |
| 9 | 43.55–43.95 | 0.0521 | EUT | – |
| 12 | 64.41–64.92 | 0.0556 | EUT | – |
| 13 | 11.76–12.27 | 0.0103 | EUT | – |
| 16 | 58.34–58.77 | 0.0070 | EUT | – |
| 18 | 2.69–2.99 | 0.0881 | EUT | *CFDP1* |
| 18 | 44.29–44.78 | 0.0324 | EUT | – |
| 20 | 23.95–24.20 | 0.0053 | EUT | *CDC20B* |
| **20** | **31.68–32.17** | **0.0472** | **EUT** | **–** |
| 21 | 33.23–33.66 | 0.0297 | IND | – |
| 23 | 39.00–39.00 | 0.0859 | EUT | *RNF144B* |
| 28 | 33.40–33.87 | 0.0160 | EUT | *KCNMA1* |
| 29 | 35.67–36.19 | 0.0072 | IND | – |

**IND, Bos indicus; AFT, African Bos taurus; and EUT, European Bos taurus, Bold regions overlap with regions identified in the genome-wide association analysis.*

Chromosome 6 plays a major role in determining protein composition of cow milk (Martin et al., 2002). The casein cluster of four tightly linked genes mapped on BTA 6 at around 87 Mbp is close to an iHS candidate region (90.70–91.12 Mbp) and lies within the scatter of points constituting the broader peak within which the candidate region sits. Buitenhuis et al. (2016) has reported several significant SNPs for variation in milk protein percentage of dairy cattle that lie within our candidate SoS region. These authors inferred their significant SNPs as possibly being in association with genes in the casein cluster. However, smallholder farmers have no information about the milk protein content, and there has been no genetic improvement program in this population, so it is unclear why polymorphisms controlling milk protein would have been under selection.

Four regions on BTA 7 identified by iHS, including the region overlapping with the GWA region spanning from 43.84 to 44.16 Mbp, have been associated with several milk traits in dairy cattle (Chamberlain et al., 2012; Marete et al., 2018). Some studies have also reported the same regions for various beef traits (Akanno et al., 2018).

The candidate regions on BTA 3, all from EUT ancestry, overlap with regions for meat-related traits (e.g., Seabury et al., 2017). The region spanning from 18.80 to 19.29 Mbp was found to have an effect on maternal weaning weight of Angus cattle (Saatchi et al., 2014). This region overlaps with several important genes involved in cell growth and proliferation (*OAZ3*), regulation of lipid metabolism (*THEM5*), and cell cycle progression and differentiation (*S100A10*) where the latter gene has also been reported as a candidate gene for residual feed intake in Angus (Al-Husseini et al., 2014).

A candidate region for selection with IND origin was mapped by iHS on BTA 2 extending from 5.46 to 6.00 Mbp. This region overlaps with the *HIBCH* gene, which is involved in amino acid metabolism in humans (Loupatty et al., 2007) and is in close proximity to bovine myostatin gene (*MSTN* at around 6.28 Mbp). *MSTN*, also known as growth and differentiation factor-8 (GDF-8), has an important role in muscle development in cattle (Sharma et al., 1999). Given that feed efficiency, muscle development, and growth are very important factors in low-input smallholder production systems, it is reasonable that these genes might have been the target of selection in the African environment.

## Adaptation

Genes with functional importance in immunity were identified on BTAs 7 (*SPOCK1*, *NLRP3*) and 21 (*CSPG4*). A candidate region on BTA 7 with a dominant IND ancestry extends from 41.40 to 42 Mbp and harbors the *NLRP3* gene. This gene encodes a protein that is involved in regulation of inflammation, immune response, and apoptosis. It is also a candidate gene for Crohn's disease (Villani et al., 2009) and Johne's disease (Scanu et al., 2007; Mallikarjunappa et al., 2018) in human and livestock populations, respectively. Other candidate regions originated from IND and associated with health traits of Kenyan admixed cattle were mapped on BTA 7 (49.91–50.25 Mbp) and 21 (33.23–33.66 Mbp) from iHS and Rsb analyses, respectively. The region on BTA 7 overlaps with a previously reported region for Mycobacterium paratuberculosis susceptibility in U.S. Holsteins (Settles et al., 2009) and encompasses the *SPOCK1* gene, which has been shown to be associated with cancer in humans (Miao et al., 2013). The region on BTA 21 has been associated with somatic cell score in Norwegian Red cattle (Sodeland et al., 2011) and contains the *CSPG4* gene, which is also linked to cancer in humans (Ilieva et al., 2017). Given that the selection sweeps harboring these genes are of IND ancestry, it is possible that the *Bos indicus* ancestors of admixed cattle may have contributed versions of genes conferring resistance to environmental disease challenges.

Evidence for EUT contribution to immunity of admixed cattle in Kenya were found on BTAs 7, 23, and 28. In a candidate region identified by iHS on chromosome 7 is the gene *LYPD8*, which has been reported to be differentially expressed between cows with vs. without subclinical mastitis (Song et al., 2016), and it provides defense against Gram-negative bacteria in the colon of non-ruminants. A candidate SNP on BTA 23 with EUT origin was found to be located in the *RNF144B* gene, which is involved in the innate immune system in humans (e.g., Ariffin et al., 2016). Further evidence for the functional importance of its surrounding region has been reported by Raphaka et al.

(2017) who found several nearby SNPs with large effects on two indicator traits for bovine tuberculosis susceptibility. Another candidate region on BTA 28 from Rsb analysis overlaps with the *POLR3A* gene, which provides instructions for making a protein that acts as a sensor to detect foreign DNA and trigger an innate immune response. The above regions are all of EUT origin, suggesting possible EUT contribution to disease resistance in the admixed population.

Heat stress can have adverse effects on reproductive performance of cattle (Folman et al., 1983). Therefore, the ability of animals to express enhanced reproduction under heat stress conditions can be deemed as an adaptive feature targeted by natural selection in the African environment. In the present study, we found several overlaps between our identified candidate regions for selection on BTAs 3, 7, 11, 12, 18, and 20 and genomic regions previously reported to affect reproduction in cattle. Chromosome 3 had the largest number of overlaps where four regions each from iHS and Rsb analyses intersected with several genomic segments from the literature. The iHS analysis identified a candidate region on this chromosome spanning from 18.80 to 19.29 Mbp. This region harbors several important genes (*TDRKH*, *OAZ3*, and *CELF3*) that are involved in spermatogenesis and early embryonic development in humans (Dasgupta and Ladd, 2012; Saxe et al., 2013) and mice (e.g., Tokuhiro et al., 2009). The same region also contains a significant peak in a large GWA on gestation lengths of U.S. Holsteins (Maltecca et al., 2011). Another region on the same chromosome (BTA 3; 9.45–9.76 Mbp) but identified by Rsb has been shown to be associated with a number of reproduction traits in Holstein cows (Cole et al., 2011). This region also covers the *IGSF8* gene, which produces a protein with the same name that has been shown to be essential in sperm-egg fusion in humans (Glazar and Evans, 2009). An iHS identified region of IND origin on BTA 7 (41.40–42.00 Mbp) overlaps with several regions reported for fertility-related traits from the literature, including genomic scans of tropical beef (Hawken et al., 2012) and Nelore (Irano et al., 2016) cattle. The iHS analysis also identified two regions of IND genetic background on BTAs 11 and 12 being important for reproduction traits of dairy cattle (Cole et al., 2011; Suchocki and Szyda, 2015; Parker Gaddis et al., 2016). The region on chromosome 12 (28.64–29.05) encompassed two genes that are especially active in ovaries (*BRCA1* and *ZAR1L*) and regulate some important functions for reproduction. These findings suggest an advantage for inheriting genes of IND origin for fertility under heat stress conditions.

The admixed cattle may have benefited from haplotypes descended from EUT ancestors on BTA 18. Chromosome 18 has been identified as an influential chromosome for fertility traits in dairy cows (e.g., Muller et al., 2017). We found two regions on this chromosome based on Rsb analysis both showing an EUT origin. The region spanning from 44.29 to 44.78 Mbp overlaps with previously reported regions for cow fertility (Parker Gaddis et al., 2016; Muller et al., 2017) and encompasses the *CHST8* gene. This gene, which is mainly expressed in the pituitary gland, encodes a protein that is involved in production of sex hormones.

## CONCLUSIONS

By explicitly mapping the regions that differentiate the exotic dairy from indigenous breeds, our GWA results, for the first time, indicate that the evolution of modern dairy breeds likely involved many genomic regions with no single region having an exceptional effect on milk production, at least under smallholder production conditions. Although clearly requiring to be validated, the results suggest that there are many regions involved in genetic variation within and between ancestral populations that might be used in genomic selection in future. The signatures of selection results provide evidence that the genome of Kenyan admixed dairy cattle has been shaped by adaptive selection in response to the low-input environment in which they exist. Exploration of genes in the candidate regions revealed a number of genes of possible functional importance. Our results also indicate that different ancestral backgrounds (indigenous vs. exotic breed genotypes) are advantageous in different regions of the genome. If confirmed, it may be possible to use beneficial haplotypes in genetic improvement of crossbred performance.

## DATA AVAILABILITY STATEMENT

The phenotypic and genotypic data for the study population were collected by the Dairy Genetics East Africa (DGEA) project. The genetic data were collected under host country agreements that anticipated Article 5 of Nagoya Protocol of the United Nations Convention on Biodiversity (CBD). The data underlying this study is archived at the International Livestock Research Institute (ILRI) and request to access the data set can be made to http://data.ilri.org/portal/dataset/dgea1-data-used-in-aliloo-et-al-2020. The reference genotypes used in this study can be accessed through direct requests to the respective data owners as indicated in the acknowledgments section.

## ETHICS STATEMENT

Ethical review and approval was not required for the animal study because in the current study we accessed to already produced genotypes and no live animals were used.

## AUTHOR CONTRIBUTIONS

AO and JG designed the data collection program, including the phenotypes used for GWA. HA and JG conceived and designed the current analyses. RM undertook the initial analyses and produced the corrected data used for GWA. HA conducted all other analyses and drafted the manuscript. JG assisted with the interpretation of results and edited the manuscript. All authors read and approved the manuscript before submission.

## FUNDING

This research was supported by Bill & Melinda Gates Foundation Grant No. OPP1130995 all funded by BMGF: OPP1071835 and OPPGD640.

## ACKNOWLEDGMENTS

The authors thank Illumina (Illumina, San Diego, CA) and Geneseek (Neogen Corporation, Lincoln, NE) for their kind contributions to genotyping costs. Special thanks to Ed Rege (PICO Eastern Africa, Nairobi, Kenya) who co-designed and helped leading the DGEA project, and to Julie Ojango, James Rao, Denis Mujibi, and Tadelle Dessie of International Livestock Research Institute (ILRI, Kenya and Ethiopia) who facilitated and undertook the field sampling that provided the data for the current research. The British Friesian genotype data was kindly provided by Scottish Rural University College (SRUC, Scotland), and the Ayrshire genotypes were kindly supplied by the Canadian Dairy Network (CDN, Canada). We also thank the smallholder farmers who participated in the DGEA project and provided samples and data on their animals.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00544/full#supplementary-material

## REFERENCES

Akanno, E. C., Chen, L., Abo-Ismail, M. K., Crowley, J. J., Wang, Z., Li, C., et al. (2018). Genome-wide association scan for heterotic quantitative trait loci in multi-breed and crossbred beef cattle. *Genet. Sel. Evol.* 50:48. doi: 10.1186/s12711-018-0405-y

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

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, 12–19. doi: 10.1111/age.12092

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, 9108–9127. doi: 10.3168/jds.2018-14621

Ariffin, J. K., Kapetanovic, R., Schaale, K., Gatica-Andrades, M., Blumenthal, A., Schroder, K., et al. (2016). The E3 ubiquitin ligase RNF144B is LPS-inducible in human, but not mouse, macrophages and promotes inducible IL-1β expression. *J. Leukocyte Biol.* 100, 155–161. doi: 10.1189/jlb.2AB0815-339R

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

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

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367. doi: 10.1093/bioinformatics/bts144

Blott, S., Kim, J. J., Moisio, S., Schmidt-Kuntzel, A., Cornet, A., Berzi, P., et al. (2003). Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* 163, 253–266.

Brown, A., Ojango, J., Gibson, J., Coffey, M., Okeyo, M., and Mrode, R. (2016). Short communication: genomic selection in a crossbred cattle population using data from the Dairy Genetics East Africa Project. *J. Dairy Sci.* 99, 7308–7312. doi: 10.3168/jds.2016-11083

Buitenhuis, B., Poulsen, N. A., Gebreyesus, G., and Larsen, L. B. (2016). Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. *BMC Genet.* 17:114. doi: 10.1186/s12863-016-0421-2

Chamberlain, A. J., Hayes, B. J., Savin, K., Bolormaa, S., McPartlan, H. C., Bowman, P. J., et al. (2012). Validation of single nucleotide polymorphisms associated with milk production traits in dairy cattle. *J. Dairy Sci.* 95, 864–875. doi: 10.3168/jds.2010-3786

Cheruiyot, E. K., Bett, R. C., Amimo, J. O., Zhang, Y., Mrode, R., and Mujibi, F. D. N. (2018). Signatures of selection in admixed dairy cattle in Tanzania. *Front. Genet.* 9:607. doi: 10.3389/fgene.2018.00607

Cole, J. B., and Silva, M. V. G. B. (2016). Genomic selection in multi-breed dairy cattle populations. *Rev. Bras. Zootecnia* 45, 195–202. doi: 10.1590/S1806-92902016000400008

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 U.S. Holstein cows. *BMC Genomics* 12:408. doi: 10.1186/1471-2164-12-408

Dasgupta, T., and Ladd, A. N. (2012). The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins. *Wiley Interdiscipl. Rev. RNA* 3, 104–121. doi: 10.1002/wrna.107

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

Folman, Y., Rosenberg, M., Ascarelli, I., Kaim, M., and Herz, Z. (1983). The effect of dietary and climatic factors on fertility, and on plasma progesterone and oestradiol-17 beta levels in dairy cows. *J. Steroid Biochem.* 19, 863–868. doi: 10.1016/0022-4731(83)90025-0

Freeman, A. R., Hoggart, C. J., Hanotte, O., and Bradley, D. G. (2006). Assessing the relative ages of admixture in the bovine hybrid zones of Africa and the Near East using X chromosome haplotype mosaicism. *Genetics* 173, 1503–1510. doi: 10.1534/genetics.105.053280

Gautier, M., Klassmann, A., and Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17, 78–90. doi: 10.1111/1755-0998.12634

Gautier, M., and Naves, M. (2011). Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.* 20, 3128–3143. doi: 10.1111/j.1365-294X.2011.05163.x

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., et al. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139, 907–920.

Glazar, A. I., and Evans, J. P. (2009). Immunoglobulin superfamily member IgSF8 (EWI-2) and CD9 in fertilisation: evidence of distinct functions for CD9 and a CD9-associated protein in mammalian sperm-egg interaction. *Reprod. Fertil. Dev.* 21, 293–303. doi: 10.1071/rd08158

Goszczynski, D. E., Corbi-Botto, C. M., Durand, H. M., Rogberg-Munoz, A., Munilla, S., Peral-Garcia, P., et al. (2018). Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. *Animal* 12, 215–223. doi: 10.1017/s1751731117001380

Hawken, R. J., Zhang, Y. D., Fortes, M. R., Collis, E., Barris, W. C., Corbet, N. J., et al. (2012). Genome-wide association studies of female reproduction in tropically adapted beef cattle. *J. Anim. Sci.* 90, 1398–1410. doi: 10.2527/jas.2011-4410

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Ilieva, K. M., Cheung, A., Mele, S., Chiaruttini, G., Crescioli, S., Griffin, M., et al. (2017). Chondroitin sulfate proteoglycan 4 and its potential as an antibody immunotherapy target across different tumor types. *Front. Immunol.* 8:1911. doi: 10.3389/fimmu.2017.01911

Irano, N., de Camargo, G. M. F., Costa, R. B., Terakado, A. P. N., Magalhães, A. F. B., Silva, R. M. O., et al. (2016). Genome-wide association study for indicator traits of sexual precocity in Nellore cattle. *PLoS ONE* 11:e0159502. doi: 10.1371/journal.pone.0159502

Iso-Touru, T., Sahana, G., Guldbrandtsen, B., Lund, M. S., and Vilkki, J. (2016). Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. *BMC Genet.* 17:55. doi: 10.1186/s12863-016-0363-8

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* 48, 811. doi: 10.1038/ng.3571

Lohmueller, K. E., Bustamante, C. D., and Clark, A. G. (2010). The effect of recent admixture on inference of ancient human population history. *Genetics* 185, 611–622. doi: 10.1534/genetics.109.113761

Loupatty, F. J., Clayton, P. T., Ruiter, J. P., Ofman, R., Ijlst, L., Brown, G. K., et al. (2007). Mutations in the gene encoding 3-hydroxyisobutyryl-CoA hydrolase results in progressive infantile neurodegeneration. *Am. J. Hum. Genet.* 80, 195–199. doi: 10.1086/510725

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

Mallikarjunappa, S., Sargolzaei, M., Brito, L. F., Meade, K. G., Karrow, N. A., and Pant, S. D. (2018). Short communication: uncovering quantitative trait loci associated with resistance to *Mycobacterium avium* ssp. paratuberculosis infection in Holstein cattle using a high-density single nucleotide polymorphism panel. *J. Dairy Sci.* 101, 7280–7286. doi: 10.3168/jds.2018-14388

Maltecca, C., Gray, K. A., Weigel, K. A., Cassady, J. P., and Ashwell, M. (2011). A genome-wide association study of direct gestation length in US Holstein and Italian Brown populations. *Anim. Genet.* 42, 585–591. doi: 10.1111/j.1365-2052.2011.02188.x

Marete, A., Sahana, G., Fritz, S., Lefebvre, R., Barbat, A., Lund, M. S., et al. (2018). Genome-wide association study for milking speed in French Holstein cows. *J. Dairy Sci.* 101, 6205–6219. doi: 10.3168/jds.2017-14067

Marshall, K., Gibson, J. P., Mwai, O., Mwacharo, J. M., Haile, A., Getachew, T., et al. (2019). Livestock genomics for developing countries - African examples in practice. *Front. Genet.* 10:297. doi: 10.3389/fgene.2019.00297

Martin, P., Szymanowska, M., Zwierzchowski, L., and Leroux, C. (2002). The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reprod. Nutr. Dev.* 42, 433–459. doi: 10.1051/rnd:2002036

Meyer, K. (2007). WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J. Zhejiang Univ. Sci. B* 8, 815–821. doi: 10.1631/jzus.2007.B0815

Miao, L., Wang, Y., Xia, H., Yao, C., Cai, H., and Song, Y. (2013). SPOCK1 is a novel transforming growth factor-beta target gene that regulates lung cancer cell epithelial-mesenchymal transition. *Biochem. Biophys. Res. Commun.* 440, 792–797. doi: 10.1016/j.bbrc.2013.10.024

Mrode, R., Tarekegn, G. M., Mwacharo, J. M., and Djikeng, A. (2018). Invited review: genomic selection for small ruminants in developed countries: how applicable for the rest of the world? *Animal* 12, 1333–1340. doi: 10.1017/S1751731117003688

Muller, 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

Nayeri, S., Sargolzaei, M., Abo-Ismail, M. K., May, N., Miller, S. P., Schenkel, F., et al. (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet.* 17:75. doi: 10.1186/s12863-016-0386-1

Ojango, J. M. K., Mrode, R., Rege, J. E. O., Mujibi, D., Strucken, E. M., Gibson, J., et al. (2019). Genetic evaluation of test-day milk yields from smallholder dairy production systems in Kenya using genomic relationships. *J. Dairy Sci.* 102, 5266–5278. doi: 10.3168/jds.2018-15807

Parker Gaddis, K. L., Null, D. J., and Cole, J. B. (2016). Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. *J. Dairy Sci.* 99, 6420–6435. doi: 10.3168/jds.2015-10444

Pryce, J. E., Bolormaa, S., Chamberlain, A. J., Bowman, P. J., Savin, K., Goddard, M. E., et al. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *J. Dairy Sci.* 93, 3331–3345. doi: 10.3168/jds.2009-2893

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

Raphaka, K., Matika, O., Sánchez-Molano, E., Mrode, R., Coffey, M. P., Riggio, V., et al. (2017). Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet.* 18:27. doi: 10.1186/s12863-017-0493-7

Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* 15:442. doi: 10.1186/1471-2164-15-442

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

Saxe, J. P., Chen, M., Zhao, H., and Lin, H. (2013). Tdrkh is essential for spermatogenesis and participates in primary piRNA biogenesis in the germline. *EMBO J.* 32, 1869–1885. doi: 10.1038/emboj.2013.121

Scanu, A. M., Bull, T. J., Cannas, S., Sanderson, J. D., Sechi, L. A., Dettori, G., et al. (2007). *Mycobacterium avium* subspecies paratuberculosis infection in cases of irritable bowel syndrome and comparison with Crohn's disease and Johne's disease: common neural and immune pathogenicities. *J. Clin. Microbiol.* 45, 3883–3890. doi: 10.1128/jcm.01371-07

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* 18:386. doi: 10.1186/s12864-017-3754-y

Settles, M., Zanella, R., McKay, S. D., Schnabel, R. D., Taylor, J. F., Whitlock, R., et al. (2009). A whole genome association analysis identifies loci associated with *Mycobacterium avium* subsp. paratuberculosis infection status in US holstein cattle. *Anim. Genet.* 40, 655–662. doi: 10.1111/j.1365-2052.2009.01896.x

Sharma, M., Kambadur, R., Matthews, K. G., Somers, W. G., Devlin, G. P., Conaglen, J. V., et al. (1999). Myostatin, a transforming growth factor-β superfamily member, is expressed in heart muscle and is upregulated in cardiomyocytes after infarct. *J. Cell. Physiol.* 180, 1–9. doi: 10.1002/(sici)1097-4652(199907)180:1<1::Aid-jcp1>3.0.Co;2-v

Sodeland, M., Kent, M. P., Olsen, H. G., Opsal, M. A., Svendsen, M., Sehested, E., et al. (2011). Quantitative trait loci for clinical mastitis on chromosomes 2, 6, 14 and 20 in Norwegian Red cattle. *Anim. Genet.* 42, 457–465. doi: 10.1111/j.1365-2052.2010.02165.x

Song, M., He, Y., Zhou, H., Zhang, Y., Li, X., and Yu, Y. (2016). Combined analysis of DNA methylome and transcriptome reveal novel candidate genes with susceptibility to bovine Staphylococcus aureus subclinical mastitis. *Sci. Rep.* 6:29390. doi: 10.1038/srep29390

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100:9440. doi: 10.1073/pnas.1533509100

Strucken, E. M., Al-Mamun, H. A., Esquivelzeta-Rabell, C., Gondro, C., Mwai, O. A., and Gibson, J. P. (2017). Genetic tests for estimating dairy breed proportion and parentage assignment in East African crossbred cattle. *Genet. Sel. Evol.* 49:67. doi: 10.1186/s12711-017-0342-1

Strucken, E. M., Swaminathan, M., Joshi, S., and Gibson, J. P. (2019). Genetic characterization of Indian indigenous cattle breeds. *Proc. Assoc. Advmt. Anim. Breed. Genet.* 23, 187–190.

Suchocki, T., and Szyda, J. (2015). Genome-wide association study for semen production traits in Holstein-Friesian bulls. *J. Dairy Sci.* 98, 5774–5780. doi: 10.3168/jds.2014-8951

Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5:e171. doi: 10.1371/journal.pbio.0050171

Tokuhiro, K., Isotani, A., Yokota, S., Yano, Y., Oshio, S., Hirose, M., et al. (2009). OAZ-t/OAZ3 is essential for rigid connection of sperm tails to heads in mouse. *PLoS Genet.* 5:e1000712. doi: 10.1371/journal.pgen.1000712

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

Viitala, S., Szyda, J., Blott, S., Schulman, N., Lidauer, M., Mäki-Tanila, A., et al. (2006). The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. *Genetics* 173, 2151–2164. doi: 10.1534/genetics.105.046730

Villani, A. C., Lemire, M., Louis, E., Silverberg, M. S., Collette, C., Fortin, G., et al. (2009). Genetic variation in the familial Mediterranean fever gene (MEFV) and risk for Crohn's disease and ulcerative colitis. *PLoS ONE* 4:e7154. doi: 10.1371/journal.pone.0007154

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi: 10.1146/annurev-genet-111212-133526

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

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford: Oxford University Press.

Weerasinghe, W. M. S. P. (2014). *The accuracy and bias of estimates of breed composition and inference about genetic structure using high density SNP markers in Australian sheep breeds* (Ph.D. Thesis). University of New England, Armidale, NSW, Australia.

**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: 04 September 2020
doi: 10.3389/fgene.2020.00752

# Haplotype Block Analysis Reveals Candidate Genes and QTLs for Meat Quality and Disease Resistance in Chinese Jiangquhai Pig Breed

*Favour Oluwapelumi Oyelami[1], Qingbo Zhao[1], Zhong Xu[1], Zhe Zhang[2], Hao Sun[1], Zhenyang Zhang[1], Peipei Ma[1], Qishan Wang[1,2]* and Yuchun Pan[1,2]**

[1] Department of Animal Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China,
[2] Department of Animal Breeding and Reproduction, College of Animal Science, Zhejiang University, Hangzhou, China

OPEN ACCESS

**Edited by:**
Marco Milanesi,
University of Tuscia, Italy

**Reviewed by:**
Zhe Zhang,
South China Agricultural University, China
Xiangdong Ding,
China Agricultural University, China
Christian H. U. W. Reimer,
University of Göttingen, Germany

***Correspondence:**
Qishan Wang
wangqishan@sjtu.edu.cn
Yuchun Pan
panyuchun1963@aliyun.com

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

**Received:** 01 December 2019
**Accepted:** 23 June 2020
**Published:** 04 September 2020

**Citation:**
Oyelami FO, Zhao Q, Xu Z, Zhang Z, Sun H, Zhang Z, Ma P, Wang Q and Pan Y (2020) Haplotype Block Analysis Reveals Candidate Genes and QTLs for Meat Quality and Disease Resistance in Chinese Jiangquhai Pig Breed. Front. Genet. 11:752. doi: 10.3389/fgene.2020.00752

The Jiangquhai (JQ) pig breed is one of the most widely recognized pig populations in China due to its unique and dominant characteristics. In this study, we examined the extent of Linkage disequilibrium (LD) and haplotype block structure of the JQ pig breed, and scanned the blocks for possible genes underlying important QTLs that could either be responsible for some adaptive features in these pigs or might have undergone some selection pressure. We compared some of our results with other Chinese and Western pig breeds. The results show that the JQ breed had the highest total block length (349.73 Mb ≈ 15% of its genome), and the coverage rate of blocks in most of its chromosomes was larger than those of other breeds except for *Sus scrofa* chromosome 4 (SSC4), SSC6, SSC7, SSC8, SSC10, SSC12, SSC13, SSC14, SSC17, SSC18, and SSCX. Moreover, the JQ breed had more SNPs that were clustered into haplotype blocks than the other breeds examined in this study. Our shared and unique haplotype block analysis revealed that the Hongdenglong (HD) breed had the lowest percentage of shared haplotype blocks while the Shanzhu (SZ) breed had the highest. We found that the JQ breed had an average $r^2 > 0.2$ at SNPs distances 10–20 kb and concluded that about 120,000–240,000 SNPs would be needed for a successful GWAS in the breed. Finally, we detected a total of 88 genes harbored by selected haplotype blocks in the JQ breed, of which only 4 were significantly enriched ($p\text{-value} \leq 0.05$). These genes were significantly enriched in 2 GO terms ($p\text{-value} < 0.01$), and 2 KEGG pathways ($p\text{-value} < 0.02$). Most of these enriched genes were related to health. Also, most of the overlapping QTLs detected in the haplotype blocks were related to meat and carcass quality, as well as health, with a few of them relating to reproduction and production. These results provide insights into the genetic architecture of some adaptive and meat quality traits observed in the JQ pig breed and also revealed the pattern of LD in the genome of the pig. Our result provides significant guidance for improving the statistical power of GWAS and optimizing the conservation strategy for this JQ pig breed.

**Keywords: linkage disequilibrium, haplotypes, conservation, complex traits, Chinese pigs**

## INTRODUCTION

The pig population in China (435 million) accounts for 45% of the total population of pigs in the world (FAOSTAT, 2017) and the Jiangquhai (JQ) pig breed is one of the most widely recognized pig populations in the country due to its unique and dominant characters. This pig breed is found in Jiangsu Province, in the eastern part of China where the giant Taihu lake is located. The JQ breed is known for its high performing economic traits like reproduction, adaptability, disease resistance, and the quality of its meat (China National Commission of Animal Genetic Resources, 2011).

The JQ pig breed has existed since the early 19th century and has many characteristics such as strong fat deposition and excellent tasting, high-quality meat. It is a well known local pig breed used in producing ham in China (China National Commission of Animal Genetic Resources, 2011), where there are three popular types of ham: Yun ham, Jinhua ham, and Rugao ham. While Yun hams are produced from three Yunnan province pig breeds, Jinhua ham is produced from Jinhua pig, and Rugao ham is produced from Jiangquhai (JQ) pig (Miao et al., 2009; Toldrá et al., 2014). Apart from JQ pigs, other pig populations such as Huaibei (HB), Hongdenglong (HD), Shanzhu (SZ), Dongchuan, Erhualian, Fengjing, Huai, Mi, and Shawutou are also distributed throughout Jiangsu province.

Recent studies had revealed high genetic diversity within the JQ pig breed (Hua et al., 2014; Xiao et al., 2017b; Xu et al., 2019). Genomic analysis also revealed that this breed might have undergone selection in the past, which could account for some adaptive traits in the breed (Meng et al., 2018; Xu et al., 2019). However, there is still a dearth of information on the genetic architecture of some economically important traits in this pig breed. Moreover, the adaptation of this breed to its environment is strongly supported by empirical evidence indicating that the genetic basis of its population differentiation is non-additive for fitness trait and that its adaptive gene complexes would be different from those of other breeds (Crnokrak and Roff, 1995). Therefore, it is imperative to understand the non-random genetic relationship between loci within the JQ pig population. This relationship is usually reflected by the pattern and extent of linkage disequilibrium (LD) that are inferred from the haplotypes in the genome.

Advancement in high-throughput genotyping technologies enables the use of large numbers of single nucleotide polymorphism (SNPs) in detecting haplotypes, which are products of introgression or selection during the domestication process of pigs (Amaral et al., 2008). These haplotypes can be inherited from one generation to the other as single units called haplotype blocks (Gabriel et al., 2002). Haplotype blocks are sections of the chromosome with high LD, low haplotype diversity, and low recombination rate (Luikart et al., 2003; Phillips et al., 2003). Many haplotype blocks may arise as a result of several factors such as chromosomal recombination, selection, population bottlenecks, population admixture, and mutations (Phillips et al., 2003; Guryev et al., 2006). Previous studies have reported a low level of admixture in the JQ pig breed, however, the degree of admixture of this breed by possible sources of admixtures is unknown. Therefore, the identification of the percentage of foreign haplotypes in the JQ breed could serve as a useful framework of future breeding actions and decisions when setting up a conservation program for the breed. Moreover, since the evolutionary history of a breed can be inferred from the pattern of LD in the genome (Hayes et al., 2003), the characterization of the patterns of LD across the genome of JQ pigs could potentially improve our understanding of the biological pathway of recombination in the breed, and also help to detect some selection footprints in the genome. Furthermore, characterizing the LD structure in the genome is particularly important for the interpretation and application of results of genome-wide association studies (GWAS) (Meuwissen and Goddard, 2000).

Over the years, haplotypes have proven to be more powerful in association studies than single-marker methods (Lin et al., 2009). Thus, they have a point of reference in GWAS, especially in the case of ungenotyped SNPs. Haplotype blocks can be used to identify significant variants in GWAS and also for predicting the genomic breeding values (GEBV) of animals in genomic selection (GS) programs (Meuwissen et al., 2001; Calus et al., 2008; Cuyabano et al., 2015; Chen et al., 2018). Therefore, the characterization of LD patterns in the genome of JQ pigs has a potential application in future studies of complex traits and the development of genomic tools for the breed (Corbin et al., 2010).

Since the extent of LD and haplotype blocks are of critical importance for genomic selection, marker-assisted selection, and conservation of animal genetic resources, the importance of constructing the haplotype blocks in the JQ pig breed and identifying the genes involved in them, especially those associated with economically important traits, cannot be overemphasized. Such information would help in understanding the genetic basis of breed distinction and adaptation and guide against incorporating haplotype blocks with deleterious gene effects into selection programs (Salem et al., 2018). To our knowledge, no haplotype block study has been conducted on this pig breed despite its unique characteristics and there is still a knowledge gap on the genetic basis of its phenotypic distinction. To this end, this research was conducted to (1) analyze the haplotype block structure of the JQ pig breed and compare it with seven other pig breeds (five Chinese and two western breeds), (2) examine the pattern of linkage disequilibrium (LD) in the JQ breed, and (3) scan the blocks for possible genes underlying important QTLs that span across the blocks. Our result provides a theoretical basis for designing breeding programs aimed at conserving economically important traits in the JQ breed and potential genetic improvement programs for this breed in the future.

# MATERIALS AND METHODS

## Animal Samples, Genotyping and Quality Control

A total of 192 pigs were used in this study. Of the total pig population, thirty-eight (38) were Jiangquhai (JQ) pigs from the pig conservation farm in Jiangsu province. Other pig breeds used as a reference population were; Huaibei (HB, $n$ = 34), Shanzhu (SZ, $n$ = 20), and Hongdenglong (HD, $n$ = 30) breeds, also from Jiangsu Province; Middle Meishan (MMS, $n$ = 20) and PudongWhite (PD, $n$ = 20) pigs from Shanghai province; and, Duroc (D, $n$ = 10) and Yorkshire (Y, $n$ = 20), which are western pig breeds. The Jiangsu pig samples from the conservation pig farms have been described in previous studies (Xiao et al., 2017b; Zhang et al., 2018). In these previous studies, the individuals were genotyped using the genotyping by genome reducing and sequencing (GGRS) protocol[1] (Chen et al., 2013). Briefly, genomic DNA samples were extracted from ear tissue, using a Lifefeng blood and tissue extraction kit [Lifefeng Biotech (Shanghai) Co., Ltd., China], digested with a restriction enzyme (*Ava*II), and then ligated with a unique adapter barcode after which the samples were pooled and enriched through PCR to construct a sequencing library. Finally, the DNA sequence libraries (fragments lengths of 300–400 bp, including the adapter barcode sequence) were sequenced using an Illumina HiSeq2500 (100 paired-end) sequencing platform according to the manufacturer's protocol.

Quality control of sequences was performed using NGS QC Toolkit v2.3 and the parameters were set according to a report from Chen et al. (2013). The sequencing reads were aligned to the pig reference genome (Sscrofa11.1) using BWA (Li and Durbin, 2009). The BAM files from the alignments were used to call and genotype SNPs using SAMtools (Li et al., 2009). These variants were then filtered and SNPs with a quality score greater than or equal to 20 (i.e., more than 99% accuracy), average sequencing depth > 5x, and minor allele frequency (MAF) greater than or equal to 0.03 were retained for imputation (Chen et al., 2013; Wang et al., 2015). To ensure the precision of imputation and density of SNPs, only those genotyped in >30% of samples were retained (Wang et al., 2015). BEAGLE v4.1 was used to impute the missing genotypes in this study with default parameters (Browning and Browning, 2016). A total of 486,018 SNPs, which passed the filtering threshold, were later separated into different populations and filtered for MAF $\geq$ 0.05. After discarding SNPs on the Y Chromosome, a total of 270,935, 223,897, 317,597, 210,277, 204,790, 237,962, 173,678, and 221,957 SNPs, with MAF $\geq$ 0.05, were retained in JQ, HB, SZ, HD, MMS, PD, D and Y breed, respectively. The alignment and variant calling statistics are presented in **Supplementary Tables S1, S2**.

## Genetic Relationships and Population Structure

To estimate the genetic distances within breeds, the average proportion of alleles shared, Dst, was calculated using PLINK v1.9 (Chang et al., 2015). The definition of Dst is as follows (Chang et al., 2015):

$$Dst = \frac{IBS_2 + 0.5^{*}IBS_1}{N}$$

$IBS_1$ and $IBS_2$ are the numbers of loci that share 1 or 2 alleles identical by state (IBS), respectively, and N is the number of loci tested. The genetic distance (D) between all pairwise combinations of individuals was calculated as follows: 1-Dst. Pairwise genetic differentiation (fixation index, $F_{ST}$) (Weir and Cockerham, 1984) between all pairs of pig breeds were calculated using the R package "diversity" (Keenan et al., 2013). Based on the matrix of pairwise $F_{ST}$ values, a Neighbor-Net tree was constructed using SplitsTree 4.14.5 software (Huson and Bryant, 2006).

To illustrate the population structure and infer genetic admixture between populations, a total of 91,092 SNPs, which discarded SNPs that were with extreme deviations from Hardy-Weinberg equilibrium ($p$-value $\leq 1 \times 10^{-6}$), MAF < 0.05, and LD (linkage disequilibrium) greater than 0.5 across populations (command: PLINK indep-pairwise 50 5 0.5), were used for population structure analysis using ADMIXTURE v1.3 software (Alexander et al., 2009). The number of ancestral clusters (K) was set from 2 to 9, and a five-fold cross-validation was run to determine the $K$ value with the lowest cross-validation error. The result was displayed using the web-based software, Clumpak[2] (Kopelman et al., 2015).

## Effective Population Size

The historical effective population size ($N_e$) of each breed was estimated using the SNP data from the admixture analysis. $N_e$ was estimated using the software SNeP (Barbato et al., 2015). SNeP estimates $N_e$ at different $t$ generations based on the LD between SNPs, where $t = [2f(c_t)]^{-1}$, and $c_t$ is the recombination rate for specific physical distance between markers, measured in Morgan (Hayes et al., 2003) (assuming 100 Mb = 1Morgan). The following options were also used in SNeP: (1) sample size correction; (2) correction to account for the occurrence of mutation; (3) Sved and Feldman's recombination rate modifier (Sved and Feldman, 1973).

## Linkage Disequilibrium

Linkage disequilibrium, $r^2$ value was used as a measure of LD between each locus because of its preference in association studies (Wall and Pritchard, 2003; Bohmanova et al., 2010). We estimated pairwise LD ($r^2$) for all retained SNPs within each breed using the command line "–ld- window-r2 0" in PLINK v1.9 (Chang et al., 2015). This procedure used a default maximum window size of 1 Mb between the estimated pair of SNPs on a chromosome. The extent and decay of LD in each breed were also predicted using the following equation (Sved, 1971; Heifetz et al., 2005; Amaral et al., 2008; Ai et al., 2013):

$$LD_{ijk} = \frac{1}{1 + 4\beta_{jk}d_{ijk}} + e_{ijk}$$

[1] http://klab.sjtu.edu.cn/GGRS/

[2] http://clumpak.tau.ac.il/distruct.html

Where $LD_{ijk}$ is the observed LD for marker pair $i$ of breed $j$ in genomic region $k$, $d_{ijk}$ is the distance in base pairs for marker pair $i$ of breed $j$ in genomic region $k$, $\beta_{jk}$ is the coefficient that describes the decline of LD with distance for breed $j$ in genomic region k and $e_{ijk}$ is a random residual. The $LD_{ijk}$, $\beta_{jk}$, and $e_{ijk}$ for each genomic region within each breed were estimated using the Beta.nonlinear fit function in $R$[3] (Amaral et al., 2008). They were fitted for the following genomic distances; 0, 4, 8, 12, 20, 30, 40, 60, 80, 100, 120, 160, 200, 250, 300, 360, 460, 620, 800, and 1000 kb. The decay of LD was plotted for both, autosomes and SSCX of each breed. To further assess the extent of LD across breeds, the LD ($r^2$) between all autosomal SNPs was, however, divided into the following bin distances: 0–10, 10–20, 20–40, 40–60, 60–100, 100–200, 200–500, and 500–1000 kb.

## Haplotype Block Construction and Haplotype Diversity

A Hidden Markov Model implemented in the program BEAGLE v4.1 software (Browning and Browning, 2016) was used to reconstruct the haplotype phase. Hereafter, haplotype blocks were estimated separately in each breed using PLINK v1.9 (Chang et al., 2015) following the default procedure in HAPLOVIEW (v4.1) (Barrett et al., 2005). The method followed for block definition was previously described by Gabriel et al. (2002). Furthermore, to investigate the pattern of LD within blocks, a haploview plot was constructed for some haplotype blocks, based on LD ($r^2$) value between SNP pairs, using the HAPLOVIEW software (Barrett et al., 2005).

As a measure of genetic diversity, we estimated haplotype diversity across breeds. First, we calculated the haplotype frequency for each breed using PLINK v1.07 (Purcell et al., 2007) (because PLINK v1.9 does not currently support the –hap flag which is needed to calculate the haplotype frequency). Afterward, the haplotype diversity across breeds was estimated. Haplotype diversity is defined as $1 - \sum f_i^2$ where $f_i$ is the frequency of the ith haplotype. To gain insight into the haplotype diversity within the block region (with the maximum number of SNPs in JQ breed), we applied the "four-gamete rule block definition algorithm" implemented in haploview software. This algorithm computes the observed frequency of the four possible two-marker haplotypes for each pair of SNPs and defines a block when the frequency is 0 (i.e., no recombination event has occurred) (Wang et al., 2002). A frequency of at least 0.01 between computed four marker-haplotypes indicates that a recombination event between the two markers likely occurred.

The shared and unique haplotype block regions between breeds were also detected. Shared haplotype blocks were defined as the overlapping block regions shared by two populations or more, while the unique haplotype blocks were the block regions specific to each population. Both the shared and unique haplotype block regions were detected and visualized using the R Bioconductor package "GenomicRanges" (Lawrence et al., 2013) and 'ggbio' (Yin et al., 2012), respectively.

[3]http://www.r-project.org/

## QTLs and Functional Gene Set Enrichment Analysis

To detect the possible genes and important QTLs that span the haplotype blocks, we hypothesized that important traits under selection for adaptation of the JQ pig breed could be harbored in haplotype block regions with the highest block length. We also theorized that haplotype blocks with the highest number of SNPs could reveal some important genetic variations in the JQ breed. Therefore, we chose the first ten haplotype block regions within each respective criterion for functional annotation. In total, 20 block regions were annotated for possible QTLs and genes.

The QTL regions spanned by the haplotype block of the JQ breed were detected by mapping selected haplotype block regions onto QTL sections using data from the Pig QTL database[4]. To ensure efficient processing and control the volume of QTLs detected, we filtered out QTL regions to lengths $\leq$ 10 Mb, afterward, a Perl homemade script was used to detect haplotype block regions with more than 50% overlap with the filtered QTL regions. The QTLs that fall within the selected haplotype block regions or the haplotype block regions that fall within the QTLs are defined as overlap.

Furthermore, we performed a gene set enrichment analysis (GSEA) to further elucidate the biological function of the selected haplotype block regions above. We mapped the selected haplotype block regions and genes using gene annotation data for pigs from the Ensembl gene database 98[5]. Thereafter, the detected genes were functionally annotated by performing the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa et al., 2012) and Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis using Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8)[6] (Huang et al., 2009). We defined a significant threshold $p$-value to be 0.05 (based on EASE score: a modified Fisher's exact test), and then selected the most significantly enriched genes with FDR (False Discovery Rate) < 15%. We also established the relationship between the likely candidate genes and QTLs detected in the haplotype blocks. This result potentially reveals the genes and characters that might have either undergone artificial or natural selection pressure in the JQ breed or the genes involved in complex traits of the breed.

# RESULTS

## Genetic Relationship and Population Structure

The average genetic distances (Dst) within the 8 populations were 0.210 (JQ), 0.181 (HD), 0.198 (HB), 0.255 (SZ), 0.164 (MMS), 0.203 (PD), 0.133 (D), and 0.167 (Y). The highest genetic differentiation (0.441) between breeds was found between MMS and D breed, while the lowest was found between MMS and PD breed (0.093) (**Supplementary Table S3**). A Neighbor-Net tree

[4]https://www.animalgenome.org/cgi-bin/QTLdb/SS/index

[5]http://asia.ensembl.org/Sus_scrofa/Info/Index

[6]https://david.ncifcrf.gov/home.jsp

constructed based on this pairwise *Fst* value between breeds is presented in **Figure 1A**. Our admixture analysis revealed some level of introgression between the Chinese breeds in this study (**Figure 1C**). We observed the lowest cross-validation error when $K = 7$ (**Figure 1B**), before PD separated from MMS into a different cluster. Suggesting a continuous gene flow between PD and the MMS breed (Xiao et al., 2017a). Consistent with previous findings and the genetic origins of worldwide pig breeds

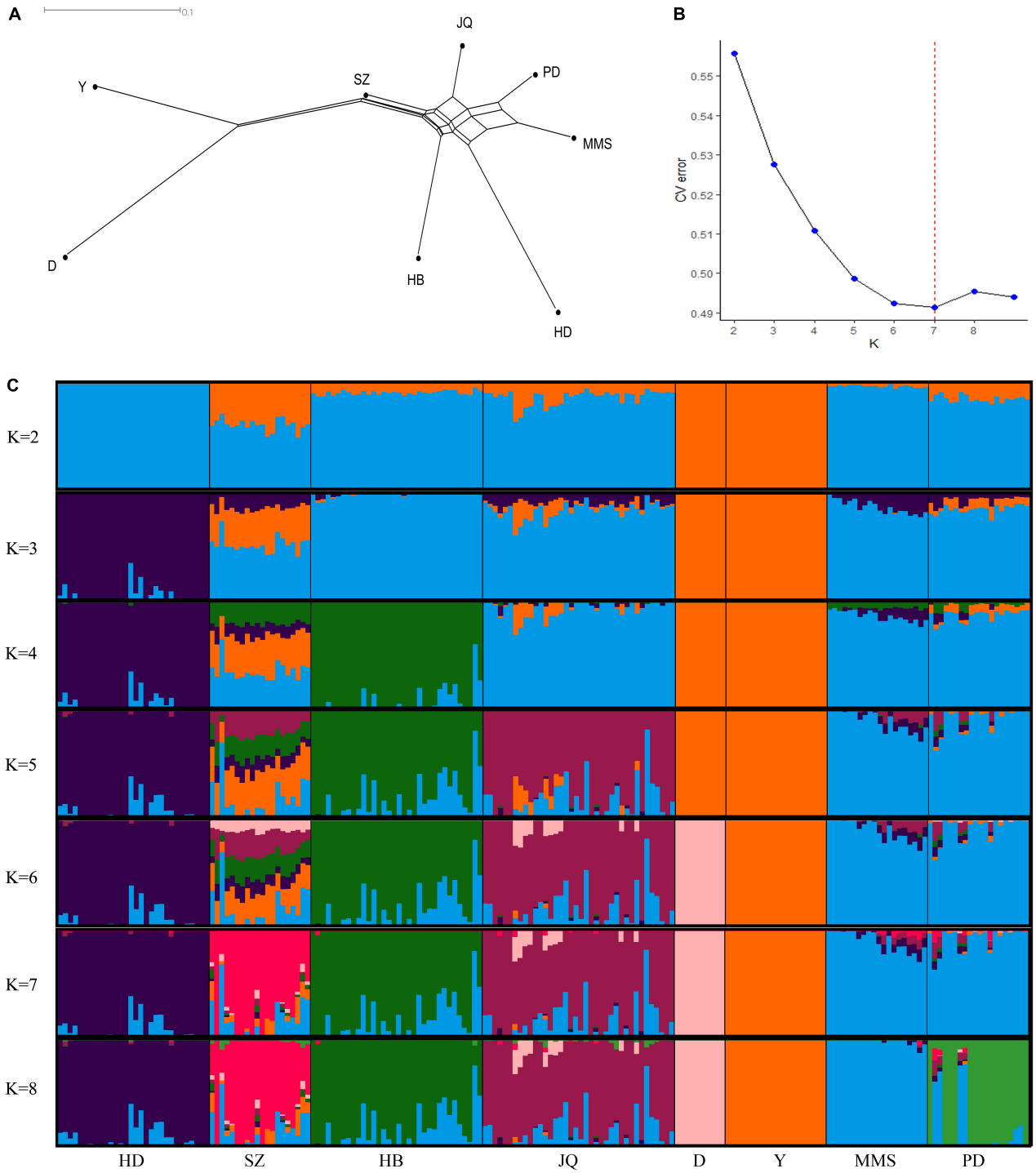


**FIGURE 1 |** Population structure of the studied pigs. **(A)** Neighbor-Net tree showing genetic differentiation (Fst) between breeds. **(B)** Cross-validation errors of the admixture analysis at different *K* values (lowest *k*-value = 7). **(C)** Population structure of the 8 breeds analyzed based on Admixture analysis. Hongdenglong (HD), Shanzhu (SZ), Huaibei (HB), Jiangquhai (JQ), Duroc (D), Yorkshire (Y), MiddleMeishan (MMS), and PudongWhite (PD).

(Fan et al., 2002; Ai et al., 2013; Zhang et al., 2018), $K = 2$ shows the ancient divergence between Asian and European pigs, indicating that the MMS breed was the ancestral population of the Chinese pigs examined in our study. This could explain why there are still some ancestral haplotypes of MMS in current Chinese pig populations.

## Effective Population Size

The estimated $N_e$ trend of each pig breed across different generations is shown in **Figure 2**. This estimate can improve our understanding of the demographic history of each population in the recent past (Barbato et al., 2015). While the extent of LD over longer recombination distances reflected more recent $N_e$, that over shorter distances provided ancestral $N_e$ (Hayes et al., 2003). The result showed that all the breeds had experienced a decrease in *Ne* estimate over time, especially from 900 to about 50 generations ago. We observed the nearest anti-climax points between 900 and 1000 generations ago, which indicated the nearest starting point of human-driven artificial selection that might have caused a population bottleneck in the breeds. In general, the western pig breeds had smaller *Ne* compared to the Chinese pigs and this can be attributed to the higher LD observed in western pig breeds (Amaral et al., 2008). In particular, we observed that the effective population size in the last 13 generations of the JQ breed was about 109 and about 3,871 in ∼1000 generations ago. This reduction might be due to an increase in inbreeding rate and a reduced genetic diversity usually observed in animals with a small population size (Food and Agriculture Organization, 2013).

## The Extent of Linkage Disequilibrium Across Breeds

A total of 268,369, 221,481, 313,100, 208,264, 202,599, 234,967, 171,952, and 219,248 autosomal SNPs were found in the JQ, HB, SZ, HD, MMS, PD, D and Y pig breeds, respectively. While, on the SSCX, we obtained 2,566, 2,416, 4,497, 2,013, 2,191, 2,995, 1,726, 2,709 SNPs in the respective breeds. These SNPs (on SSCX) were only utilized in characterizing the LD and haplotype block structure of the breeds.

The average $r^2$ between adjacent SNPs on the autosomes was largest for D breed ($r^2 = 0.39$), followed by Y ($r^2 = 0.34$), whereas other pigs exhibited a smaller average $r^2$, ranging from 0.22 (SZ) to 0.32 (HD). The average autosomal LD ($r^2$) for the following bin distances 0–10, 10–20, 20–40, 40–60, 60–100, 100–200, 200–500, and 500–1000, is presented in **Figure 3**. On the SSCX, the average $r^2$ value observed for both D ($r^2 = 0.45$) and Y ($r^2 = 0.37$) breed was also the highest. However, the average LD ($r^2$) decreased in other breeds; PD ($r^2 = 0.32$), HD ($r^2 = 0.32$), SZ ($r^2 = 0.31$), MMS ($r^2 = 0.29$), HB ($r^2 = 0.27$), and JQ ($r^2 = 0.23$). Overall, on the

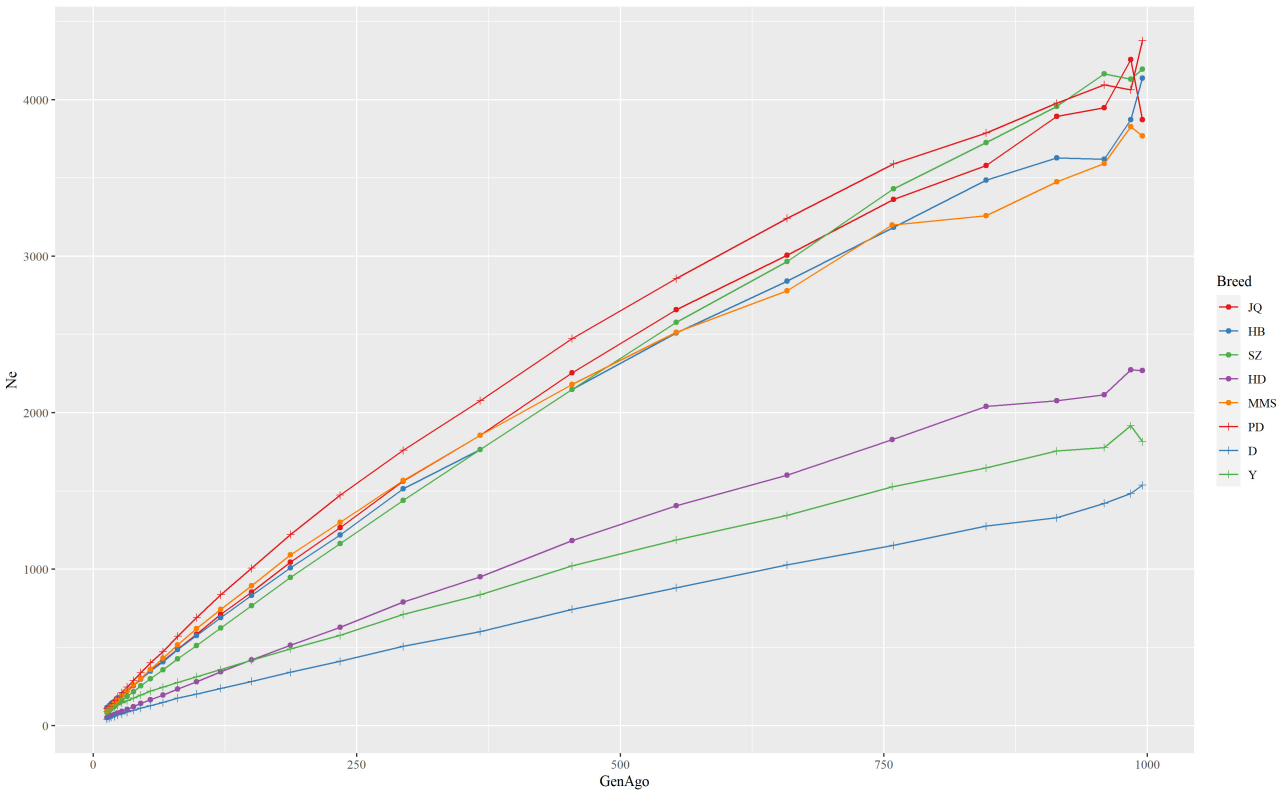


**FIGURE 2 |** The estimate of the effective population size ($N_e$) trend of each pig breed from 13 to about 1000 generations ago. The genome-wide estimate of $N_e$ was based on the linkage disequilibrium between SNPs and corrected for sample size, mutation, and recombination rate. Each line shows the trend in effective population size across generations. The result showed that the JQ breed had experienced a rapid decline in its population, including the most recent generation.

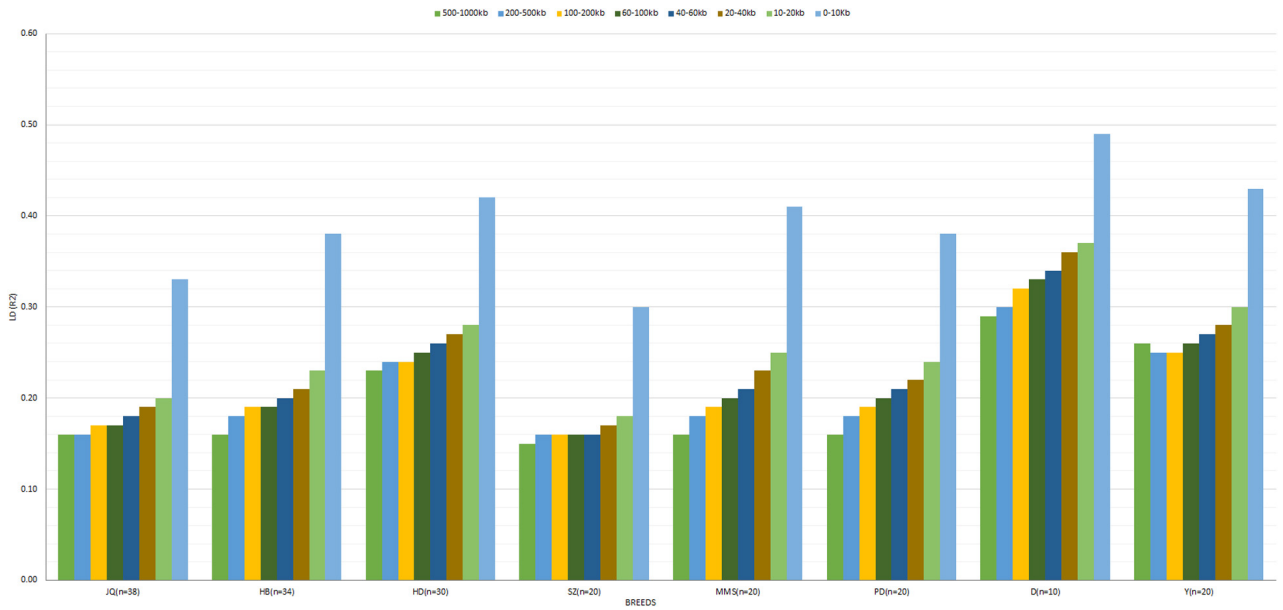


**FIGURE 3 |** The averaged LD ($r^2$) value at SNP distances 0–10, 10–20, 20–40, 40–60, 60–100, 100–200, 200–500, and 500–1000 kb across breeds. Jiangquhai (JQ), Huaibei (HB), Hongdenglong (HD), Shanzhu (SZ), MiddleMeishan (MMS), PudongWhite (PD), Duroc (D), and Yorkshire (Y). This LD estimate is non-fitted.

autosome of the JQ breed, about 28% of adjacent SNP pairs had $r^2 > 0.3$ and 36% had $r^2 > 0.2$. The corresponding percentages for HB, SZ, HD, MMS, PD, D, and Y were about 32 and 41%, 25 and 33%, 40 and 41%, 34 and 41%, 32 and 39%, 48 and 56%, 42 and 50%, respectively.

In general, the genome-wide average LD ($r^2$) across breeds decreased with increasing SNP pair distance (**Figure 3** and **Supplementary Figure S1**). A lower LD, which rapidly decayed with increasing genomic distance, especially for distances greater than 10 kb, was observed across breeds. As expected, large LD differences were observed between the Western (especially D) and Chinese breeds. Interestingly, we found that the LD decay on SSCX, across each breed, was slower compared to the autosomes' (**Supplementary Figure S1**). We also observed that the LD decay on the SSCX of D breed was slower than other breeds, while that of the JQ breed was faster. Apart from D and Y breeds, the PD breed also had a slower LD decay on the SSCX compared to other breeds in the study.

## Haplotype Block Structure and Haplotype Diversity

To gain insight into the systematic difference in the level of LD across each pig breeds, we characterized their haplotype blocks. Among all the pig breeds analyzed in this study, the JQ breed had the highest total autosomal block length, 345.30 Mb (14.18% of its total genome) while HB, SZ, HD, MMS, PD, D, and Y had 300.83 Mb (12.35%), 92.20 Mb (3.79%), 330.41 Mb (13.57%), 167.88 Mb (6.90%), 211.38 Mb (8.68%), 33.04 Mb (1.36%), and 176.35 Mb (7.24%) total autosomal block length, respectively. Moreover, fewer haplotype blocks (2,286) were observed on the autosome of D breed compared to others (**Table 1**), possibly due to a bias in its small sample size and a high percentage of fixed markers that were not involved in the haplotype block construction. On the SSCX, the total lengths of block (and average block size) were 4.43 Mb (15.32 kb), 6.49 Mb (23.09 kb), 2.75 Mb (12.31 kb), 5.61 Mb (21.50 kb), 2.62 Mb (22.75 kb), 1.96 Mb (14.52 kb), 1.51 Mb (47.05 kb), and 3.40 Mb (15.76 kb) for JQ, HB, SZ, HD, MMS, PD, D, and Y breed, respectively (**Table 2** and **Supplementary Tables S4–S10**). We also found that the number of maximum haplotype block size per chromosome across breeds was larger in both JQ and HB except for SSC3, SSC4, SSC5, SSC8, SSC9, SSC12, SSC15, SSC16, SSC18, and SSCX (**Figure 4**).

Furthermore, the coverage rate of blocks per chromosome in the JQ breed was higher than those of other breeds except for *Sus scrofa* chromosome 4 (SSC4), SSC6, SSC7, SSC8, SSC10, SSC12, SSC13, SSC14, SSC17, SSC18, and SSCX (**Supplementary Table S11**). The average block size in JQ breed was 10.90 kb (ranging from 0.002 to 199.97 kb) (**Table 2**). The average block size distribution across breeds is presented in **Figure 5B**. We also investigated the pattern of LD in the haplotype block region with the highest number of SNPs in the JQ breed. This block displayed a moderate LD (**Figure 6**) and high haplotype diversity (**Figure 7C**) suggesting that several recombination events might have occurred in this haplotype block. Furthermore, we observed a low LD level in the haplotype block with the maximum block length (199.97 kb), and a complete LD in the block with the minimum number of SNPs and block size (0.002 kb) (**Figures 7A,B**).

Generally, we found that the haplotype frequency and diversity across breeds (**Table 1**) were lower in all the Jiangsu pig breeds (JQ, HB and HD) except for SZ breed which

**TABLE 1 |** The number of haplotype blocks, haplotype frequency, and diversity across breeds.

| Breed | JQ | HB | SZ | HD | MMS | PD | D | Y |
|---|---|---|---|---|---|---|---|---|
| No. of haplotype blocks | 31146 | 26619 | 19362 | 25645 | 12312 | 14861 | 2286 | 16649 |
| Haplotype frequency | 0.25 | 0.26 | 0.34 | 0.27 | 0.29 | 0.29 | 0.36 | 0.30 |
| Haplotype diversity | 0.464 | 0.465 | 0.483 | 0.467 | 0.482 | 0.489 | 0.525 | 0.484 |

*Results in the table are derived from autosomal blocks.*

**TABLE 2 |** Block statistics of JQ breed.

| | | | Block size (kb) | | | | SNPs (*n*) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SSC | Blocks (*n*) | Total block length (kb) | Mean | Min | Max | No. of SNPs in blocks (*n*) | Mean | Min | Max | % of SNPs in blocks |
| 1 | 2336 | 43176.07 | 18.48 | 0.002 | 199.95 | 10587 | 4.53 | 2 | 44 | 7.50 |
| 2 | 1908 | 28761.04 | 15.07 | 0.002 | 199.56 | 9326 | 4.89 | 2 | 52 | 6.61 |
| 3 | 2385 | 24841.31 | 10.42 | 0.002 | 198.34 | 10883 | 4.56 | 2 | 37 | 7.71 |
| 4 | 1741 | 18151.28 | 10.43 | 0.002 | 198.69 | 7621 | 4.38 | 2 | 40 | 5.40 |
| 5 | 1662 | 14397.77 | 8.66 | 0.002 | 197.18 | 7353 | 4.42 | 2 | 51 | 5.21 |
| 6 | 2838 | 34404.52 | 12.12 | 0.002 | 199.91 | 13751 | 4.85 | 2 | 55 | 9.75 |
| 7 | 2003 | 17262.21 | 8.62 | 0.002 | 198.72 | 8767 | 4.38 | 2 | 48 | 6.21 |
| 8 | 1330 | 13453.34 | 10.12 | 0.002 | 199.18 | 5594 | 4.21 | 2 | 32 | 3.96 |
| 9 | 1955 | 22753.54 | 11.64 | 0.002 | 199.19 | 9091 | 4.65 | 2 | 43 | 6.44 |
| 10 | 1398 | 8856.84 | 6.34 | 0.002 | 195.37 | 5995 | 4.29 | 2 | 35 | 4.25 |
| 11 | 997 | 9979.21 | 10.01 | 0.002 | 195.15 | 4327 | 4.34 | 2 | 27 | 3.07 |
| 12 | 1514 | 8364.55 | 5.53 | 0.002 | 178.76 | 6515 | 4.30 | 2 | 26 | 4.62 |
| 13 | 1681 | 22458.83 | 13.36 | 0.002 | 199.97 | 7277 | 4.33 | 2 | 41 | 5.16 |
| 14 | 2362 | 25559.27 | 10.82 | 0.002 | 199.91 | 10891 | 4.61 | 2 | 43 | 7.72 |
| 15 | 1688 | 24085.86 | 14.27 | 0.002 | 199.60 | 7493 | 4.44 | 2 | 33 | 5.31 |
| 16 | 1064 | 11134.99 | 10.47 | 0.002 | 198.01 | 4456 | 4.19 | 2 | 33 | 3.16 |
| 17 | 1268 | 10090.86 | 7.96 | 0.002 | 199.97 | 5647 | 4.45 | 2 | 55 | 4.00 |
| 18 | 1016 | 7565.22 | 7.45 | 0.002 | 195.75 | 4349 | 4.28 | 2 | 80 | 3.08 |
| X | 289 | 4427.64 | 15.32 | 0.002 | 198.11 | 1177 | 4.07 | 2 | 23 | 0.83 |
| **Total** | **31435** | **349724.35** | **10.90** | | | **141100** | | | | **100.00** |

exhibited a higher haplotype frequency and diversity similar to that of Y.

## Distribution of SNPs in Haplotype Blocks

The density of SNPs in the Chinese and western pig population in this study is presented in **Supplementary Figures S2, S3**. The summary of SNPs distribution and proportion involved in the haplotype block formation per chromosome across breeds was also presented in **Table 2** and **Supplementary Tables S4–S10**. In summary, a total of 139,923, 116,377, 74,629, 113,762, 61,489, 77,729, 12,080, 75,997 SNPs located on the autosomes of JQ, HB, SZ, HD, MMS, PD, D, and Y breed were clustered into haplotype blocks respectively. These SNPs account for about 52.14, 52.55, 23.84, 54.62, 30.35, 33.08, 7.03, and 34.66% of all the autosomal SNPs in the respective breeds.

The frequency distribution of SNPs in the haplotype blocks for each breed is presented in **Figure 5A**. Generally, we observed a small proportion of haplotype blocks with more than 10 SNPs across each breed in this study. However, JQ and HD breeds had the highest number of blocks, with at least 10 SNPs. Intriguingly, among all the Chinese breeds in our study, the JQ breed had the highest number of SNPs in a block, with 80 SNPs in Block 22 of Chromosome 18 (923784 bp – 1002867 bp). This block overlaps the protein tyrosine phosphatase receptor (*PTPRN2*) gene, which suggests that it is associated with oncogenic processes (Bourgonje et al., 2016). This gene is also predominantly expressed in endocrine and neuronal cells, where it functions in exocytosis (Sorokin et al., 2015). Generally, JQ pigs are known for their high resistance to porcine reproductive and respiratory syndrome virus (PRRSV) infection (Meng et al., 2018).

This study also discovered that the highest amount of SNPs involved in block formation on chromosomes is observed in the JQ and other Jiangsu pig breeds (13,751, 12,403, 6,966, and 11,738 SNPs on Chromosome 6 of JQ, HB, SZ, and HD, respectively), while the lowest amount of SNPs on the autosomes of these pig breeds was 4,327, 3,083, and 3,013 on chromosome 11 of JQ, HB, and HD; and 2,276 on chromosome 16 of SZ breed. Conversely, the highest number of SNPs in other pigs was 7,940 and 6,274 on chromosome 6 of PD and MMS; 1,183 on chromosome 1 of D; and 7,799 on Chromosome 6 of Y breed. However, the lowest (total) number of SNPs in blocks (formed on the autosome) was 2,488, 1,580, 338, and 2,070, and was found on chromosome 16, 16, 18, and 16

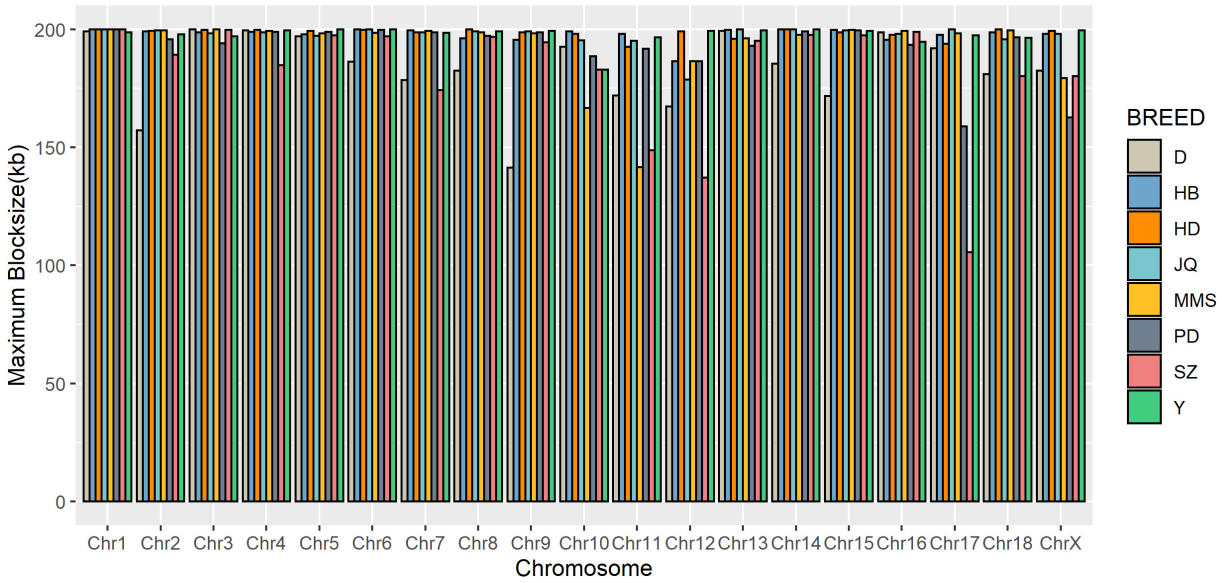


**FIGURE 4 |** Distribution of maximum haplotype block length per chromosome across breeds. Jiangquhai (JQ) and Huaibei (HB) had more chromosomes with maximum block length compared to other breeds.

of PD, MMS, D, and Y breed, respectively (**Supplementary Tables S7–S10**).

## Shared and Unique Haplotype Block Regions Between Breeds

As shown in **Table 3**, among all the pig breeds considered in this study, HD had the lowest percentage of shared haplotype blocks (with other breeds) while the SZ and MMS breed had the highest percentage of shared haplotype blocks. This result could be linked to the ancestral origin of MMS and the high admixture observed in SZ (**Figure 1C**). Among all the Chinese pig breeds in our study, the SZ breed had the highest percentage of shared haplotype block (18.12%) with Y (**Table 3**), indicating a high introgression of Y haplotype into the SZ breed. This result is also in line with our admixture analysis (**Figure 1C**) which suggests that the SZ breed might have been introgressed with different breeds in the past. JQ, HB, and PD breeds also shared a considerably high percentage of haplotype blocks with the western breeds (D and Y), suggesting an introgression between western pigs and these Chinese breeds. All the pig breeds included in our study, shared the highest percentage of their haplotype block with the JQ breed, which suggests a common ancestry (or introgression) between JQ, the western breeds (Bosse et al., 2014), and other Chinese pigs in our study. This could also be because the JQ breed had the highest number of haplotype blocks (**Table 3**).

In general, the total length of haplotype block shared across the autosome of all the Jiangsu province pigs (JQ, HB, SZ, and HD) was 7.89 Mb (**Supplementary Data**). These shared haplotype blocks could indicate the existence of conserved genomic regions that are a product of intensive and directional natural or artificial selection in the Jiangsu pig population. The plot of shared and unique haplotype block region in the Jiangsu pig population is presented in **Supplementary Figures S4–S6**.

## Functional Annotation of Overlapping QTLs and Genes

We detected the QTLs spanned by the haplotype block regions of JQ breed by finding the overlapping regions with 25,388 QTLs (length $\leq$ 10 Mb) downloaded from the pig QTL database. Consequently, 112 porcine QTLs were detected to overlap with the haplotype block regions of this breed. Interestingly, we found that most of the detected QTLs were related to meat and carcass quality, health, and a few reproduction and production-related QTLs. We detected QTLs related to traits such as feed conversion ratio, loin muscle area, body weight, intramuscular fat content, scrotal/inguinal hernia, teat number, total number born alive, change in *Mycoplasma hyopneumoniae* antibody titer, and toll-like receptor 9 level (**Table 4**), which suggest that the pig breed might have previously undergone selection for meat quality and health (an indication of the environmental adaptability of the breed). Specifically, about 13% of the QTLs (based on QTL IDs reported in the Pig QTL database), overlapping in the 20 scanned blocks in the JQ pig breed were related to drip loss (DRIPL) (water holding capacity of pork meat), and loin muscle area (LMA).

Furthermore, we identified a total of 88 genes harbored by the selected haplotype blocks (**Supplementary Table S12**), of which only 4 were significantly enriched (*p-value* $\leq$ 0.05). These genes were significantly enriched in 2 GO terms and 2 KEGG pathways (**Table 5**) which were related to a variety of molecular functions linked to immunity. It is of note that two of the enriched genes (*ACVRL1* and *ACVR1B*) found on SSC5 were enriched (GO:0003840) in the molecular functional process

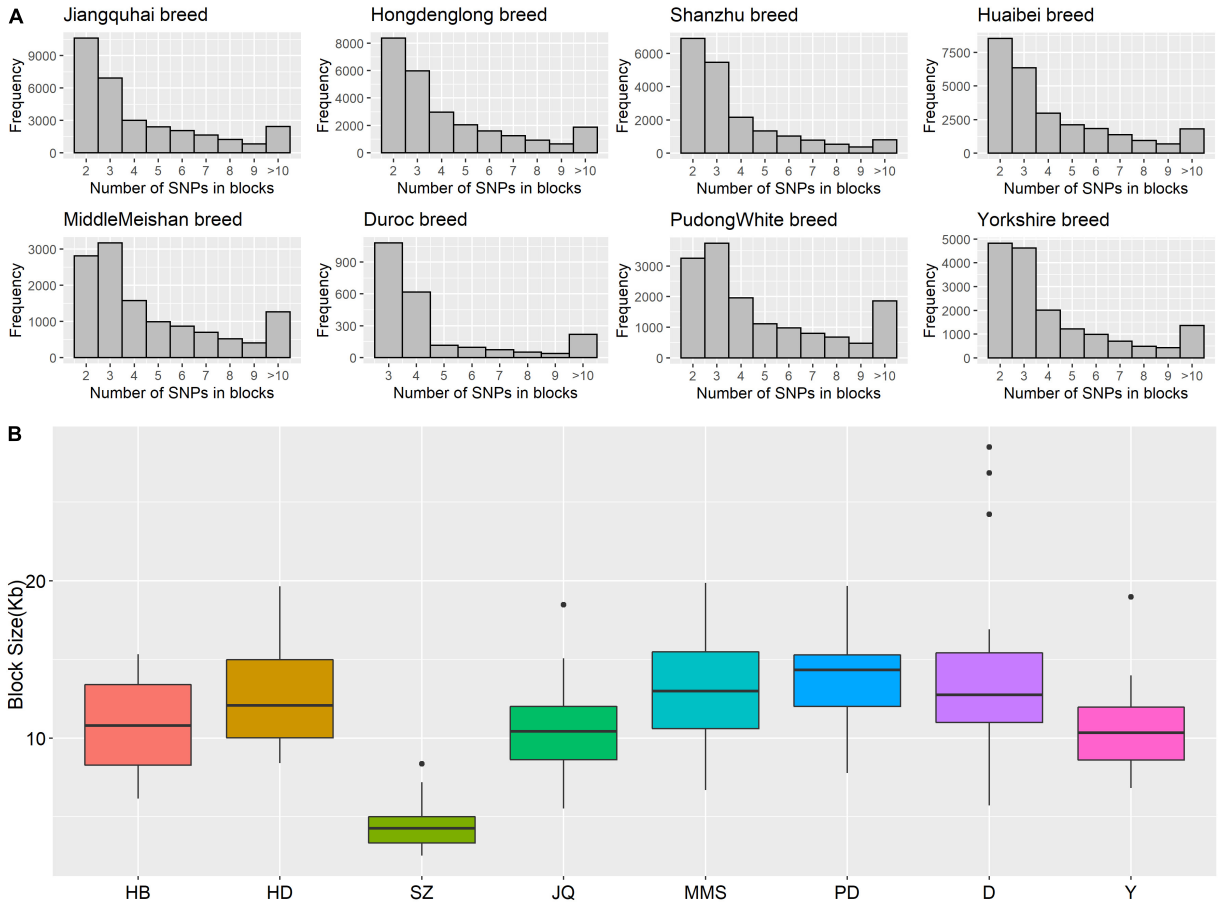


**FIGURE 5 | (A)** Histogram plot showing SNP distribution in haplotype blocks across breeds. Jiangquhai (JQ) had more SNPs clustered into haplotype blocks than other breeds. **(B)** Box plot of haplotype block size distribution in different breeds. Shanzhu (SZ) had the shortest average haplotype block size while PudongWhite (PD) had the longest.

(*p-value* $\leq$ 0.01) associated with activin receptor activity, type I. However, the other 2 genes (*GGT5* and *GGT1*) which are located on SSC14 were enriched in the molecular functional process gamma-glutamyltransferase activity (*p-value* $\leq$ 0.01). These 2 genes (*GGT5* and *GGT1*) were also enriched in the signaling pathway related to ssc00460: Cyanoamino acid metabolism, and ssc00430: Taurine and hypotaurine metabolism (*p-value* $\leq$ 0.02).

## DISCUSSION

The evolutionary history of some pigs in certain regions of developing countries like China is poorly understood. The emergence of new breeds or sub-populations is a result of natural (adaptation) or artificial selection and this selection pressure plays a major role in shaping the genetic architecture and gene pool of extensively raised livestock species (Amaral et al., 2008; Khanyile et al., 2015). Despite recent research on the JQ pig population, there is still a dearth of information on the genetic architecture of economically important breed traits. This study, as one of the first reports on the haplotype block structure in the JQ pig breed, aimed to reveal the effects of selection pressure on its genome. We characterized the pattern of LD in the genome of the breed and detected various QTLs and genes spanned by haplotype blocks. We compared most of our results with the ones obtained in three other breeds from the same province (region) (HB, SZ, and HD breeds); two from Shanghai province (MMS and PD); and two western breeds (D and Y). From this comparison, JQ showed a higher level of variation in block structure and the number of SNPs involved in the block formation.

### Overlapping QTLs and Genes Detected in the JQ Breed

Conservation of animal genetic resources from a global perspective, focuses not only on endangered breeds but also on those that are not well utilized. Locally adapted breeds are always at risk of extinction, particularly when local populations have a preference for imported breeds. Generally, only a small proportion of breeds, particularly in developing countries, are involved in planned genetic improvement programs that aim to ensure efficient and sustainable utilization of these breeds.

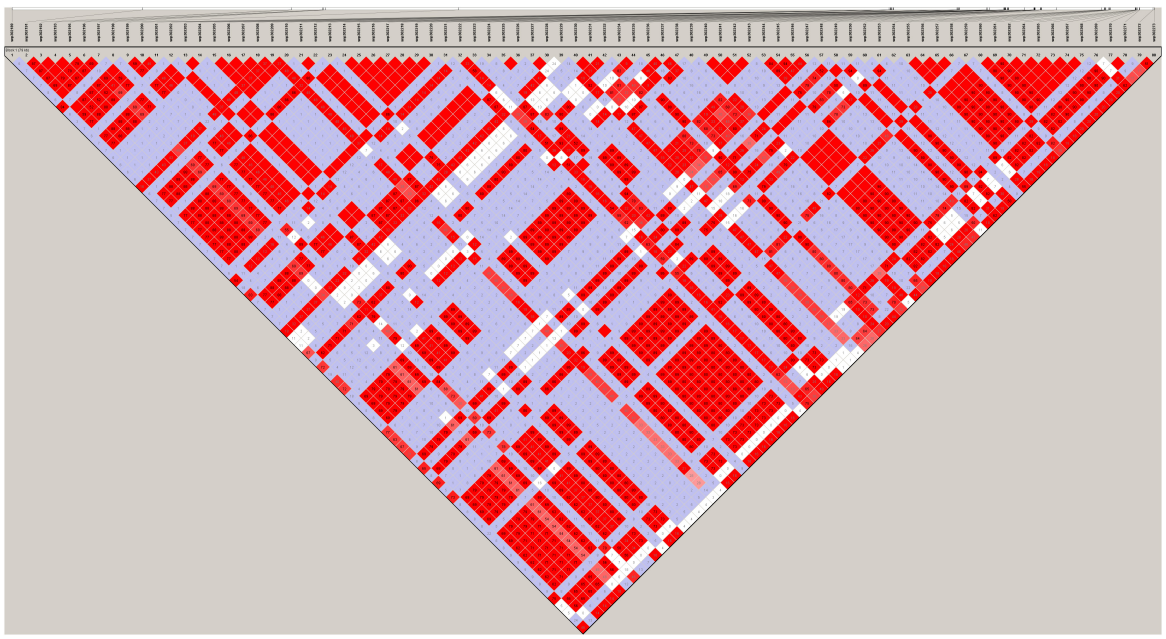

**FIGURE 6 |** Haploview plot of linkage disequilibrium ($r^2$) between SNPs on chromosome 18 of JQ breed. This block is 79.084 kb in size and has the maximum number of SNPs (80) in JQ blocks. Values in the diamond are LD values in percentages and diamonds without a value shows a complete LD ($r^2$ = 1) between SNPs.

Therefore, developing countries like China should ensure that commercial pig strains are developed, while also maintaining the genetic diversity within the purebred population.

In our study, we detected a high percentage of haplotype blocks overlapping QTLs related to meat and carcass quality, and a few related to health. This suggests that these haplotype blocks may be potentially associated with economic traits like side fat thickness, intramuscular fat content, average backfat thickness, cooking loss, meat firmness, lean meat weight, and response to *Mycoplasma hyopneumoniae* in the JQ breed (**Table 4**). Various studies have already reported most of these QTLs in the pig quantitative trait loci (QTL) database. For example, on SSC1, Stratz et al. (2018) reported a highly significant QTL for Dressing percentage (ID = 161054); on SSC6, Le et al. (2017) identified significant QTL for top line conformation (ID = 126140); Choi et al. (2010) detected highly significant QTL for lean meat weight (ID = 16910); Liu et al. (2008) detected significant QTL for average backfat thickness (ID = 5980); and Choi et al. (2011) also detected highly significant QTL for meat firmness (ID = 21367). On SSC17, Stratz et al. (2013) detected significant QTL for pH 24 hr post-mortem (loin) (ID = 21865), while on SSC18, Uddin et al. (2010) identified significant QTL for *Mycoplasma hyopneumoniae* antibody titer (ID = 12330) and changes in *Mycoplasma hyopneumoniae* antibody titer (ID = 12331). Generally, JQ pigs are excellent producers of quality meat, used in the production of Rugao ham, and characterized by large body size and high lean percentage (Toldrá et al., 2014). We believe that this result might aid the further genomic study of meat quality and health-related traits in the JQ breed.

In our gene enrichment analysis, we also detected 4 health-related genes involved in various molecular functions in JQ pigs. These include *ACVRL1*, *ACVR1B*, *GGT5*, and *GGT1* gene. The *ACVRL1* gene is a TGFb/BMP type I receptor that plays a key role in the regulation of endothelial cell proliferation and maintenance of vascular integrity (Tual-Chalot et al., 2014), while *ACVR1B* acts in a paracrine manner on skin epithelial cells to suppress tumorigenesis (Qiu et al., 2011). These two genes also play essential roles in bone growth and morphogenesis (Merino et al., 1999), suggesting a pleiotropic SNP in the haplotype block region harboring these genes (Solovieff et al., 2013; Zhang et al., 2016). Furthermore, *GGT5* has been found to code for a cell surface protein that helps in the hydrolysis of the gamma-glutamyl bond of glutathione and glutathione *S*-conjugates (Wickham et al., 2011). It is expressed by macrophages throughout the body and may play an important role in the immune system (Hanigan et al., 2015). An increase in the expression of *GGT5* has also been found to impair testicular steroidogenesis by deregulating local oxidative stress (Li et al., 2016). The *GGT1* gene plays a major role in cleaving glutathione and its conjugate (Hanigan et al., 2015). In our study, *GGT5* and *GGT1* genes were also found to be enriched in the KEGG term (KEGG: ssc00430) related to Taurine and hypotaurine metabolism. Taurine is known to affect the cholesterol level in the body and can be found in various meat products (Laidlaw et al., 1990; Woollard and Indyk, 1993; Wójcik et al., 2010; Ripps and Shen, 2012). Interestingly, we found that this genomic region (Chromosome 14: 49585363 bp – 49722310 bp) overlapped the QTL that is suggestively associated with cholesterol levels in meat (CHOL) (**Table 4**). This suggests an association of this haplotype block with some meat quality traits in the JQ breed.

Although the haplotype block with the highest number of SNPs in this study was found within the Protein Tyrosine

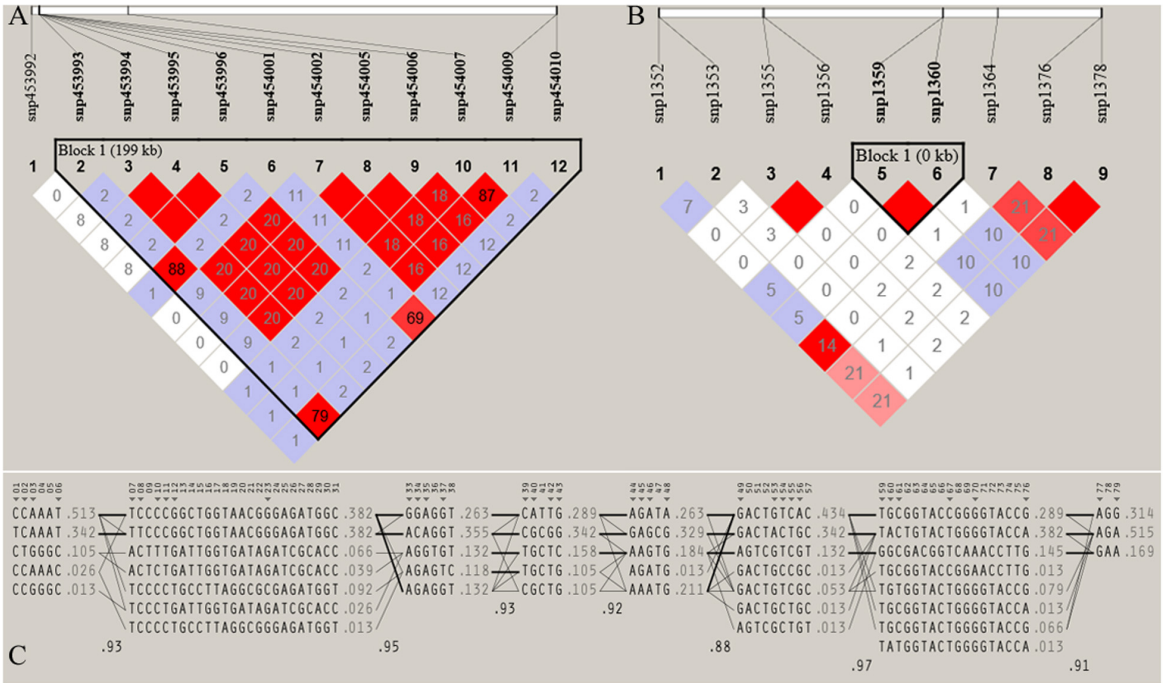


**FIGURE 7 |** Haploview plot of linkage disequilibrium ($r^2$) between SNPs located on **(A)** Chromosome 17 (42665098 bp – 42865069 bp) of JQ breed, with the maximum block length (199.97 kb) and **(B)** chromosome 1 (1936544 bp – 1936545 bp) of the same breed, with the minimum number of SNPs and block size (0.002 kb). This block overlapped the *UNC93A* gene **(C)** Haplotype block structure of chromosome 18 of JQ breed (with the Maximum number of SNPs (80snps). Marker numbers are shown across the top, with highlighted tag SNPs. The population frequencies of each haplotype are shown next to them with lines showing the most common crossings from one block to the next. The thicker lines indicate more common crossings than thinner lines and below the crossing lines is the multi-locus *D* prime between two blocks. Lower *D* prime value indicates a greater amount of historical recombination between two blocks (Barrett et al., 2005).

Phosphatase Receptor Type N2 (*PTPRN2*) gene in the JQ breed, to our surprise, it was not significantly enriched in any pathway or ontology. We infer that there could be more health-related genes in this block region (**Figure 6**) that are yet to be annotated as there were a lot of health-related QTLs, like Mycoplasma hyopneumoniae antibody titer (MHT), Change in Mycoplasma hyopneumoniae antibody titer (MHTC), and Toll-like receptor 9 level (TLR9), spanned by the block (**Table 4**). A previous study had investigated the degree of resistance to *M. hyopneumoniae* in JQ porcine lean strain (JQHPL) and concluded that JQHPL pigs exhibited higher resistance to *M. hyopneumoniae* than the western strains in the study, possibly due to the faster and stronger mucosal immunity phenotype of the strain (Hua et al., 2014). However, we also premise that this haplotype block could be harboring some disease susceptibility traits in the JQ breed, as the average MAF of its SNPs was about 0.20 (ranging from 0.07 to 0.5) (**Supplementary Table S13**). Generally, common variants (MAF > 5%) have been found to contribute to complex diseases more than rare variants (MAF < 1%) (Gibson, 2012; Bomba et al., 2017). Therefore, further study of this genomic region could help in understanding the genomic architecture of complex diseases in the JQ breed and also prevent the incorporation of such haplotype block into selection programs.

In a bid to establish a relationship between the candidate genes and QTLs detected in haplotype blocks, we linked the genomic regions of the detected candidate genes to the corresponding QTL region. Surprisingly, we found that only one gene overlapped these QTL regions. This is because the length of each QTL in these regions is greater than 10 Mb (ranging from 33.11 to 131.16 Mb). We premise this observation on the filtering of all QTLs with length > 10 Mb during our QTL annotation. However, we made some compromises (in QTL length) to enable us to have an overview of the quantitative traits in these genomic regions. In summary, *ACVRL1* and *ACVR1B* gene overlapped QTLs linked to CD4-positive leukocyte percentage (CD4LP), C3c concentration (C3C), and Hemolytic complement activity (alternative pathway) (AH50), while, *GGT5*

**TABLE 3 |** The percentage of common haplotypes shared across populations.

| | JQ | HB | SZ | HD | MMS | PD | D | Y |
|---|---|---|---|---|---|---|---|---|
| **JQ** | **100** | 27.81 | 12.09 | 25.86 | 18.73 | 22.17 | 2.60 | 13.00 |
| **HB** | 31.93 | **100** | 10.65 | 27.04 | 18.16 | 20.74 | 2.73 | 12.95 |
| **SZ** | 45.28 | 34.76 | **100** | 27.46 | 23.10 | 27.72 | 2.34 | 18.12 |
| **HD** | 27.03 | 24.62 | 7.66 | **100** | 16.43 | 18.28 | 2.36 | 12.00 |
| **MMS** | 38.52 | 32.55 | 12.69 | 32.34 | **100** | 31.12 | 2.44 | 12.46 |
| **PD** | 36.21 | 29.51 | 12.09 | 28.57 | 24.72 | **100** | 2.45 | 13.20 |
| **D** | 27.15 | 24.82 | 6.54 | 23.55 | 12.41 | 15.68 | **100** | 15.25 |
| **Y** | 25.44 | 22.09 | 9.48 | 22.48 | 11.86 | 15.82 | 2.86 | **100** |

*Each row represents the percentage of common haplotype blocks between the breed (on the first column) and other breeds across the row.*

**TABLE 4 |** QTLs associated with haplotype block regions in the JQ Breed.

| SSC | No. of Haplotypes | Location (bp) | Size (kb) | No. of SNPs | QTLs |
|---|---|---|---|---|---|
| 1 | 6 | 57754743–57954471 | 199.730 | 9 | DRESS%, LMA, FEEDIN, AFR, SHEAR, FA-C20:1, FA-C18:0 |
| 1 | 17 | 100151813–100351758 | 199.946 | 26 | DRIPL |
| 1 | 17 | 145630575–145788505 | 157.931 | 44 | – |
| 2 | 23 | 150363020–150561606 | 198.587 | 52 | LMA, BLACT |
| 5 | 24 | 17274912–17439428 | 164.517 | 51 | – |
| 6 | 4 | 48308117–48507976 | 199.860 | 5 | TOPLC, IHERN |
| 6 | 8 | 52898408–53098319 | 199.912 | 5 | TOPLC, FEEDIN |
| 6 | 18 | 64203148–64347705 | 144.558 | 55 | IMF, HAPT, C3C, NEUT, LEANWT, EBPC, FIRM, BFT, SHOUFATD, CTISSP, BFS, GLYPO, LMA, COOKL |
| 6 | 27 | 65204715–65346283 | 141.569 | 41 | LEANCUTP, LEANP, DRESS% |
| 6 | 9 | 107929215–108129083 | 199.869 | 12 | FAPC |
| 7 | 20 | 57703699–57797277 | 93.579 | 48 | TNUM |
| 9 | 20 | 41012098–41114458 | 102.361 | 43 | LVNUM, HDL |
| 13 | 7 | 71557016–71756980 | 199.965 | 19 | – |
| 14 | 12 | 47714888–47914794 | 199.907 | 18 | ANDR |
| 14 | 13 | 49585363–49722310 | 136.948 | 43 | TVNUM, CHOL |
| 14 | 9 | 50940408–51140288 | 199.881 | 11 | SCF, AGEP, RTNUM, TNUM, TNUMD |
| 14 | 21 | 51322222–51522048 | 199.827 | 33 | PLTCT |
| 17 | 8 | 42665098–42865069 | 199.972 | 11 | ACTH2, 34RIBBFT |
| 17 | 13 | 62089019–62255688 | 166.670 | 55 | pH |
| 18 | 23 | 923784–1002867 | 79.084 | 80 | MHT, MHTC, TLR9, DIAMF, FIB1DIAM, FIB2ADIAM, LIVWT, FEEDCON, ADG, BW, WWT, NBA, TNB |

*List of full names of QTLs: DRESS%, dressing percentage; LMA, loin muscle area; FEEDIN, daily feed intake; AFR, average feeding rate; SHEAR, Shear force; FA-C20:1, cis-11-Eicosenoic acid content; FA-C18:0, stearic acid content; DRIPL, drip loss; BLACT, lactate level; TOPLC, top line conformation; IMF, intramuscular fat content; HAPT, haptoglobin concentration; C3C, C3c concentration; NEUT, neutrophil count; IHERN, scrotal/inguinal hernia; LEANWT, lean meat weight; EBPC, empty body protein content; FIRM, firmness; BFT, average backfat thickness; SHOUFATD, shoulder subcutaneous fat thickness; CTISSP, connective tissue protein; BFS, side fat thickness; GLYPO, average glycolytic potential; LMA, loin muscle area; COOKL, cooking loss; LEANCUTP, lean cuts percentage; LEANP, lean meat percentage; FAPC, fat area percentage in carcass; TNUM, teat number; LVNUM, lumbar vertebra number; HDL, HDL cholesterol; TVNUM, thoracic vertebra number; AGEP, age at puberty; TNUM, teat number; TNUMD, teat number, difference between sides; RTNUM, right teat number; PLTCT, platelet count; SCF, backfat between 3rd and 4th last ribs; ANDR, androstenone, laboratory; CHOL, cholesterol level in meat; ACTH2, post-stress ACTH level; 34RIBBFT, backfat thickness between 3rd and 4th rib; pH, pH 24 h post-mortem (loin); TLR9, toll-like receptor 9 level; DIAMF, diameter of muscle fibers; FIB1DIAM, diameter of type I muscle fibers; FIB2ADIAM, diameter of type IIa muscle fibers; BW, body weight (birth); FEEDCON, feed conversion ratio; LIVWT, liver weight; ADG, average daily gain; WWT, body weight (weaning); MHT, mycoplasma hyopneumoniae antibody titer; MHTC, change in mycoplasma hyopneumoniae antibody titer; TNB, litter size; NBA, total number born alive.*

**TABLE 5 |** Candidate Genes detected in the haplotype blocks of JQ pig population.

| SSC | Haplotype block position (Mb) | ID | Term | *P*-value | Candidate Genes |
|---|---|---|---|---|---|
| 5 | 17.275–17.439 | GO:0016361 | Activin receptor activity, type I | 0.009 | *ACVRL1*, *ACVR1B* |
| 14 | 49.585–49.722 | GO:0003840 | Gamma-glutamyltransferase activity | 0.009 | *GGT5*, *GGT1* |
| | | KEGG:ssc00460 | Cyanoamino acid metabolism | 0.011 | *GGT5*, *GGT1* |
| | | KEGG:ssc00430 | Taurine and hypotaurine metabolism | 0.017 | *GGT5*, *GGT1* |

and *GGT1* gene overlapped QTLs linked to Interferon-gamma to interleukin-10 ratio (IFNGIL10), Calcium level (BCAL), Creatinine level (CREAT), Potassium level (BPOTASS), C3c concentration (C3C), Haptoglobin concentration (HAPT), and Melanoma susceptibility (MELAN). This suggests an association of the haplotype blocks in these genomic regions with health-related traits in the JQ breed.

## The Extent of Linkage Disequilibrium in the Pigs and Application in GWAS

A full understanding of the LD properties in domesticated animals like pigs is of importance because it underlies all forms of genetic mapping (Nordborg and Tavaré, 2002) and can be used for fine mapping genes associated with complex diseases in pigs. To increase the power of SNP-based association studies (GWAS), the extent of LD in a breed must be considered. A knowledge of this can be used to predict the average number of markers required in quantitative trait association studies (GWAS).

In this study, we looked at the extent of LD in the JQ breed and compared it to the one obtained in other breeds. We used $r^2$ value as a measure of LD between each locus of a chromosome. Generally, we observed a lower LD level, at larger SNP distances, on both autosomes and SSCX of JQ breed. A similar result was also found in the SZ breed (**Figure 3** and **Supplementary Figure S1**). However, our result contradicts that of Xu et al. (2019), which reported a higher LD extent greater

than 0.3, at SNPs distance of 99.66 kb in the JQ breed using Porcine 80 K SNP chips. This difference might be due to the larger sample size, density, and type of SNP data used in the study (as reviewed by Qanbari, 2020). Generally, SNP chips (or genotyping array) data tends to underrepresent rare variants that are likely to be detected in sequence data like the one used in this study. Since the extent of LD ($r^2$) depends on MAF, it is expected that there would be a little difference in the $r^2$ value obtained from both studies, partly due to SNP ascertainment bias on SNP chip data (Lachance and Tishkoff, 2013; Geibel et al., 2019). This kind of bias was reduced in a recent study by Huang et al. (2020) which had more Chinese breeds represented in the design of the SNP array used. The study reported an LD ($r^2 > 0.3$) at SNPs-distance of 36.10 kb for the JQ breed. Moreover, using the MUC4 (Mucin 4, Cell Surface Associated) gene sequences, Yang et al. (2012) also reported that $r^2 > 0.3$ extended up to 20 kb distance in the JQ breed, validating to some extent, the reliability of the LD value obtained in our study.

To assess the differences in LD extent on the autosome and SSCX across breeds, we predicted the extent of LD decay for different genomic distances. The result showed a noticeable difference in the LD extent across breed, especially on the SSCX (**Supplementary Figure S1**). As expected, we observed a longer LD extent on the SSCX compared to the autosome (Schaffner, 2004). Generally, the SSCX is known to have a low recombination rate and tends to preserve demographic events longer than the autosome (Schaffner, 2004; Laan et al., 2005). Among the Chinese pigs in this study, we observed a longer LD extent on the SSCX of the PD breed, suggesting that this breed might have recently evolved or experienced a bottleneck. This result is also in line with our admixture analysis which had the lowest cross-validation error when $K = 7$ (**Figure 1B**), before PD separated from MMS into a different cluster (**Figure 1C**). This result might also be useful in mapping sex-related traits in the Chinese pigs in our study.

According to previous studies, a mean $r^2 \geq 0.3$ is considered as a strong LD sufficient for QTL mapping (Farnir et al., 2000). However, to detect a QTL in GWAS and estimate the genomic breeding value (GEBV) of an animal, an average $r^2$ of at least 0.2 is required to achieve power and accuracy $\geq 0.8$ (Meuwissen et al., 2001; Meuwissen et al., 2001). In our study, we found that moderate LD ($r^2 \geq 0.2$) extended up to 500–1000 kb in HD (0.23), D (0.29), and Y (0.26) breeds (**Figure 3**). This suggests that an association study performed within these breeds using an average inter-marker $r^2 \geq 0.20$ would require about 12,000 SNPs. Although the average $r^2$ (for bin distance 500–1000 kb), reported for the D and Y in our study was higher compared to the one reported by Grossi et al. (2017) using a 60 K SNP panel (0.23 and 0.17 for D and Y, respectively), we found that the average $r^2$ for the autosomes of these breeds is still comparable to a previous study that used larger sample sizes (>100) (Badke et al., 2012). These differences in LD could be attributed to population structure, selection, sample size or density of the markers used in the study. On the other hand, $r^2 \geq 0.20$ extended only up to 0–10 kb in the SZ breed, and 10–20 kb in JQ (**Figure 3**), indicating that about 120,000 to 240,000 SNPs would be required for effective GWAS in these breeds. For the HB breed, the average inter-marker $r^2$ extended up to 40–60 kb, meaning that about 40,000 to 60,000 SNPs would be needed for GWAS. While for MMS and PD, $r^2$ extended up to 60–100 kb distances, and about 24,000–40,000 evenly spaced SNPs would be sufficient for a successful association study in the breeds. This result could be particularly useful in designing breed-specific SNP array panels for future genomic study and selection programs for these pig breeds.

## Haplotype Block Structure

The ability of an animal to survive in a changing environment, and also keep up with changes in selection preference, depends on the richness (genetic diversity) of its gene pool. This can be affected by several occurrences, including natural and artificial selection. There are various parameters for measuring genetic diversity in a population, including population-gene-frequency based statistics like average expected heterozygosity (He), the proportion of polymorphic loci (Pn), and allelic richness (Ar). However, alternative statistics based on allelic diversity (i.e., number of different allele types present at a locus) can also provide insight into the genetic diversity in a population and be a better predictor of long-term adaptation and total response to selection in an unpredictable future scenario (Caballero and García-Dorado, 2013; Vilas et al., 2015). Therefore, our study examined haplotype diversity as a measure of genetic diversity across breeds, since haplotypes are multi-allelic markers and can be treated as an allele in a haplotype-based study. To our surprise, we found that D pigs had higher haplotype diversity than the other Chinese pig breeds in this study (**Table 1**) despite having experienced high selection pressure in the past. Interestingly, among the Jiangsu province pigs in our study, the SZ breed had the highest haplotype diversity compared to JQ, HB, and HD (**Table 1**). Its haplotype diversity (0.483) (**Table 1**) was similar to that of Y with which it shared the highest haplotype block (18.12%) in comparison with other breeds (**Table 3**). This result is in agreement with our previous research (Zhang et al., 2018), which suggested that Y might have been used to improve the SZ breed and that the SZ breed might have originated from different genomic sources, therefore increasing the diversity of haplotypes in the breed's genome. Furthermore, the higher haplotype diversity observed in the highly selected western pig breeds in our study could be because the sample size was selected from a larger population of western breeds (more than 1,000) in China, and a limited population of Jiangsu pig breeds (about 140 individuals) kept in the conservation pig farm in Jiangsu province, China. Therefore, we can also infer that the diversity of haplotypes in a population can be influenced by its population size. Moreover, recent studies had reported ongoing selection processes in Chinese pig breeds (Quan et al., 2020), which calls for the strategic management of these breeds to prevent the loss of important traits or genes. Although many studies had reported high genetic diversity in the JQ breed (Fan et al., 2002; Xu et al., 2019), most of these reported metrics do not perfectly reflect the ability of the breed to cope with a future unexpected change in breeding preference or disease outbreak. Therefore, the lower haplotype diversity observed in

the JQ breed could be an indicator of a reduction in their genetic diversity, indicating a need for proper management of the JQ population in conservation pig farming. This result is also in line with the findings of Quan et al. (2020), which reported a lower haplotype diversity (0.752) in the mtDNA sequences of the JQ breed compared to that of Duroc (0.794), Yorkshire (0.837), and Meishan pigs (0.811). In line with our admixture and haplotype sharing result (**Figure 1C** and **Table 3**), the conservation farms could design breeding programs that restrict the level of contribution of the highly admixed JQ individuals to the next generation of the breed.

Furthermore, our block analysis revealed that the JQ breed had more SNPs that were clustered into haplotype blocks than any other breed in this study (**Figure 5A**), an indication that the breed contains a lot of variants that are inherited in the form of haplotype blocks. This could improve the fine mapping of QTLs and association studies in this indigenous pig breed. Our result also revealed that the JQ breed had a moderate block size (**Figure 5B**), which implies that the breed could have undergone moderate selection. This is also in line with the extent of LD observed in the breed (**Figure 3**). Besides, the smaller average haplotype block size observed in the SZ breed could be the result of historical admixture with western pig breeds, which is strongly supported by our Neighbor-Net tree (**Figure 1A**), admixture analysis (**Figure 1C**), and haplotype block sharing result (**Table 3**). Although our results showed variations in haplotype block structure across breeds, we were unable to detect some known haplotype blocks on the SSCX of each breed (Reimer et al., 2018) (shown in **Supplementary Figure S7**). This might be because the density of the SNPs on the SSCX (**Supplementary Figures S2, S3**) was not enough to track these SNPs and the resulting haplotype blocks.

The variation in the number of unique haplotype blocks within a population can shed more light on the independent genomic sub-structuring and evolution of such populations (Khanyile et al., 2015). Moreover, haplotype sharing allows the generational transfer of genomic materials between breeds (Khanyile et al., 2015). Our shared and unique haplotype block result showed different variations across breeds and reveals the level of uniqueness of each breed in our study. To strategically manage a pig population for conservation purposes, it is necessary to perform a SWOT (Strength, Weakness, Opportunity, and Threat) analysis of each breed. The strengths of a breed might be, for example, its genetic uniqueness, its adaptation to a particular system of production, or its past and present function in human culture (Food and Agriculture Organization, 2013). JQ pigs are generally well adapted to their local environment while they serve as a major source of meat for ham production in China (Toldrá et al., 2014). Breed-specific haplotype blocks in this pig could be considered a useful tool in characterizing and protecting its genetic diversity, as they potentially indicate a genomic source of unique phenotypic characters in the breed. Consistent with our admixture result (**Figure 1C**), a higher percentage of JQ blocks was (found in or) shared with D (**Table 3**), suggesting that the JQ breed might have experienced high introgression from D breed. This result is also in line with the known history of the breed. Generally, Chinese pigs, despite their superior meat quality and high prolificacy, have a slower growth rate compared to western pig breeds, and local pig farmers tend to supplement this by crossing indigenous pig breeds with commercial lines. However, if this is not properly managed in the JQ population, it might lead to a complete genetic erosion in the breed and also minimize their natural ability to adapt to local environmental stresses and disease outbreak.

The present study confirms previous findings (Fan et al., 2002; Xiao et al., 2017b; Xu et al., 2019) and contributes additional evidence that suggests that the JQ breed is indeed an important genetic resource. However, continuous genetic erosion and decline in the population, increases the risk of losing some economically important traits of the breed. We believe our result has provided more information that could guide the development of breeding programs to ensure the conservation and utilization of this genetic resource. In addition, the SZ breed showed the highest level of introgression from the Y breed and might be at the point of losing its genetic uniqueness. This indicates the weakness of this breed and should be taken into consideration when planning conservation programs in the future.

## CONCLUSION

We analyzed the LD and haplotype block structure of the JQ pig breed and also detected some underlying QTLs and genes spanned by these blocks. The present study revealed some blocks that might be associated with some quantitative or adaptive traits of the JQ pig and also provides useful information that could contribute to more informed, strategic management decisions in conserving and utilizing this breed. This result might also be useful in selecting variants for further association studies of these traits. We also reported a high level of introgression of Y haplotypes into the SZ pig breed and concluded that the later breed might be at the point of losing its genetic uniqueness.

## DATA AVAILABILITY STATEMENT

All SNP data in the present study can be found in the FigShare Repository: doi: 10.6084/m9.figshare.11984010.

## ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of Shanghai Jiao Tong University (contract no. 2011-0033).

## AUTHOR CONTRIBUTIONS

YP and QW conceived and supervised the study. FO analyzed the data and wrote the manuscript while other authors helped FO during the analysis and also revised the manuscript. All the authors revised and approved the manuscript.

## FUNDING

This study was supported by the National Natural Science Foundation of China (grant nos. 31672386, 31872321, and 31772552) and the Interdisciplinary Program of Shanghai Jiao Tong University (ZH2018ZDA16).

## ACKNOWLEDGMENTS

We would like to thank Mr. S. O. Jimoh (Grassland Research Institute, Chinese Academy of Agricultural Sciences, Hohhot, China), who reviewed the first draft of this manuscript. Also, our special thanks to Dr. Mirte Bosse (Wageningen University and Research), Dr. Arsen Dotsev (L.K. Ernst Federal Science Center for Animal Husbandry, Podolsk, Russia), and the reviewers whose suggestions and contribution have gone a long way to improve this work.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00752/full#supplementary-material

## REFERENCES

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PLoS One* 8:e56001. doi: 10.1371/journal.pone.0056001

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., and Steibel, J. P. (2012). Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* 13:24. doi: 10.1186/1471-2156-13-24

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

Bohmanova, J., Sargolzaei, M., and Schenkel, F. S. (2010). Characteristics of linkage disequilibrium in North American holsteins. *BMC Genomics* 11:421. doi: 10.1186/1471-2156-13-421

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* 18:77.

Bosse, M., Megens, H. J., Frantz, L. A. F., Madsen, O., Larson, G., Paudel, Y., et al. (2014). Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat. Commun.* 5:5392.

Bourgonje, A. M., Verrijp, K., Schepens, J. T. G., Navis, A. C., Piepers, J. A. F., Palmen, C. B. C., et al. (2016). Comprehensive protein tyrosine phosphatase mRNA profiling identifies new regulators in the progression of glioma. *Acta Neuropathol. Commun.* 4:96.

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

Caballero, A., and García-Dorado, A. (2013). Allelic diversity and its implications for the rate of adaptation. *Genetics* 195, 1373–1384. doi: 10.1534/genetics.113.158410

Calus, M. P. L., Meuwissen, T. H. E., De Roos, A. P. W., and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553–561. doi: 10.1534/genetics.107.080838

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Chen, Q., Ma, Y., Yang, Y., Chen, Z., Liao, R., Xie, X., et al. (2013). Genotyping by genome reducing and sequencing for outbred animals. *PLoS One* 8:e67500. doi: 10.1371/journal.pone.0067500

Chen, Z., Yao, Y., Ma, P., Wang, Q., and Pan, Y. (2018). Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins. *PLoS One* 13:e0192695. doi: 10.1371/journal.pone.00192695

China National Commission of Animal Genetic Resources (2011). *Animal Genetic Resources in China Pigs*. China: China Agriculture Press.

Choi, I., Steibel, J. P., Bates, R. O., Raney, N. E., Rumph, J. M., and Ernst, C. W. (2010). Application of alternative models to identify QTL for growth traits in an F2Duroc x Pietrain pig resource population. *BMC Genet.* 11:97. doi: 10.1186/1471-2156-13-97

Choi, I., Steibel, J. P., Bates, R. O., Raney, N. E., Rumph, J. M., and Ernst, C. W. (2011). Identification of carcass and meat quality QTL in an F2 Duroc × Pietrain pig resource population using different least-squares analysis models. *Front. Genet.* 2:18. doi: 10.3389/fgene.2015.00018

Corbin, L. J., Blott, S. C., Swinburne, J. E., Vaudin, M., Bishop, S. C., and Woolliams, J. A. (2010). Linkage disequilibrium and historical effective population size in the thoroughbred horse. *Animal Genetics* 41(Suppl. 2), 8–15. doi: 10.1111/j.1365-2052.2010.02092.x

Crnokrak, P., and Roff, D. A. (1995). Dominance variance: associations with selection and fitness. *Heredity* 75, 530–540. doi: 10.1038/hdy.1995.169

Cuyabano, B. C. D., Su, G., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47:61.

Fan, B., Wang, Z. G., Li, Y. J., Zhao, X. L., Liu, B., Zhao, S. H., et al. (2002). Genetic variation analysis within and among Chinese indigenous swine populations using microsatellite markers. *Anim. Genet.* 33, 422–427. doi: 10.1046/j.1365-2052.2002.00898.x

FAOSTAT (2017). Available online at: http://www. fao. org/faostat/en/# data QC (accessed May 2019). doi: 10.1046/j.1365-2052.2002.00898.x

Farnir, F., Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., et al. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.* 10, 220–227. doi: 10.1101/gr.10.2.220

Food and Agriculture Organization (2013). *In Vivo Conservation of Animal Genetic Resources*. Rome: FAO.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424

Geibel, J., Reimer, C. H. U. W., Weigend, S., Weigend, A., Pook, T., and Simianer, H. (2019). How array design affects SNP ascertainment bias. *bioRxiv* [Preprint],

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Grossi, D. A., Jafarikia, M., Brito, L. F., Buzanskas, M. E., Sargolzaei, M., and Schenkel, F. S. (2017). Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. *BMC Genet.* 18:6. doi: 10.1186/1471-2156-13-06

Guryev, V., Smits, B. M. G., Van De Belt, J., Verheul, M., Hubner, N., and Cuppen, E. (2006). Haplotype block structure is conserved across mammals. *PLoS Genet.* 10:19. doi: 10.1371/journal.pone.0000019

Hanigan, M. H., Gillies, E. M., Wickham, S., Wakeham, N., and Wirsig-Wiechmann, C. R. (2015). Immunolabeling of gamma-glutamyl transferase 5 in normal human tissues reveals that expression and localization differ from

gamma-glutamyl transferase 1. *Histochem. Cell Biol.* 143, 505–515. doi: 10.1007/s00418-014-1295-x

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13, 635–643. doi: 10.1101/gr.387103

Heifetz, E. M., Fulton, J. E., O'Sullivan, N., Zhao, H., Dekkers, J. C. M., and Soller, M. (2005). Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* 171, 1173–1181. doi: 10.1534/genetics.105.040782

Hua, L. Z., Wu, Y. Z., Bai, F. F., William, K. K., Feng, Z. X., Liu, M. J., et al. (2014). Comparative analysis of mucosal immunity to *Mycoplasma hyopneumoniae* in *Jiangquhai porcine* lean strain and DLY piglets. *Genet. Mol. Res.* 13, 5199–5206. doi: 10.4238/2014.july.7.13

Huang, D. W., Sherman, B. T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., et al. (2009). Extracting biological meaning from large gene lists with DAVID. *Curr. Protoc. Bioinform.* 27, 13.11.1–13.11.13.

Huang, M., Yang, B., Chen, H., Zhang, H., Wu, Z., Ai, H., et al. (2020). The fine-scale genetic structure and selection signals of Chinese indigenous pigs. *Evol. Appl.* 13, 458–475. doi: 10.1111/eva.12887

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114.

Keenan, K., Mcginnity, P., Cross, T. F., Crozier, W. W., and Prodöhl, P. A. (2013). DiveRsity: an R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol. Evol.* 4, 782–788. doi: 10.1111/2041-210x.12067

Khanyile, K. S., Dzomba, E. F., and Muchadeyi, F. C. (2015). Haplo-block structure of Southern African village chicken populations inferred using genome-wide SNP data. *Genet. Mol. Res.* 14, 12276–12287. doi: 10.4238/2015.october.9.16

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi: 10.1111/1755-0998.12387

Laan, M., Wiebe, V., Khusnutdinova, E., Remm, M., and Pääbo, S. (2005). X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations. *Eur. J. Hum. Genet.* 13, 452–462. doi: 10.1038/sj.ejhg.5201340

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, 780–786. doi: 10.1002/bies.201300014

Laidlaw, S. A., Grosvenor, M., and Kopple, J. D. (1990). The taurine content of common foodstuffs. *J. Parenter. Enter. Nutr.* 14, 183–188. doi: 10.1177/014860719001400218

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

Le, T. H., Christensen, O. F., Nielsen, B., and Sahana, G. (2017). Genome-wide association study for conformation traits in three Danish pig breeds. *Genet. Sel. Evol.* 49:12. doi: 10.1186/s12711-017-0289-2.

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

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

Li, W., Wu, Z. Q., Zhang, S., Cao, R., Zhao, J., Sun, Z. J., et al. (2016). Augmented expression of gamma-glutamyl transferase 5 (GGT5) impairs testicular steroidogenesis by deregulating local oxidative stress. *Cell Tissue Res.* 366, 467–481. doi: 10.1007/s00441-016-2458-y

Lin, S., Zhao, H., Zhang, Y., and Niu, T. (2009). "Haplotype structure," in *Handbook on Analyzing Human Genetic Data*, (Berlin: Springer), 25–79. doi: 10.1007/978-3-540-69264-5_2

Liu, G., Kim, J. J., Jonas, E., Wimmers, K., Ponsuksili, S., Murani, E., et al. (2008). Combined line-cross and half-sib QTL analysis in duroc-pietrain population. *Mamm. Genome* 19, 429–438. doi: 10.1007/s00335-008-9132-y

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 981–994. doi: 10.1038/nrg1226

Meng, C., Su, L., Li, Y., Zhu, Q., Li, J., Wang, H., et al. (2018). Different susceptibility to porcine reproductive and respiratory syndrome virus infection among Chinese native pig breeds. *Arch. Virol.* 163, 2155–2164. doi: 10.1007/s00705-018-3821-y

Merino, R., Macias, D., Gañan, Y., Rodriguez-Leon, J., Economides, A. N., Rodriguez-Esteban, C., et al. (1999). Control of digit formation by activin signalling. *Development* 126, 2161–2170.

Meuwissen, T. H., and Goddard, M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155, 421–430.

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.

Miao, Z. G., Wang, L. J., Xu, Z. R., Huang, J. F., and Wang, Y. R. (2009). Developmental changes of carcass composition, meat quality and organs in the Jinhua pig and Landrace. *Animal* 2009, 468–473. doi: 10.1017/s1751731108003613

Nordborg, M., and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18, 83–90. doi: 10.1016/s0168-9525(02)02557-x

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., et al. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* 33, 382–387. doi: 10.1038/ng1100

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., and Bender, D. (2007). PLINK: a tool set for whole genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:95.

Qanbari, S. (2020). On the extent of linkage disequilibrium in the genome of farm animals. *Front. Genet.* 10:1304. doi: 10.3389/fgene.2015.01304

Qiu, W., Li, X., Tang, H., Huang, A. S., Panteleyev, A. A., Owens, D. M., et al. (2011). Conditional activin receptor Type 1B (Acvr1b) knockout mice reveal hair loss abnormality. *J. Invest. Dermatol.* 131, 1067–1076. doi: 10.1038/jid.2010.400

Quan, J., Gao, C., Cai, Y., Ge, Q., Jiao, T., and Zhao, S. (2020). Population genetics assessment model reveals priority protection of genetic resources in native pig breeds in China. *Glob. Ecol. Conserv.* 21:e00829. doi: 10.1016/j.gecco.2019.e00829

Reimer, C., Rubin, C. J., Sharifi, A. R., Ha, N. T., Weigend, S., Waldmann, K. H., et al. (2018). Analysis of porcine body size variation using re-sequencing data of miniature and large pigs. *BMC Genomics* 19:687. doi: 10.1186/1471-2156-13-687

Ripps, H., and Shen, W. (2012). Review: taurine: a "very essential" amino acid. *Mol. Vis.* 18, 2673–2686.

Salem, M. M. I., Thompson, G., Chen, S., Beja-Pereira, A., and Carvalheira, J. (2018). Linkage disequilibrium and haplotype block structure in Portuguese holstein cattle. *CZECH J. Anim. Sci.* 63, 61–69. doi: 10.17221/56/2017-cjas

Schaffner, S. F. (2004). The X chromosome in population genetics. *Nat. Rev. Genet.* 5, 43–51. doi: 10.1038/nrg1247

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Sorokin, A. V., Nair, B. C., Wei, Y., Aziz, K. E., Evdokimova, V., Hung, M. C., et al. (2015). Aberrant expression of proPTPRN2 in cancer cells confers resistance to apoptosis. *Cancer Res.* 75, 1846–1858. doi: 10.1158/0008-5472.can-14-2718

Stratz, P., Baes, C., Rückert, C., Preuss, S., and Bennewitz, J. (2013). A two-step approach to map quantitative trait loci for meat quality in connected porcine F2 crosses considering main and epistatic effects. *Anim. Genet.* 44:360.

Stratz, P., Schmid, M., Wellmann, R., Preuß, S., Blaj, I., Tetens, J., et al. (2018). Linkage disequilibrium pattern and genome-wide association mapping for meat traits in multiple porcine F2 crosses. *Anim. Genet.* 49, 403–412. doi: 10.1111/age.12684

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2:e90011-16.

Sved, J. A., and Feldman, M. W. (1973). Correlation and probability methods for one and two loci. *Theor. Popul. Biol.* 9, 321–340.

Toldrá, F., Hui, Y. H., Astiasarán, I., Sebranek, J. G., and Talon, R. (2014). *Handbook of Fermented Meat and Poultry: Second Edition*. Hoboken, NJ: Wiley, doi: 10.1002/9781118522653

Tual-Chalot, S., Mahmoud, M., Allinson, K. R., Redgrave, R. E., Zhai, Z., Oh, S. P., et al. (2014). Endothelial depletion of Acvrl1 in mice leads to arteriovenous malformations associated with reduced endoglin expression. *PLoS One* 9:e98646. doi: 10.1371/journal.pone.098646

Uddin, M. J., Grosse-Brinkhaus, C., Cinar, M. U., Jonas, E., Tesfaye, D., Tholen, E., et al. (2010). Mapping of quantitative trait loci for mycoplasma and tetanus antibodies and interferon-gamma in a porcine F2 Duroc × Pietrain resource population. *Mamm. Genome* 21, 409–418. doi: 10.1007/s00335-010-9269-3

Vilas, A., Pérez-Figueroa, A., Quesada, H., and Caballero, A. (2015). Allelic diversity for neutral markers retains a higher adaptive potential for quantitative traits than expected heterozygosity. *Mol. Ecol.* 24, 4419–4432. doi: 10.1111/mec.13334

Wall, J. D., and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 4, 587–597. doi: 10.1038/nrg1123

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* 71, 1227–1234. doi: 10.1086/344398

Wang, Z., Chen, Q., Yang, Y., Liao, R., Zhao, J., Zhang, Z., et al. (2015). Genetic diversity and population structure of six Chinese indigenous pig breeds in the Taihu Lake region revealed by sequencing data. *Anim. Genet.* 46, 697–701.

Weir, B., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.

Wickham, S., West, M. B., Cook, P. F., and Hanigan, M. H. (2011). Gamma-glutamyl compounds: Substrate specificity of gamma-glutamyl transpeptidase enzymes. *Anal. Biochem.* 414, 208–214.

Wójcik, O. P., Koenig, K. L., Zeleniuch-Jacquotte, A., Costa, M., and Chen, Y. (2010). The potential protective effects of taurine on coronary heart disease. *Atherosclerosis* 208:19.

Woollard, D. C., and Indyk, H. E. (1993). The determination and distribution of taurine in dairy products. *Food Chem.* 46, 429–437.

Xiao, Q., Zhang, Z., Sun, H., Wang, Q., and Pan, Y. (2017a). Pudong white pig: a unique genetic resource disclosed by sequencing data. *Animal* 11, 1117–1124.

Xiao, Q., Zhang, Z., Sun, H., Yang, H., Xue, M., Liu, X., et al. (2017b). Genetic variation and genetic structure of five Chinese indigenous pig populations in Jiangsu Province revealed by sequencing data. *Anim. Genet.* 48, 596–599.

Xu, P., Wang, X., Ni, L., Zhang, W., Lu, C., Zhao, X., et al. (2019). Genome-wide genotyping uncovers genetic diversity, phylogeny, signatures of selection, and population structure of Chinese Jiangquhai pigs in a global perspective. *J. Anim. Sci.* 97, 1491–1500.

Yang, M., Yang, B., Yan, X., Ouyang, J., Zeng, W., Ai, H., et al. (2012). Nucleotide variability and linkage disequilibrium patterns in the porcine MUC4 gene. *BMC Genet.* 13:57. doi: 10.1186/1471-2156-13-57

Yin, T., Cook, D., and Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* 13:R77.

Zhang, Z., Wang, Z., Yang, Y., Zhao, J., Chen, Q., Liao, R., et al. (2016). Identification of pleiotropic genes and gene sets underlying growth and immunity traits: a case study on Meishan pigs. *Animal* 10, 550–557.

Zhang, Z., Xiao, Q., Zhang, Q., Sun, H., Chen, J., Li, Z., et al. (2018). Genomic analysis reveals genes affecting distinct phenotypes among different Chinese and western pig breeds. *Sci. Rep.* 8:13352.

**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: 15 January 2021
doi: 10.3389/fgene.2020.531218

# Hierarchical Modelling of Haplotype Effects on a Phylogeny

*Maria Lie Selle*[1]*, *Ingelin Steinsland*[1], *Finn Lindgren*[2], *Vladimir Brajkovic*[3], *Vlatka Cubric-Curik*[3] and *Gregor Gorjanc*[4]

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [2] School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom, [3] Department of Animal Science, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia, [4] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom

We introduce a hierarchical model to estimate haplotype effects based on phylogenetic relationships between haplotypes and their association with observed phenotypes. In a population there are many, but not all possible, distinct haplotypes and few observations per haplotype. Further, haplotype frequencies tend to vary substantially. Such data structure challenge estimation of haplotype effects. However, haplotypes often differ only due to few mutations, and leveraging similarities can improve the estimation of effects. We build on extensive literature and develop an autoregressive model of order one that models haplotype effects by leveraging phylogenetic relationships described with a directed acyclic graph. The phylogenetic relationships can be either in a form of a tree or a network, and we refer to the model as the haplotype network model. The model can be included as a component in a phenotype model to estimate associations between haplotypes and phenotypes. Our key contribution is that we obtain a sparse model, and by using hierarchical autoregression, the flow of information between similar haplotypes is estimated from the data. A simulation study shows that the hierarchical model can improve estimates of haplotype effects compared to an independent haplotype model, especially with few observations for a specific haplotype. We also compared it to a mutation model and observed comparable performance, though the haplotype model has the potential to capture background specific effects. We demonstrate the model with a study of mitochondrial haplotype effects on milk yield in cattle. We provide R code to fit the model with the INLA package.

**Keywords: genealogy, haplotype, DAG, autoregression, INLA, Bayesian**

OPEN ACCESS

***Edited by:***
*Gábor Mészáros, University of Natural Resources and Life Sciences Vienna, Austria*

***Reviewed by:***
*Yongzhen Huang, Northwest A and F University, China*
*Maulana Naji, BOKU-University of Natural Resources and Life Sciences Vienna, Austria*

***Correspondence:**
*Maria Lie Selle*
*maria.selle@ntnu.no*

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

***Received:*** *31 January 2020*
***Accepted:*** *15 December 2020*
***Published:*** *15 January 2021*

***Citation:***
*Selle ML, Steinsland I, Lindgren F, Brajkovic V, Cubric-Curik V and Gorjanc G (2021) Hierarchical Modelling of Haplotype Effects on a Phylogeny. Front. Genet. 11:531218. doi: 10.3389/fgene.2020.531218*

## 1. INTRODUCTION

This paper develops a hierarchical model to estimate haplotype effects based on phylogenetic relationships between haplotypes and their association with observed phenotypes. With current technology we can readily obtain genome-wide information about an individual, either through single-nucleotide polymorphism array genotyping or sequencing platforms. Since the genome-wide information has become abundant, modelling this data has become the standard in animal and plant breeding as well as human genetics. The application of this modelling has been shown to improve genetic gains in breeding (Meuwissen et al., 2001; Ibanez-Escriche and Simianer, 2016; Hickey et al., 2017), and has potential for personalised prediction in human genetics and medicine (de los Campos et al., 2018; Lello et al., 2018; Maier et al., 2018; Begum, 2019).

Geneticists aim to infer which mutations are causing variation in phenotypes and what are their effects. This aim is nowadays approached with genome-wide association studies of regressing observed phenotypes on mutation genotypes (Morris and Cardon, 2019). However, mutations arise on specific haplotypes passed between generations, which limits accurate estimation due to low frequency of mutations, correlation with other mutations and limited ability to observe all mutations with a used genomic platform (e.g., see Gibson, 2018; Simons et al., 2018; Uricchio, 2019). Further, most mutations do not affect phenotypes, while some mutations have background (haplotype) specific effects (e.g., Chandler et al., 2017; Steyn et al., 2019; Wojcik et al., 2019).

Instead of focusing on mutation effects we here focus on haplotype effects and their differences to estimate the effect of mutations on specific haplotypes. There is extensive literature on estimating haplotype effects (Balding, 2006; Thompson, 2013; Morris and Cardon, 2019). One issue with estimating haplotype effects is that there is usually an uneven distribution of haplotypes in a population (Ewens, 1972, 2004; Walsh and Lynch, 2018), and estimating the effects of rare haplotypes is equally challenging as estimating the effect of rare mutations. However, the described genetic processes in the previous paragraph create a "network" of haplotypes (sometimes referred to as *genealogy* or *phylogeny*), which suggests that effects of similar haplotypes are similar. This observation inspired (Templeton et al., 1987) to cluster phylogenetically similar haplotypes. Others have used similar approaches to account or leverage haplotype similarities (Balding, 2006; Thompson, 2013; Morris and Cardon, 2019).

We here approach the problem of estimating haplotype effects by leveraging phylogenetic relationships between haplotypes described with a directed acyclic graph (DAG) (Koller and Friedman, 2009) and developing a hierarchical model of haplotype effects on this graph. We were inspired by recent advances in building phylogenies on large data sets (Kelleher et al., 2019), and aimed to develop a hierarchical model that could scale to a large number of haplotypes. Our work extends the phylogenetic mixed modelling of the whole genome (Lynch, 1991; Pagel, 1999; Housworth et al., 2004; Hadfield and Nakagawa, 2010) to a specific region. This region specific modelling could be applied either across species (macroevolution) or within a species (microevolution).

A potentially important modelling aspect with respect to across and within species modelling is that the phylogenetic mixed model assumes Brownian motion for evolution of phenotypes along a phylogeny (Felsenstein, 1988; Huey et al., 2019). Brownian motion is a continuous random-walk process with variance that grows over time (is non-stationary) (Gardiner, 2009; Blomberg et al., 2019), which makes it a plausible model of evolution due to mutation and drift. There are alternatives to Brownian motion, in particular the Ornstein-Uhlenbeck process that can accommodate various forms of selection (Lande, 1976; Hansen and Martins, 1996; Martins and Hansen, 1997; Paradis, 2014). The Ornstein-Uhlenbeck process is also a continuous random-walk, but with an additional parameter that reverts the process to the mean (is a stationary process; e.g., Gardiner, 2009; Blomberg et al., 2019). Both of these models imply Gaussian distributions for the initial state and increments. The differences between the two processes might be important in the context of modelling haplotypes that likely manifest less variation than whole genomes, particularly when considering haplotypes within a species or even a specific population.

The aim of this paper is to develop a hierarchical model for haplotype effects by leveraging phylogenetic relationships between haplotypes. We assume that such relationships are described with a DAG encoded network and therefore call the model the haplotype network model. Since haplotypes differ due to a small number of mutations and very few mutations have an effect we expect that phylogenetically similar haplotypes will have similar effects. Furthermore, the small discrete number of mutation differences suggest discrete-time analogues of Brownian and Ornstein-Uhlenbeck processes. Therefore, we have modelled the effect of a mutated haplotype given its parental haplotype with a stationary autoregressive model of order one following the phylogenetic structure encoded with a DAG. The results show that the haplotype network model improves the estimation of haplotype effects compared to an independent haplotype model due to sharing of information. The results also show that it is comparable to a mutation model, but has the potential to capture background specific effects.

## 2. MATERIALS AND METHODS

We present the haplotype network model and show how to use it as a component in a phenotype model. We also describe simulations, a case study of modelling mitochondrial effects on milk yield in cattle, and the chosen method to perform inference and model evaluation.

### 2.1. The Haplotype Network Model

We present the haplotype network model, which is a hierarchical model for haplotype effects based on phylogenetic relationships between haplotypes encoded with a DAG. The phylogenetic relationships can be either in a form of a tree or a more general network. We also present two generalisations of the model—first due to multiple parental haplotypes and second due to genetic recombination. By multiple parental haplotypes we mean the situation where two different haplotypes in a phylogeny mutate into the same haplotype.

We assume throughout that the phylogeny between haplotypes is known and that it can be encoded with a DAG. The haplotype network model can in principle deal with different types of mutations, but for simplicity we focus only on biallelic mutations with the code 0 used for the ancestral/reference allele (commonly at a higher frequency in a population), and the code 1 used for the alternative allele that arose due to a mutation.

#### 2.1.1. Motivating Example

To motivate the haplotype network model, we use the example from Kelleher et al. (2019) that presents 5 haplotypes spanning 7 biallelic polymorphic sites (**Table 1**). Note that the 5 haplotypes are just a sample of the $2^7 = 128$ possible haplotypes over the 7 sites. An example of a phylogeny for the haplotypes is shown in **Figure 1**, where haplotypes are denoted as nodes (we also show their allele sequence), relationships between haplotypes

**TABLE 1 |** Example of 5 haplotypes spanning 7 mutations from Kelleher et al. (2019).

| | | Site | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Haplotype | a | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| | b | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| | c | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| | d | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | e | 0 | 1 | 1 | 0 | 0 | 0 | 1 |

*The ancestral (reference) alleles are coded as 0 and alternative alleles are coded as 1.*

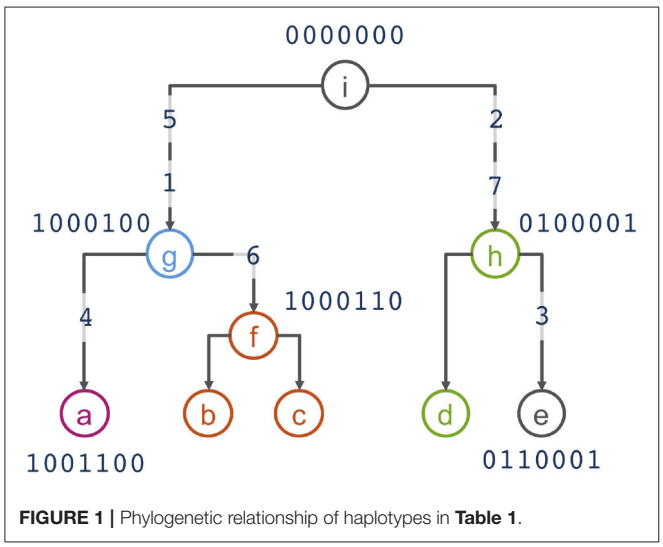


**FIGURE 1 |** Phylogenetic relationship of haplotypes in **Table 1**.

are denoted as edges, and mutated sites are denoted with a number on edges. For example, the ancestral haplotype $i$ has allele sequence `0000000`, and the haplotype $g$ with sequence `1000100` differs from the ancestral haplotype due to mutations at the sites 5 and 1.

Assuming that similar haplotypes have similar effects, we model dependency between parent-progeny pairs of haplotypes with an autoregressive Gaussian process of order one. For haplotypes in **Table 1** and **Figure 1** this model implies the following set of conditional dependencies:

$$\begin{aligned}
h_i &\sim N\left(0, \sigma^2_{h_m}\right)\\
h_{g'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right)\\
h_g|h_{g'} &\sim N(\rho h_{g'}, \sigma^2_{h_c})\\
h_a|h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right)\\
h_f, h_b, h_c, |h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right)\\
h_{h'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right)\\
h_h, h_d|h_{h'} &\sim N\left(\rho h_{h'}, \sigma^2_{h_c}\right)\\
h_e|h_h &\sim N\left(\rho h_h, \sigma^2_{h_c}\right)
\end{aligned}$$

where $h_i, h_g, \ldots, h_e$ indicate the effect of haplotypes $i, g, \ldots, e$, and $h_{*'}$ indicates the effect of haplotypes that occur between haplotypes separated by multiple mutations, for example, $g'$ is the additional haplotype between the haplotypes $i$ and $g$ due to two mutations between $i$ and $g$; we describe the other model parameters $\left(\rho, \sigma^2_{h_m}, \sigma^2_{h_c}\right)$ in the following.

### 2.1.2. The Model

Assume a known general phylogenetic network of haplotypes described with a DAG with haplotype effects as nodes and relationships between the haplotype effects as edges as in **Figure 1**, and let repeated identical haplotypes be handled as the same haplotype. We model the effect of a chosen "starting" (this could be either a central, ancestral, most common or some other choice) haplotype 1 with mean-zero and marginal variance $\sigma^2_{h_m}$:

$$h_1 \sim \mathcal{N}(0, \sigma^2_{h_m}), \tag{1}$$

and any other haplotype $j$ in the phylogenetic network as a function of its one-mutation-removed parental haplotype $p(j)$ assuming the autoregressive Gaussian process of order one with the autocorrelation between haplotype effects of $\rho$ ($|\rho| < 1$ to ensure stationarity) and conditional variance of $\sigma^2_{h_c}$ as:

$$h_j|h_{p(j)} \sim \mathcal{N}(\rho h_{p(j)}, \sigma^2_{h_c}). \tag{2}$$

We consider the autoregressive Gaussian process of order one that is stationary both in mean and variance, which is achieved by setting the marginal variance to $\sigma^2_{h_m} = \sigma^2_{h_c}/(1-\rho^2)$, so $\sigma^2_{h_c} = \sigma^2_{h_m}(1-\rho^2)$. The variance parameter is capturing scale (spread) of haplotype effects and the autocorrelation parameter is capturing dependency between haplotype effects. This is the standard autoregressive model of order one used in time-series analysis (e.g., Rue and Held, 2005). The difference here is that we are applying the model onto a phylogenetic network described with a tree or more generally with a DAG (Basseville et al., 1992; Wu et al., 2020).

The set of distributions in Equation (1) and Equation (2) give a system of equations for all $n$ haplotype effects $\boldsymbol{h} = (h_1, ..., h_n)^T$:

$$\boldsymbol{h} = \mathbf{T}(\rho)\,\varepsilon, \tag{3}$$

$$\mathbf{T}(\rho)^{-1}\boldsymbol{h} = \boldsymbol{\varepsilon}, \tag{4}$$

where the matrices $\mathbf{T}(\rho)$ and $\mathbf{T}(\rho)^{-1}$ of dimension $n \times n$ respectively represent marginal and conditional phylogenetic regression between haplotype effects $\boldsymbol{h}$ and the vector $\boldsymbol{\varepsilon}$ represents haplotype effect deviations, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}(\rho)\,\sigma^2_{h_c})$. The expression $\mathbf{T}(\rho)$ indicates that the matrix $\mathbf{T}$ depends on the value of $\rho$. Since haplotype effect deviations are independent, the matrix $\mathbf{D}(\rho)$ is diagonal and has value $1/(1-\rho^2)$ for the "starting" haplotype and 1 for the other haplotypes. Following the assumed autoregressive process of order one (2), the non-zero elements of $\mathbf{T}(\rho)^{-1}$ are 1 along the diagonal and $-\rho$ between a haplotype effect (row index) and its parental haplotype effect (column index). This simple sparse lower-triangular structure

of the matrix $\mathbf{T}(\rho)^{-1}$ arises from the Markov properties of the autoregressive process (Rue and Held, 2005).

From Equation (3), the covariance between haplotype effects is:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{h}) &= \mathrm{Var}(\mathbf{T}(\rho)\,\boldsymbol{\varepsilon}), &(5)\\ &= \mathbf{T}(\rho)\,\mathrm{Var}(\boldsymbol{\varepsilon})\mathbf{T}(\rho)^T = \mathbf{T}(\rho)\,\mathbf{D}(\rho)\,\mathbf{T}(\rho)^T\sigma^2_{h_c} &(6)\\ &= \mathbf{H}(\rho)\,\sigma^2_{h_c} = \mathbf{V}_h\left(\rho,\sigma^2_{h_c}\right), &(7)\end{aligned}$$

The covariance expression in Equation (5) shows that haplotype covariances $\mathbf{V}_h\left(\rho,\sigma^2_{h_c}\right)$ depend on the autocorrelation and variance parameters, while the covariance coefficients $\mathbf{H}(\rho)$ depend only on the autocorrelation parameter. Note that the parameters $\rho$ and $\sigma^2_{h_c}$ are correlated by definition $\sigma^2_{h_c} = \sigma^2_{h_m}(1-\rho^2)$. When $\rho = 0$ there is no covariance between haplotype effects due to phylogenetic relationships, which suggests a model where haplotype effects are identically and independently distributed, $\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_{h_m})$. When $\rho \neq 0$ effects of phylogenetically related haplotypes covary due to shared mutations.

For completeness, the joint density of all $n$ haplotype effects $\boldsymbol{h}$ is multivariate Gaussian:

$$\boldsymbol{h}|\rho,\sigma^2_{h_c} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_h\left(\rho,\sigma^2_{h_c}\right)), \tag{8}$$

with the probability density function:

$$p(\boldsymbol{h}|\rho,\sigma^2_{h_c}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}_{h_c}(1-\rho^2)^{1/2} \exp\left(-\frac{1}{2\sigma^2_{h_c}}\boldsymbol{h}^T\mathbf{H}(\rho)^{-1}\boldsymbol{h}\right). \tag{9}$$

The expression in Equation (9) involves inverse of the covariance coefficient (precision) matrix $\mathbf{H}(\rho)^{-1}$, which we can obtain without computationally expensive inverse of the $\mathbf{H}(\rho)$ (5). Following the definition in Equation (5), inverting both sides and using the described structure of $\mathbf{T}(\rho)^{-1}$ available from the DAG and $\mathbf{D}(\rho)$, we can efficiently get this inverse by:

$$\mathbf{H}(\rho)^{-1} = \frac{1}{\sigma^2_{h_c}}\mathbf{T}(\rho)^{-1^T}\mathbf{D}(\rho)^{-1}\mathbf{T}(\rho)^{-1}. \tag{10}$$

Inspection of the structure of Equation (10) shows that this is a very sparse matrix with a structure. We can compute the non-zero elements of $\sigma^2_{h_c}\mathbf{H}(\rho)^{-1}$ directly with the following simple algorithm where we loop over all haplotypes:

**if** the haplotype is the "starting" haplotype **then**
  add $1-\rho^2$ to the diagonal element
**else**
  add 1 to the diagonal element
**end if**
**if** the haplotype has a parental haplotype **then**
  set off-diagonal element between the haplotype and its parental haplotype to $-\rho$
  add $\rho^2$ to the diagonal element of the parental haplotype
**end if**

To fully specify the model for $\boldsymbol{h}$ in Equation (8), prior distributions must be assigned to the autocorrelation parameter $\rho$ and the marginal variance $\sigma^2_{h_m}$ or the conditional variance $\sigma^2_{h_c}$. Because most mutations do not have an effect we can expect that most parent-progeny pairs of haplotypes will have similar effects, which suggests that the autocorrelation parameter will be close to 1. This knowledge can be incorporated in the prior distribution for $\rho$. For the variance parameters there may be some prior knowledge about the size of haplotype effects relative to other effects, which can also be taken into account when choosing the prior distribution. We will specify prior distributions for these parameters in later sections.

### 2.1.3. Multiple Parental Haplotypes

Sometimes phylogenetic inference cannot resolve bifurcating trees with dichotomies (one parental haplotype and two progeny haplotypes) and outputs a multifurcating tree with polytomies (one parental haplotype and multiple progeny haplotypes) or even just a network [multiple parent haplotypes and multiple progeny haplotypes (e.g., Schliep et al., 2017; Uyeda et al., 2018)]. The multiple progeny case works out of the box with the initial model, and we will here present an extension of the model presented in section 2.1.2, that can accommodate the multiple parent haplotypes and multiple progeny haplotypes case where the trees or networks can be described with a DAG.

We assume that the effects of all ancestral haplotypes, the haplotypes at the top of the network, are independent and come from the same Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_{h_m})$. We further assume conditional independence between a haplotype and all previous haplotypes in the network given the parents of that haplotype. In the model where each haplotype had only a single parent haplotype it was assumed that the haplotype effect was $\rho$ times the parental haplotype effect plus some Gaussian noise. When a haplotype has multiple parents, we now assume that the effect is the average over each of these processes from each parental haplotype.

We illustrate this with a small example which implies the model construction used. Let haplotype segment $d$ have parental haplotypes segments $a$, $b$, and $c$. We denote the contribution from each of these parental segments $h_{d_a}, h_{d_b}, h_{d_c}$, and assume:

$$\begin{aligned}h_{d_a} &= \rho h_a + \varepsilon_{d_a}\\ h_{d_b} &= \rho h_b + \varepsilon_{d_b}\\ h_{d_c} &= \rho h_c + \varepsilon_{d_c}\end{aligned}$$

where $(\varepsilon_{d_a}, \varepsilon_{d_b}, \varepsilon_{d_c})^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_{h_c})$. Further, we assume that the resulting effect of haplotype $h_d$ is the average over all parent processes:

$$h_d = \frac{\rho}{3}(h_a + h_b + h_c) + \frac{1}{3}(\varepsilon_{d_a} + \varepsilon_{d_b} + \varepsilon_{d_c}).$$

The distribution of $h_d$ conditional on $h_a$, $h_b$, and $h_c$ becomes:

$$h_d|h_{d_a}, h_{d_b}, h_{d_c} \sim \mathcal{N}\left(\frac{\rho}{3}(h_a + h_b + h_c), \frac{\sigma^2_{h_c}}{3}\right).$$

In general this means that $h_i|h_1, ..., h_k \sim \mathcal{N}(\frac{\rho}{k}\sum_{j=1}^{k} h_j, \frac{\sigma^2_{h_c}}{k})$, for haplotype $i$ with parental haplotypes $1, ..., k$. This model construction corresponds to a model where one first takes every path down through the DAG and assigns separate stationary autoregressive processes of order one to each such path, and then assume conditionally independent but identical autoregressive processes of order one, that is, the processes have the same parameters.

Multiple parental haplotypes change the structure of the $\mathbf{T}(\rho)^{-1}$ matrix to having $-\rho/k_i$ value between a haplotype effect (row index) and its parental haplotype effect (column index) and $\mathbf{D}(\rho)^{-1}$ matrix diagonals for "non-starting" haplotypes to $k_i$, where $k_i$ is the number of parental haplotypes of the haplotype $i$. The algorithm to setup the $\sigma^2_{h_c}\mathbf{H}(\rho)^{-1}$ matrix is then (looping over all haplotypes)

**if** the haplotype is the "starting" haplotype **then**
  add to the diagonal element $1 - \rho^2$
**else**
  add $k_i$ to the diagonal element
**end if**
**if** the haplotype has a parental haplotype **then**
  set off-diagonal element between the haplotype and its parental haplotype to $-\rho$
  set off-diagonal elements between all parental haplotypes that share that progeny haplotype to $\rho^2/k_i$
  add $\rho^2/k_i$ to the diagonal element of the parental haplotype(s)
**end if**

The model presented in this section is only one of many possible choices for a model accommodating multiple parental haplotypes. There are other options that could model such graph structures, for example by modelling it as a mixture distribution with variable probabilities between parental haplotypes.

### 2.1.4. Expanding to Multiple Regions Due to Recombination

Haplotype phylogeny can differ along genome regions due to recombination—the process of swapping genome regions between haplotypes during meiosis. We accommodate this in the haplotype network model by considering each haplotype region separately, but still within the framework of the same model. This means that the effect of haplotype $h_i$ is modelled as the sum of effects for all haplotype regions. Consider haplotypes spanning three regions. The effect of haplotype $i$, is then assumed to be the sum of the effects of haplotype segments in each of the three regions:

$$h_i = h_{1,i} + h_{2,i} + h_{3,i}.$$

We assume the haplotype network model for each haplotype region, but with joint hyper-parameters $\left(\rho, \sigma^2_{h_c}\right)$. Let $\boldsymbol{h} = (h_{1,1}, \ldots, h_{1,n_1}, h_{2,1}, \ldots, h_{m,n_m})$ be the effect of all haplotypes in all regions, where $m$ is the number of regions and $n$ is the number of haplotypes in each region. The joint probability density for the haplotype effects $\boldsymbol{h}$ is then:

$$p(\boldsymbol{h}|\rho, \sigma^2_{h_c}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+\ldots+n_m} \sigma^{-(n_1+\ldots+n_m)}_{h_c}(1-\rho^2)^{m/2} \exp\left(-\frac{1}{2\sigma^2_{h_c}}\boldsymbol{h}^T\mathbf{H}(\rho)^{-1}\boldsymbol{h}\right),$$

with:

$$\mathbf{H}(\rho)^{-1} = \begin{pmatrix} \mathbf{H}(\rho)_1^{-1} & & \\ & \ddots & \\ & & \mathbf{H}(\rho)_m^{-1} \end{pmatrix}. \quad (11)$$

Although recombination is common, we have focused on the special case of no recombination in this study, where the haplotypes are connected in one phylogeny, as presented in section 2.1.2. We address recombination in discussion.

## 2.2. Phenotype Model With Haplotype Effects

We now show how the haplotype effects can be included in a model for phenotypic observations. We also present a phenotype model that includes independent haplotype effects or mutation effects rather than the haplotypes.

Let $\boldsymbol{y}_{p\times 1}$ be phenotype observations of $p$ individuals and let $\boldsymbol{h}_{n\times 1}$ be the effect of $n$ haplotypes obtained from phasing genotypic data of the individuals. We assume the following model (Gaussian likelihood) for the centred and scaled phenotypic observations:

$$\boldsymbol{y}_{p\times 1} = \mathbf{X}_{p\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{f}^1_{p\times 1} + \ldots + \boldsymbol{f}^s_{p\times 1} + \mathbf{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \quad (12)$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}1000)$ is a vector of $r$ fixed effects with covariate matrix $\mathbf{X}$, $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_f)$ are random effects, $\boldsymbol{h}$ are the haplotype effects with incidence matrix $\mathbf{Z}$ that maps haplotypes to individuals, and the residual effect is $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_e)$. In the case of diploid individuals there will be two entries in every row of $\mathbf{Z}$, and a single entry for haploid individuals or male sex chromosome or mitogenome.

We have assumed three different models for the haplotype effects $\boldsymbol{h}$. The first is a base model with independent haplotype effects (IH model), where $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2_{h_m}\right)$. The second is the haplotype network model presented in section 2.1.2 (HN model), where $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_h\left(\rho, \sigma^2_{h_c}\right)\right)$. The third is an alternative way of estimating haplotype effects via a linear combination of mutation effects (mutation model). Assume $\boldsymbol{h} = \mathbf{U}\boldsymbol{v}$ with $\boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2_v\right)$ being mutation effects and $\mathbf{U}$ is the matrix containing the haplotype allele sequence with reference alleles coded as 0 and alternative alleles coded as 1. The effects described so far consist of the latent field of a Bayesian hierarchical model, and are assigned Gaussian prior distributions.

The models do not have a common intercept because a common intercept and the mean level in the haplotype effects are not identifiable when $\rho$ approaches 1. Instead the mean level in the observations is captured by the haplotype effects, for computational reasons. A sum-to-zero constraint can be specified for the haplotype network part of the model if a common intercept is required, though this changes the model interpretation if $\rho$ is close to 1. This problem is not specific to this model, but occurs for all autoregressive models when they are used as part of a structured mixed effects model. When the goal is to make predictions about the haplotype effects, this model choice will not influence the results.

#### 2.2.1. Prior Distributions

We assigned penalised complexity (PC) prior distributions to the variance parameters and the autocorrelation parameter. PC priors are proper prior distributions developed by Simpson et al. (2017) that penalise increased complexity as measured by deviation from a simpler base model to avoid over-fitting. For a random effect with a variance parameter the base model has variance of this random effect zero. For the autoregressive model of order one we have assumed a base model with $\rho = 1$. We could have assumed a base model with $\rho = 0$, but it is more likely that phylogenetically similar haplotypes have similar effects than completely independent effects. The PC prior can be specified through a parameter $u$ and a probability $\alpha$ which satisfy $\text{Prob}(x > u_x) = \alpha_x$ for the parameter $x$. We emphasise that the parameter $u$ here is not an element of the allele sequence matrix $\mathbf{U}$ mentioned above.

Although the precision matrix of the haplotype effects is specified with the conditional variance in Equation (10), the prior is specified for the marginal variance since we often have a better intuition for the marginal variance than for the conditional variance. Specifically, we specify the prior for the marginal standard deviation $\sigma_{h_m}$, and assume the conditions $u_{\sigma_{h_m}} > 0$ and $0 < \alpha_{\sigma_{h_m}} < 1$. For the autocorrelation parameter we use the PC prior developed for stationary autoregressive processes (Sørbye and Rue, 2017) with base model at $\rho = 1$, and parameters satisfying $-1 < u_\rho < 1$ and $\sqrt{(1-u_\rho)/2} < \alpha_\rho < 1$. We highlight that the prior by Sørbye and Rue (2017) was developed for a stationary autoregressive process with different model assumptions than the models presented in this paper. Ideally, the prior for the autoregressive parameter would be tailored to the haplotype network model.

## 2.3. Inference and Evaluation

We describe the used method for statistical inference—the Integrated nested Laplace approximations (INLA)—and the methods used for evaluating model fit in the simulation study.

#### 2.3.1. Inference

All models in this study fit in the framework of hierarchical latent Gaussian models, which makes INLA (Rue et al., 2009) a suitable choice to perform inference as implemented in the R (R Core Team, 2018) package `INLA` (available at `www.r-inla.org`). We give a brief introduction to latent Gaussian models and how INLA is used to approximate the marginal posterior distributions in such models. For an in-depth description of INLA (see Rue et al., 2009, 2017; Blangiardo and Cameletti, 2015).

The class of latent Gaussian models includes several models, for example generalised linear (mixed) models, generalised additive (mixed) models, spline smoothing methods, and the models presented in this article. Latent Gaussian models are hierarchical models where observations $\boldsymbol{y}$ are assumed to be conditionally independent given a latent Gaussian random field $\boldsymbol{x}$ and hyper-parameters $\boldsymbol{\theta}_1$, meaning $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) \sim \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta}_1)$. The latent field $\boldsymbol{x}$ includes both fixed and random effects and is assumed to be Gaussian distributed given hyper-parameters $\boldsymbol{\theta}_2$, that is $p(\boldsymbol{x}|\boldsymbol{\theta}_2) \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{\Sigma}(\boldsymbol{\theta}_2))$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are known as hyper-parameters and control the Gaussian field and the likelihood for the data. These are usually variance parameters for simple models, but can also include other parameters, for example the $\rho$ parameter in the autoregressive model. We must also assign prior distributions to the hyper-parameters to completely specify the model.

The main aim of Bayesian inference is to estimate the marginal posterior distribution of the variables of interest, that is, $p(\theta_j|\boldsymbol{y})$ for hyper-parameters and $p(x_i|\boldsymbol{y})$ for the latent field. INLA computes fast approximations to these densities with high accuracy. The INLA methodology is based on numerical integration and utilising Markov properties. Hence, for the computations to be both fast and accurate, the latent Gaussian models have to satisfy some assumptions. The number of non-Gaussian hyper-parameters $\boldsymbol{\theta}$ should be low, typically less than 10, and not exceeding 20. Further, the latent field should not only be Gaussian, it must be a Gaussian Markov random field. The conditional independence property of a Gaussian Markov random field yields sparse precision matrices which makes computations in INLA fast due to the use of efficient algorithms for sparse matrices. Lastly, each observation $y_i$ should depend on the latent Gaussian field only through one component $x_i$.

The R package INLA is run using the `inla()` function with three mandatory arguments: a data frame or stack object containing the data, a formula much like the formula for the standard `lm()` function in R, and a string indicating the likelihood family. Prior distributions for the hyper-parameters are specified through additional arguments. Several tools to manipulate models and likelihoods exist as described in tutorials at `www.r-inla.org` and the books by Blangiardo and Cameletti (2015), and Krainski et al. (2018). In the **Supplementary Material** (Supplemental 1), we have included a script showing how we simulated the data from the haplotype network model and how we fitted the model to the data.

#### 2.3.2. Evaluation of Model Performance

We evaluated the model fit with the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007). The CRPS is a proper score which takes into account the whole posterior distribution. It is negatively oriented, so the smaller the CRPS the closer the posterior distribution is to the true value. The full Bayesian posterior output from `inla()` for these models are mixtures of Gaussians, for which there is no closed form expression for CRPS. The mixtures here are similar to plain Gaussians, so we

approximate the exact CRPS with the Gaussian CRPS using only the posterior mean and variances provided in the results.

We calculated the CRPS for estimated haplotype effects with the IH, HN and mutation models. To ease the comparison we have then calculated a relative CRPS (RCRPS) score as the log of the ratio between the averages of the CRPS from the HN model and IH model, and correspondingly for the mutation model relative to the IH model. The score is computed as:

$$\log\left(\frac{\sum_{i=1}^{n}\text{CRPS}(\hat{h}_i)_{HN}}{\sum_{i=1}^{n}\text{CRPS}(\hat{h}_i)_{IH}}\right),$$

where $\text{CRPS}(\hat{h}_i)_{HN}$ is the CRPS of the posterior distribution for haplotype effect $h_i$ with the HN model. We will refer to this score as the RCRPS.

We also calculated the root mean square error (RMSE) between the mean posterior haplotype effect and true haplotype effects, but the results for the relative RMSE and RCRPS were qualitatively the same. We therefore only present the RCRPS results.

In addition to comparing the haplotype estimates, we compared the estimated mutation effects from the HN model and the mutation model, using the RCRPS (HN model vs. mutation model). Although the HN model estimates the haplotype effects $\boldsymbol{h}$, we can obtain mutation effects via $\boldsymbol{v} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{h}$. We could also obtain mutation effects through linear combinations of haplotype effects.

## 2.4. Simulation Study

To test the proposed HN model, we first used simulated data. Here, we present data simulated from two different models—the HN model with varying degree of autocorrelation, and a more realistic mutation model where only some mutations have causal effect. We also present the models that were fitted to the simulated data, and how the model fit was evaluated. In the **Supplementary Material** (Supplemental 1), we provide an R script and the data file to simulate from and fit the haplotype network model.

### 2.4.1. Simulation From the Haplotype Network Model

We used the coalescent simulator `msprime` (Kelleher et al., 2016) to simulate the phylogeny shown in **Figure 2** with $n = 107$ unique haplotypes. A script showing how this was performed is provided in the **Supplementary Material** (Supplemental 1) We then simulated phenotypes $\boldsymbol{y}$ for $p = 400$ individuals from the model:

$$\boldsymbol{y}_{p\times 1} = \mathbf{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{13}$$

where $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_h\left(\rho, \sigma^2_{h_c}\right)\right)$ with $\mathbf{V}_h\left(\rho, \sigma^2_{h_c}\right)$ built from the DAG describing the phylogeny (**Figure 2** Equation 5), and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_e)$.

We tested 15 parameter sets, from weak to strong haplotype effect dependency, and from low to high residual variance relative to the conditional haplotype variance:

$$\rho = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\sigma^2_e/\sigma^2_{h_c} = \{0.5, 1, 2\}.$$

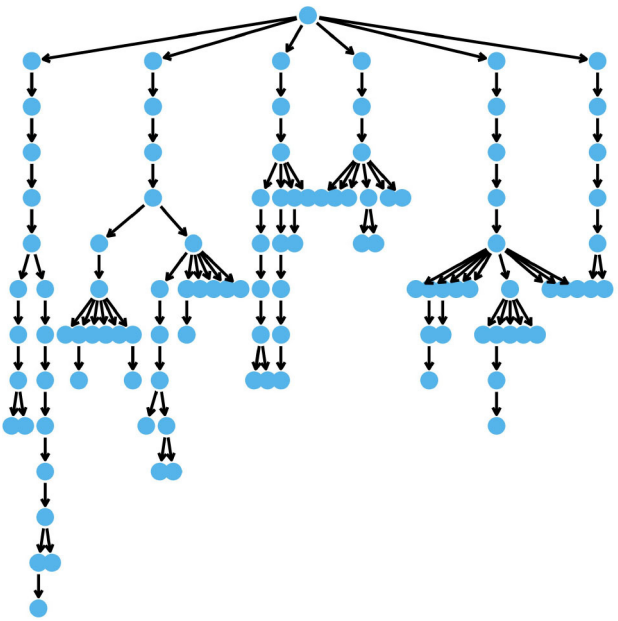

**FIGURE 2 |** The DAG describing the phylogeny of simulated haplotypes.

We simulated a haploid system for simplicity, so the incidence matrix $\mathbf{Z}$ was a zero matrix with a single 1 on each row indicating which individuals had which haplotype. We were particularly interested in estimating the haplotype effect with few or no direct phenotype observations. This is the extreme scenario where the haplotype network model could be beneficial. To achieve this, we designed the incidence matrix to create two different scenarios. In the first scenario, all haplotypes had associated phenotype observation, but some haplotypes only had one observation. We assigned a random sample of 15% of the haplotypes only to one individual each and the rest of the haplotypes randomly to the remaining individuals. In the second scenario, some haplotypes did not have phenotype observations. We selected a random sample of 15% of the haplotypes that did not have phenotype observations and assigned phenotype observations to the rest of the haplotypes. The values of the simulated observations ranged between −7.2 and 7.3.

### 2.4.2. Simulation From the Mutation Model

We also simulated haplotype effects from a mutation model using the same phylogeny as in the previous section, shown in **Figure 2**, and using $p = 400$ individuals. For the 107 unique haplotypes we had 106 mutations in the haplotypes. We used the variants at these mutations to simulate haplotype effects and phenotypes according to the model:

$$\boldsymbol{y}_{p\times 1} = \mathbf{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{14}$$

where $\boldsymbol{h} = \mathbf{U}_{n\times 106}\boldsymbol{v}_{106\times 1}$, $\boldsymbol{v}$ was the mutation effect, $\mathbf{U}$ a matrix containing ancestral (reference) alleles coded as zero and alternative alleles coded as 1, and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$. We sampled the mutation effect $v$ from:

$$v = \begin{cases} \mathcal{N}(0, \sigma_v^2), \text{ with probability } \lambda \\ 0, \text{ with probability } (1-\lambda) \end{cases}$$

where we chose $\sigma_v^2$ so that the empirical variance of $\boldsymbol{h}$, Var($\boldsymbol{h}$), was 1.

Again, we tested 15 parameter sets, from few to many causal variants, and from low to high residual variance relative to empirical haplotype variance:

$$\lambda = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\sigma_e^2/\text{Var}(\boldsymbol{h}) = \{0.5, 1, 2\}.$$

We again simulated haploid individuals, so the incidence matrix $\mathbf{Z}$ was a zero matrix with a single 1 on each row indicating which individuals had which haplotype. The incidence matrix was designed to create the same scenarios as for the data simulated from the HN model in section 2.4.1. The values of the simulated observations ranged between −8.4 and 8.9.

### 2.4.3. Models Fitted to the Simulated Data

We fitted the HN model, IH model and the mutation model to the simulated data:

$$\boldsymbol{y}_{p\times 1} = \mathbf{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{15}$$

where $\boldsymbol{h}$ was assumed to be distributed according to:

$\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_h\left(\rho, \sigma_{h_c}^2\right)\right)$ for the HN model,
$\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_I^2\right)$ for the IH model and
$\boldsymbol{h} = \mathbf{U}\boldsymbol{v}$, $\boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_v^2\right)$ for the mutation model.

The residual effect was $\boldsymbol{e} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$. We used PC priors for the $\rho$ parameters with $u_\rho = 0.7$ and $\alpha_\rho = 0.8$, and for all variance parameters with $u = 0.1$ and $\alpha = 0.8$.

### 2.4.4. Evaluation

For each parameter set, we performed the same experiment 50 times. In 4% of the experiments when the data was simulated from the HN model, the optimisation method with the HN model did not converge. We report results only for cases where all models were successfully fitted. There was no trend for any parameter set causing the inference method to break down.

Since we created different scenarios for how phenotype observations were distributed among the haplotypes, we stratified the results for haplotype effects based on how many times a haplotype was phenotyped. For the first scenario, where some haplotypes were phenotyped either once or multiple times, we have computed the RCRPS for these two groups separately. For the second scenario, where some haplotypes were not phenotyped, we present the RCRPS only for haplotypes that were not phenotyped. In both cases, RCRPS less than zero indicates that the HN/mutation model was better than the IH model on average. We present the RCRPS for estimated mutation effects only for the mutation model simulation, because the true mutation effects were not generated when simulating from the haplotype network model.

## 2.5. Case Study: Mitochondrial Haplotypes in Cattle

We present a case study using the haplotype network model to estimate the effect of mitochondrial haplotypes on milk yield in cattle. We first briefly describe the data and then the fitted model.

### 2.5.1. Data

We demonstrate the use of the haplotype network model with a case study estimating the effect of mitochondrial haplotypes on milk yield in cattle from Brajković (2019). We chose this case study because mitochondrial haplotypes are passed between generations without recombination and are as such a good case for the haplotype network model. The phenotyped data comprised of information about the first lactation milk yield, age at calving, county, herd-year-season of calving for 381 cows. Additionally, the data comprised of pedigree information with 6,336 individuals (including the 381 cows) and information about mitochondrial haplotypes (whole mitogenome with 16,345 bp) variation between maternal lines in the pedigree. We inferred the mitochondrial haplotypes by first sequencing mitogenome, aligning it to the reference sequence and calling 363 single-nucleotide mutations as described in detail in Brajković (2019). We used PopART (Leigh and Bryant, 2015) to build a phylogentic network of mitochondrial haplotypes. For simplicity we used the median-joining method to show that the haplotype network model can be fit to the output of a standard phylogentic method. In this process we assumed that the ancestral alleles were the most frequent alleles. The phylogeny contained 63 unique mitochondrial haplotypes each separated by one mutation. Of the 63 haplotypes only 16 haplotypes were observed in the 381 phenotyped cows. There were five haplotypes that did not have a parent haplotype, meaning we treated them as a "starting" haplotype in the haplotype network model.

### 2.5.2. Model

Let $\boldsymbol{h}_{n\times 1}$ be the effect of the $n = 63$ mitochondrial haplotypes, and let $\boldsymbol{y}_{p\times 1}$ be the phenotypes of the $p = 381$ cows. We fitted the following model to centred and scaled phenotypes:

$$\boldsymbol{y}_{p\times 1} = \mathbf{X}_{p\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{c}_{p\times 1} + \boldsymbol{a}_{p\times 1} + \mathbf{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}1000)$ contained effects of age at calving as a continuous covariate effect and county as a categorical covariate effect with corresponding design matrix $\mathbf{X}$, $\boldsymbol{c} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_c^2\right)$ was the random effect of herd-year-season of calving (contemporary group), $\boldsymbol{a} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}\sigma_a^2\right)$ was additive genetic effect for the whole nuclear genome with the covariance coefficient matrix $\boldsymbol{A}$ derived from the pedigree (Henderson, 1976; Quaas, 1988), and lastly the mitochondrial haplotype effects were fitted with the haplotype network model $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_h\left(\rho, \sigma_{h_c}^2\right)\right)$ with the covariance matrix $\mathbf{V}_h\left(\rho, \sigma_{h_c}^2\right)$ derived from the phylogeny and

using the expanded model that accommodates multiple parental haplotypes from section 2.1.3. We assumed that residuals were distributed as $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$. We assigned PC priors to the $\rho$ parameter with $u_\rho = 0.7$ and $\alpha_\rho = 0.8$ and to the $\sigma^2_{h_m}$ parameter with $u_{\sigma_{h_m}} = 0.1$ and $\alpha_{\sigma_{h_m}} = 0.3$, and to all remaining variance parameters with $u_{\sigma_*} = 0.1$ and $\alpha_{\sigma_*} = 0.8$.

# 3. RESULTS

We present results from the simulation study testing the behavior of the haplotype network model and the case study estimating the effect of mitochondrial haplotypes on milk yield in cattle. In the results from the simulation study, we present the RCRPS between the haploptype network (HN) model and the independent haplotype (IH) model, and between the mutation model and the IH model for the different parameter sets. In the results from the case study, we present the mean and standard deviation of the posterior mitochondrial haplotype effects mapped onto the phylogenetic network, and posterior estimates for the hyper-parameters.

## 3.1. Simulation Study

### 3.1.1. Simulation From the Haplotype Network Model

We start by considering the results with the data simulated from the HN model from section 2.4.1 that were fitted with the models from section 2.4.3.

The RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) is presented in **Figure 3**. This figure has three panels denoting haplotypes that were observed in (**Figure 3A**) several phenotyped individuals, (**Figure 3B**) only one phenotyped individual and (**Figure 3C**) were not observed in a phenotyped individual. The full lines show the RCRPS between the HN model and the IH model, while the dashed lines show the RCRPS between the mutation model and the IH model. Along the $x$-axis the autocorrelation parameter $\rho$ for the simulated haplotype effects increases from weak to strong phylogenetic dependency, and the three colored lines indicate the amount of phenotypic variation due to residual relative to the variation from haplotype effects.

**Figure 3** shows that (1) the HN model outperforms the IH model across a range of parameter values, (2) the HN model is more important for haplotypes with fewer phenotypic observations, (3) the HN model is more important for noisy phenotypic data, and (4) when haplotypes are more phylogenetically dependent, the HN model and the mutation model have similar performance. We go through each of these findings in detail.

The HN model outperforms the IH model for almost all 15 parameter sets. In all panels of **Figure 3** almost all points with the full line are below zero, meaning that the HN model gave better estimates of haplotype effects than the IH model. When the haplotype dependency due to phylogeny was low, the RCRPS was around zero, meaning that the two models performed similarly, which was expected. As the phylogenetic dependency became stronger, the HN model improved relative to the IH model, as seen from the decreasing RCRPS as $\rho$ approaches 0.9.

The improvement in CRPS with the HN model relative to the IH model increased when haplotypes were observed in a smaller number of phenotyped individuals. This is indicated by the decreasing RCRPS when we compare panels (A), (B), and (C) in **Figure 3**. The decrease in RCRPS was the largest in **Figure 3C** followed by **Figure 3B** and **Figure 3A**. This means that modelling phylogenetic dependency between haplotypes is most useful when there are some haplotypes with few phenotypic observations, or if we want to predict the effect of new haplotypes. Especially for haplotypes that do not have a direct link to observed phenotypes, the IH model is not useful, because it assigns the average effect of haplotypes with direct link to observed phenotypes to haplotypes without such links, whereas the HN model can assign the haplotype effect based on a phylogenetic network. When the haplotype effects had low phylogenetic dependency ($\rho$ is low), there was not much difference in RCRPS between the three panels.

The improvement with the HN model relative to the IH model increased when the phenotypic data was noisier. In **Figures 3A,B**, the RCRPS was lower with larger residual variance. This indicates that the HN model does a better separation of the environmental and genetic sources of variation than the IH model. We did not observe the same in **Figure 3C**, because the IH model performed equally poorly in predicting new haplotypes regardless of the amount of residual variance. The HN model on the other hand, performed slightly better as there was less variation due to residual effects for some values of $\rho$ and similar for other values of $\rho$ compared to the IH model.

As haplotypes became phylogentically more dependent with the increasing $\rho$, the HN model and the mutation model performed similarly. In all panels the dashed lines indicate a worse fit for the mutation model than for the IH model and HN model when $\rho$ was low. When $\rho$ increased, the mutation model improved relative to the IH model, but not better than the HN model.

### 3.1.2. Simulated Data From the Mutation Model

Now, we consider the results with the haplotype effects simulated from a more realistic mutation model in section 2.4.2, and fitted with the models from section 2.4.3. Here we varied the probability of mutations having a causal effect $\lambda$ and we present results using only $\lambda = 0.1$, since the results were qualitatively similar for all tested $\lambda$ values.

The RCRPS is presented in **Figure 4** for the three different levels of phenotype observations per haplotype and three different values of residual variance relative to the empirical haplotype variance which was always 1. The full lines show the RCRPS between the HN model and the IH model, while the dashed lines show the RCRPS between the mutation model and the IH model.

In general, the results align with the results from the previous section except for the mutation model; (1) the HN model outperforms the IH model, (2) the HN model is more important for haplotypes with few phenotypic observations, (3) the HN model is more important for noisy phenotypic data and (4) the mutation model was marginally better than the HN model in

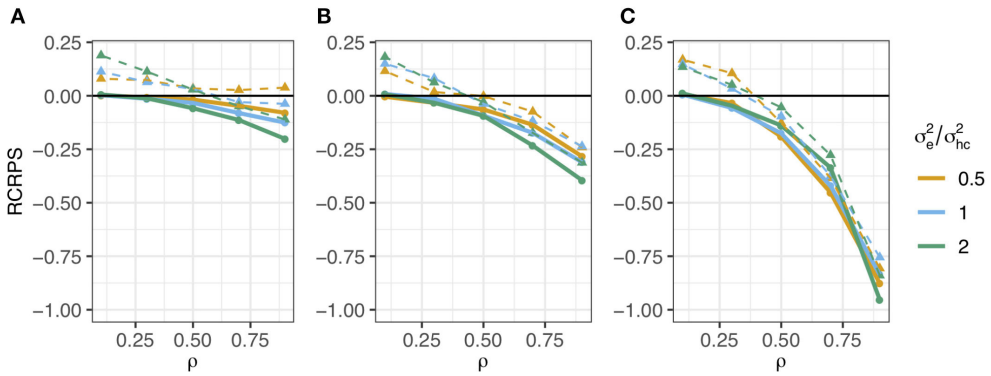


**FIGURE 3 |** RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) between the HN model and the IH model (solid line) and between the mutation model and the IH model (dashed line) for data simulated from the HN model with varying $\rho$ parameter and ratio between the residual $\sigma^2_e$ and conditional haplotype variance $\sigma^2_{h_c}$. The three panels show RCRPS for the haplotypes that were observed in **(A)** several phenotyped individuals, **(B)** only one phenotyped individual, and **(C)** were not observed in a phenotyped individual.

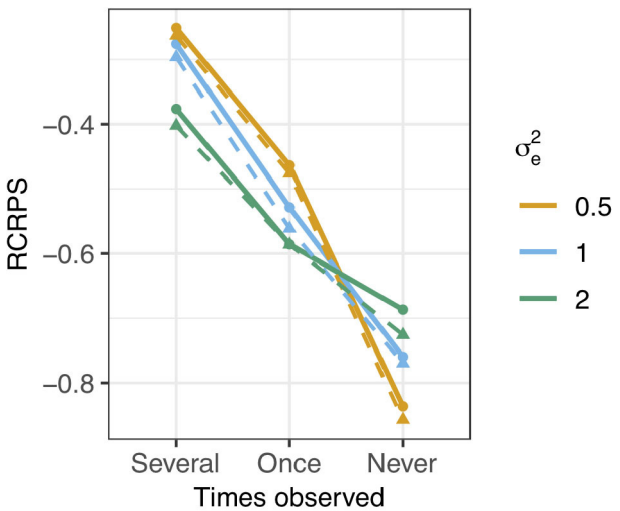


**FIGURE 4 |** RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) between the HN model and the IH model (solid line) and between the mutation model and the IH model (dashed line) for data simulated from the mutation model with varying residual variance $\sigma^2_e$ and empirical haplotype variance 1 [Var($\boldsymbol{h}$) = 1]. The three scenarios show RCRPS for the haplotypes that were observed in (Several) several phenotyped individuals, (Once) only one phenotyped individual, and (Never) were not observed in a phenotyped individual.

estimating haplotype effects. We go through each of the findings in detail.

The HN model outperformed the IH model for all tested parameter sets. In **Figure 4**, all RCRPS values, are well below zero. For haplotypes observed in several or one phenotyped individual, the RCRPS was lower than what was seen in **Figures 3A,B**. For haplotypes with no direct links to phenotype observations, the RCRPS was not improving as much as seen in **Figure 3C**.

The improvement with the HN model relative to the IH model increased with fewer phenotype observations per haplotype. The RCRPS in **Figure 4** is lowest for haplotypes with no direct links to phenotype observations, second lowest for haplotypes with one direct link to a phenotype observation, and highest for haplotypes that were observed in several phenotyped individuals.

The improvement with the HN model relative to the IH model increased with increasing residual variation. In **Figure 4**, the RCRPS for haplotypes observed in several or one phenotyped individual decreases with increasing residual variance. This was again not the case for haplotypes with no direct links to phenotype observations. As mentioned in the previous section, the IH model is predicting new haplotypes equally poorly irrespective of the residual variance. The HN model on the other hand, improves the prediction of new haplotypes when the phenotypic data is less noisy.

The mutation model was marginally better than the HN model in estimating haplotype effects. The dashed lines in **Figure 4** indicate the RCRPS between the mutation model and the IH model, and the full lines indicate the RCRPS between the HN model and the IH model. The dashed lines and full lines follow each other closely, and the dashed lines are slightly lower than the full lines, indicating that the mutation model was slightly better than the HN model, although not by much.

In **Table 2**, we present the average RCRPS between the HN model and the mutation model for the estimated mutation effects. This table has the RCRPS for the two scenarios where either all haplotypes had associated phenotype observation, or most haplotypes had associated phenotype observation and the rest did not, with different proportions of mutations with causal effect and for different residual variance. RCRPS above zero indicate that the mutation model had better CRPS, and averages below zero indicate that the HN model had better CRPS. Overall the difference between the two models is small. The mutation model had the best performance when there were few causal mutations, and the HN model had the best performance when there were many causal mutations.

**TABLE 2 |** RCRPS between the HN model and the mutation model for mutation effects by different values of residual variance $\sigma_e^2$, proportion of causal mutations and for the two scenarios where either all or most haplotypes have direct links to observed phenotypes.

| Prop. of causal mut. | All observed | Most observed |
|---|---|---|
| $\sigma_e^2 = 0.5$ | | |
| 0.1 | 0.060 | 0.071 |
| 0.3 | 0.019 | 0.025 |
| 0.5 | −0.002 | −0.004 |
| 0.7 | −0.019 | −0.021 |
| 0.9 | −0.027 | −0.029 |
| $\sigma_e^2 = 1$ | | |
| 0.1 | 0.123 | 0.111 |
| 0.3 | 0.043 | 0.037 |
| 0.5 | 0.004 | 0.000 |
| 0.7 | −0.024 | −0.022 |
| 0.9 | −0.041 | −0.034 |
| $\sigma_e^2 = 2$ | | |
| 0.1 | 0.168 | 0.214 |
| 0.3 | 0.067 | 0.101 |
| 0.5 | 0.006 | 0.018 |
| 0.7 | −0.025 | −0.026 |
| 0.9 | −0.042 | −0.048 |

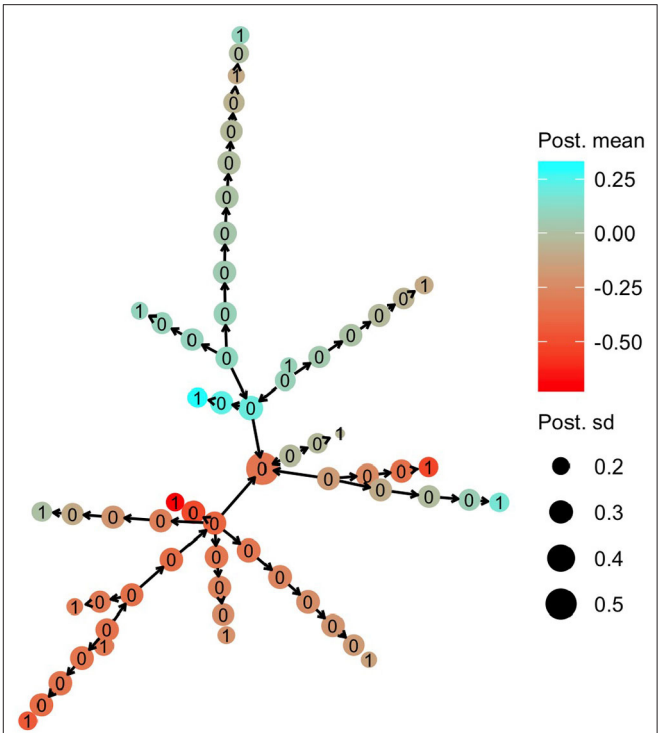


**FIGURE 5 |** Posterior mean and standard deviation for mitochondrial haplotype effects on milk yield in cattle. Posterior means are denoted with node color, while posterior deviations are denoted by the node size. The numbers on each haplotype node indicate if the haplotype had a direct link to the observed phenotype (1) or not (0).

## 3.2. Case Study: Mitochondrial Haplotypes in Cattle

We present results for the case study of estimating the effect of mitochondrial haplotypes on milk yield in cattle presented in section 2.5. We present the posterior mean and standard deviation for the effect of mitochondrial haplotypes mapped onto the phylogeny, the posterior distribution for the autocorrelation parameter $\rho$, and the mean and 95% confidence interval of the posterior variances in the model.

In summary, the results show (1) that there was sharing of information between the mitochondrial haplotypes, (2) that haplotypes without a direct link to observed phenotyopes were estimated with larger uncertainty, (3) indications of strong phylogenetic dependency between the haplotypes, and (4) a significant proportion of the total phenotypic variation explained by mitochondrial haplotypes.

The HN model enabled sharing of information from the haplotypes that had a direct link with observed phenotypes to the other haplotypes. In **Figure 5**, we present the posterior mean for the effect of mitochondrial haplotypes with node color. Haplotype effect estimates are similar for phylogenetically similar haplotypes, meaning that there was sharing of information between the haplotypes, even though haplotypes that had direct links with phenotype observations (nodes labelled with 1) were separated from each other with a substantial number of mutations.

Haplotypes without direct links to observed phenotypes were estimated with larger uncertainty. In **Figure 5**, we present the posterior standard deviation for the effect of mitochondrial haplotypes with node size. Haplotypes with direct links to observed phenotypes (nodes labelled with 1) have smaller posterior standard deviation than the other haplotypes (nodes labelled with 0). The posterior standard deviation decreased slightly as the haplotypes without direct links were closer (in number of mutations) to the haplotypes with direct links, which was expected.

The posterior distribution for the autoregression parameter $\rho$ indicated strong dependency between haplotype effects. The posterior distribution (full line) of $\rho$ is shown in **Figure 6** together with the prior distribution (dashed line). The mode of the distribution lies around 0.85, and the mean lies around 0.73, indicating that neighboring haplotypes had similar effects, which is related to the sharing of information between haplotypes seen in **Figure 5**. We also note that the posterior distribution shifted to slightly lower values of $\rho$ than the prior distribution. This means that the data contained information that the model could learn from.

A significant amount of the total phenotypic variation was explained by the mitochondrial haplotypes. In **Table 3**, we present the posterior mean and 95% confidence interval of each variance component in the model, and how much of the total variation in the model ($\sigma_c^2 + \sigma_a^2 + \sigma_{h_m}^2 + \sigma_e^2$) was explained by each variance component. The posterior

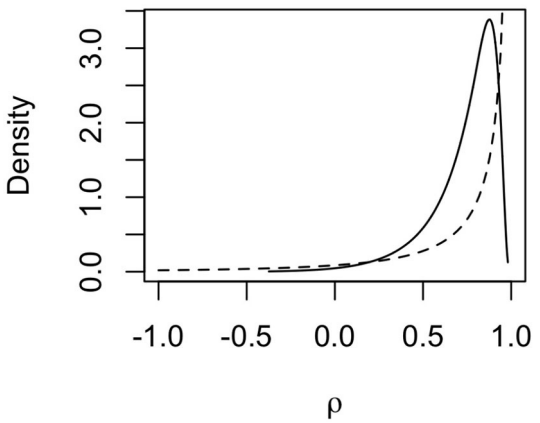


**FIGURE 6 |** Prior (dashed line) and posterior (solid line) distribution for the autocorrelation parameter $\rho$ for mitochondrial haplotype effects on milk yield in cattle.

**TABLE 3 |** Posterior mean, 95% confidence interval (CI) for variance parameters, and the proportion of variation explained by each variance component for the case study estimating mitochondrial haplotype effects on milk yield in cattle.

| Variance parameter | Mean | 95% CI | Prop. of variance explained |
|---|---|---|---|
| $\sigma^2_c$ | 0.035 | (0.005, 0.090) | 0.047 |
| $\sigma^2_a$ | 0.329 | (0.194, 0.533) | 0.444 |
| $\sigma^2_{h_m}$ | 0.113 | (0.033, 0.264) | 0.152 |
| $\sigma^2_{h_c}$ | 0.048 | (0.007, 0.154) | 0.065 |
| $\sigma^2_e$ | 0.265 | (0.171, 0.416) | 0.357 |

*$\sigma^2_c$, variance of contemporary group effects; $\sigma^2_a$, variance of nuclear-genome additive effects; $\sigma^2_{h_m}$, marginal variance of mitogenome haplotype effects; $\sigma^2_{h_m}$, conditional variance of mitogenome haplotype effects; $\sigma^2_e$, variance of residuals.*

distribution of the conditional haplotype variance was obtained by computing $\sigma^2_{h_c} = \sigma^2_{h_m}(1-\rho^2)$, using 10,000 samples from the posterior distributions of the marginal haplotype variance and the autocorrelation parameter. We see that the marginal haplotype variance $\sigma^2_{h_m}$ and conditional haplotype variance $\sigma^2_{h_c}$ is smaller compared to the additive genetic variance $\sigma^2_a$, and the residual variance $\sigma^2_e$. This was expected as the mitogenome ($\sim 1 \times 16Kbp$) is much smaller than the nuclear genome ($\sim 2 \times 3Gbp$). In the light of this difference we can say that mitochondrial haplotypes captured a significant amount of phenotypic variation. The variance for the random effect of herd-year-season of calving $\sigma^2_c$ was also smaller compared to $\sigma^2_a$ and $\sigma^2_e$.

It should be noted that this is a small data set with few haplotypes with direct links to observed phenotypes, which means that the posterior standard deviations for haplotype effects were relatively large. This also causes posterior estimates to be strongly influenced by the prior distributions, especially the posterior for $\rho$ which we can see in **Figure 6**. However, we still chose to assign an informative prior to $\rho$, since it is expected that most mutations have no causal effect and that phylogenetically similar haplotypes have similar effects.

## 3.3. Computation Time

The models were run on a computation server with Linux operating system, 24 cores (4x6 core 2.66 GHz Intel Xeon X7542) and 256 GB memory, fitting up to seven models in parallel. The R version used to produce the results was 3.6.0, and the INLA package version was 18.07.12. INLA was allowed to use as many threads as were available.

In the simulation study, the average computation time was 359.3 s with the HN model, 3.4 s with the IH model, and 4.4 s with the mutation model when the data were simulated from the haplotype network model. When the data were simulated from the mutation model, the average computation time was 310.4 s for HN model, 3.2 s for the IH model and 1.4 s for the mutation model. For the case study with mitochondrial haplotypes, the computation time with the HN model was 119 s.

# 4. DISCUSSION

The objective of this paper was to propose a hierarchical model that leverages haplotype phylogeny to improve the estimation of haplotype effects. We have presented the haplotype network model, evaluated it using simulated data from two different generative models, and applied it in a case study of estimating the effect of mitochondrial haplotypes on milk yield in cattle. We highlight three points for discussion in relation to the proposed haplotype network model: (1) the importance of the haplotype network model, (2) future development and possible extensions and (3) limitations.

## 4.1. The Importance of the Haplotype Network Model

We see three important advantages of the haplotype network model; the ability to share information between related haplotypes, computational advantages when modelling a single region of a genome, and the potential to capture background specific mutation effects.

The haplotype network model utilises phylogenetic relationships between haplotypes and with this improves estimation of their effects. From the simulation study, we saw the importance of this information sharing when there is limited information per haplotype. In the haplotype network model the autocorrelation parameter $\rho$ and the conditional variance parameter $\sigma^2_{h_c}$ reflect the covariance between effects of phylogenetically similar haplotypes. As the autocorrelation approaches 1, haplotype effects become more dependent. Further, if the conditional variance is small the large dependency and small deviations lead to similar effects for phylogenetically similar haplotypes, suggesting that mutations separating the haplotypes have very small or no effect compared to other shared mutations between haplotypes. If on the other hand conditional variance is large, the large dependency and large deviations lead to haplotype effects that change rapidly along the phylogeny, suggesting that mutations separating the haplotypes have large effects. On the other hand, if the autocorrelation parameter approaches 0, the covariance between effects of phylogenetically

similar haplotypes is decreasing, suggesting that haplotypes should be modelled independently.

The three extreme scenarios of hyper-parameter values could denote three real cases. The first case with high autocorrelation and small conditional variance could reflect a situation where the whole haplotype sequence would be used to build a phylogeny and since most mutations do not have a causal effect, but some do, it is expected that similar haplotypes will have similar effects with small differences between the haplotypes. The second case with high autocorrelation and large conditional variance could reflect the situation when the number of causal mutations would be high compared to all mutations (because only such mutations are analysed) and therefore change of effects along the phylogeny would be larger. The third scenario with no autocorrelation could reflect the situation where phylogeny does not correlate with phenotype change.

As mentioned in the introduction, modelling phenotypic variation as a function of haplotype variation has extensive literature (Templeton et al., 1987; Balding, 2006; Thompson, 2013; Morris and Cardon, 2019). The prime motivation for this work is the recent growth in the generation of large scale genomic data sets and methods to build phylogenies (Kelleher et al., 2019). We aimed to develop a general model that could exploit phylogenetic relationships between haplotypes in a computationally efficient way. The computational benefits come from the sparse precision matrix $\mathbf{V}_h^{-1}$, which is due to the conditional independence structure encoded in the DAG of a network of haplotypes (Rue and Held, 2005). The computational benefits are not critical when the number of haplotypes is small. In that case the matrix $\mathbf{V}_h$ is small and easy to invert, though for the autoregressive model we would have to invert it many times during the estimation procedure due to dependency on the autocorrelation parameter. However, it is better to avoid inversions if possible because it can lead to numerical errors and loss of precision (e.g., Misztal, 2016).

While the haplotype network model is different to the pedigree mixed model (Henderson, 1976; Quaas, 1988) (where we model the inheritance of whole genomes in a pedigree without (fully) observing the genomes) or the phylogenetic mixed model (Lynch, 1991; Pagel, 1999; Housworth et al., 2004; Hadfield and Nakagawa, 2010) (where we model the inheritance of whole genomes in a phylogeny without (fully) observing the genomes), the principles of conditional dependence between genetic effects and the resulting sparsity are the same (Rue and Held, 2005). The key difference of the haplotype network model is that it estimates the effect of observed haplotype sequences as compared to unobserved or partially observed inheritance of whole genomes in a pedigree or phylogeny. To improve the estimation of the haplotype effects we take into account the phylogenetic relationships. A similar model has also been used in spatial disease mapping (Datta et al., 2019), showing potential of this kind of model in several applications.

While the use of phylogenetic relationships might seem redundant if we know (most of) the haplotype sequence, the simulations showed that it improves estimation in most cases, even marginally compared to the mutation model where we directly model mutation effects. The haplotype network model can be seen as a hybrid between the mutation model (that models variation between the columns of a haplotype matrix) and the independent haplotype model (that models variation between the rows of a haplotype matrix). This hybrid view might improve genome-wide association studies (see reviews by Gibson, 2018; Simons et al., 2018; Morris and Cardon, 2019; Uricchio, 2019).

The haplotype network model has the potential to capture background specific mutation effects, which are effects observed when the effect of a mutation depends on other mutations present in an individual (e.g., Chandler et al., 2017; Steyn et al., 2019; Wojcik et al., 2019). If there are background specific mutation effects the haplotype effect differences will capture this, while a mutation model only estimates an average effect of a mutation across multiple backgrounds (haplotypes). However, we must point that the haplotype network model captures only local effects, that are due to interactions between mutations present on a haplotype (e.g., Clark, 2004; Liu et al., 2019). We have not evaluated how well the model captures background specific mutation effects in this study, and more simulations to a range of data are needed to evaluate this aspect.

## 4.2. Future Development and Possible Extensions

There is a number of areas for future development with the haplotype network model. We are looking into four areas: making the model more flexible in the number of mutations separating phylogentically similar haplotypes, modelling haplotype differences in a continuous way utilising branch lengths, incorporating biological information and phylogenetic aspects of haplotype relationships.

We have developed the haplotype network model by assuming the differences between similar haplotypes is due to one mutation to simplify model definition. However, in the observed data there might not be haplotypes that are separated for just one mutation. We handle this situation by inserting phantom haplotypes, to ensure that we do not model haplotypes as more similar than they actually are. The order of mutations in such situations is uncertain and a model could be generalised to account for these larger number of mutations between haplotypes. However, the current "one-mutation" difference model setup has a useful property of inferring the value of unobserved haplotypes and the sparse model definition does not increase computational complexity of the model.

The haplotype network model could be generalised to utilise time calibrated distances between haplotypes rather than using the number of mutations. The Ornstein-Uhlenbeck (OU) process is the continuous-time analogue of the autoregressive process of order one used in this study, and plays a major role in the analysis of the evolution of phenotypic traits along phylogenies (Lande, 1976; Hansen and Martins, 1996; Martins and Hansen, 1997; Paradis, 2014). Relatedly, if the autocorrelation parameter of the autoregressive process of order one is set to 1 we get the non-stationary discrete random walk process, whose continuous-time analogue is the Brownian process that is the basic model of phylogenetic comparative analysis (Felsenstein, 1988; Huey et al., 2019). There is a scope to improve computational aspects for

these continuous models too by employing recent developments from the statistical analysis of irregular time-series (Lindgren and Rue, 2008).

In the haploptype network model presented in this study, the same autocorrelation parameter has been assumed for all mutations. However, the autocorrelation parameter could be allowed to vary as Beaulieu et al. (2012) did in the context of adaptive evolution. For example, different autocorrelation parameters for different types of mutations could incorporate biological information, which could combine the quantitative analysis of mutation and haplotype effects with molecular genetic tools such as Variant Effect Predictor (McLaren et al., 2016).

We have assumed that the phylogenetic network is given and described with a DAG. There is a large body of literature on inferring phylogenies in the form of strict bifurcating trees, more general trees or networks and recent developments in genomics are rapidly advancing the field (e.g., Anisimova, 2012; Puigbò et al., 2013; Schliep et al., 2017; Uyeda et al., 2018). The haplotype network model can work both with phylogenetic bifurcating and multifurcating trees and phylogenetic networks. The only condition is that we describe the haplotype relationships with a DAG, an output provided by many tools (e.g., Leigh and Bryant, 2015; Suchard et al., 2018; Kelleher et al., 2019). We have generalised the model construction to allow for network structures. This generalisation enables the model to describe haplotype relationships without paying attention to the directionality as long as there are no directed loops in the graph. The proposed model does not depend on which allele is ancestral, major or minor, but we believe that the most logical is to work with ancestral alleles as the starting point.

It is beneficial to know the order of mutations, and therefore which haplotypes are parental to other haplotypes, because this leads to a tree structure and a sparse precision matrix structure in the model (Rue and Held, 2005). An example of non-optimal sparsity can be seen in our case study. In **Figure 5**, the "central" haplotype with the largest uncertainty is modelled as a progeny haplotype of four surrounding haplotypes, which means that there is a dense $5 \times 5$ block in the precision matrix $\mathbf{V}_h^{-1}$. The block is dense because the "central" haplotype is modelled as a function of the other four "parental" haplotypes. If however the "central" haplotype was used as the parental haplotype the $5 \times 5$ block would be sparse since all other haplotypes would be conditionally independent given the "central/parental" haplotype. The same applies also for the other parts of the haplotype network in **Figure 5**.

The haploptype network model could also work with probabilistic networks where edges have associated uncertainty (weights). By encoding such a network with a DAG, the edge weights can be used in model construction—for example, in the same way uncertain parentage is handled in pedigree models (Henderson, 1976). An alternative would be to construct a model for each possible realisation of a network, run separate models and combine haplotype estimates in the spirit of Bayesian model averaging.

## 4.3. Limitations

The haplotype network model also has some limitations that merit further development. We highlight three areas: is the haplotype network model necessary given that we can model mutation effects, Gaussian assumption and causal mutations, and modelling recombining haplotypes.

For the haplotype network model to achieve its full potential, the data need to have a certain structure. We saw from fitting the haplotype network model to a real data set, that having few haplotypes with direct links to observed phenotypes and many haplotypes without, lead to large uncertainty in estimated haplotype effects. We also saw from fitting simulated data, that the mutation model was slightly better at estimating the mutation effects than the haplotype network model, when the data were simulated from a mutation model, but the magnitude of difference was minimal. In the future, different data structures should be tested to find optimal scenarios, in order for the haplotype network model to achieve its full potential.

The haplotype network model assumes that the haplotype effects follow a Gaussian distribution. If all, or very many, of the haplotypes have the same effect, the distribution may be quite different from Gaussian, which breaks the model assumptions and perhaps other models should be proposed. Blomberg et al. (2019) describe the underlying theory behind the common Gaussian processes, such as Brownian motion and Ornstein-Uhlenbeck process, and present general methods for deriving new stochastic models, including non-Gaussian models of quantitative trait macroevolution. See also (Landis et al., 2012; Schraiber and Landis, 2015; Duchen et al., 2017; Bastide et al., 2020).

Scaling the haplotype network model to multiple recombining haplotype regions is challenging for two reasons. First, while phasing methods have improved substantially in the last years (Marchini, 2019), determining a recombination breakpoint is challenging due to a limited resolution to resolve exact locus where recombination occurred (Johnsson et al., 2020). Second, the sparsity of the haplotype network model comes from the sparsity of the precision matrix $\mathbf{V}_h^{-1}$. In the extension for recombining haplotypes the sparsity in the prior is maintained also for multiple consequitive haplotype regions along a chromosome as shown in Equation (11) in section 2.1.4. However, the design matrices that link phenotype observations with multiple haplotype regions create dense cross-products in the system of equations as we increase the number of regions and the sparsity advantage is lost. To this end we are exploring alternative ways of formulating the haplotype network model following data structures in Kelleher et al. (2019), with the aim to improve upon the existing haplotype based genomic modelling of whole genomes (e.g., Villumsen et al., 2009; Hickey et al., 2013).

# DATA AVAILABILITY STATEMENT

The datasets analysed in this article are not publicly available. Requests to access the datasets should be directed to Vladimir Brajkovic, vbrajkovic@agr.hr.

## AUTHOR CONTRIBUTIONS

MLS, FL, and GG conceived and derived the haplotype network model. MLS, IS, and GG designed the analysis, and evaluated the results. MLS simulated data and performed all analyses. VB and VC-C provided case study data. MLS wrote the manuscript. FL, IS, and GG commented on and edited the manuscript. All authors have read and approved the final manuscript.

## FUNDING

MLS and IS acknowledge the support from The Research Council of Norway, Grant Number: 250362. This work by VB and VC-C was supported by the Croatian science Foundation under the Project MitoTAUROmics-IP-11-2013_9070 Utilisation of the whole mitogenome in cattle breeding and conservation genetics and project ANAGRAMS-IP-2018-01-8708 Application of NGS in assessment of genomic variability in ruminants (https://angen.agr.hr/). GG acknowledges the support from the Biotechnology and Biological Sciences Research Council (BBSRC; Swindon, UK) funding to The Roslin Institute (BBS/E/D/30002275), and The University of Edinburgh's Data-Driven Innovation Chancellors fellowship.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.531218/full#supplementary-material

**Supplemental 1 |** R script and data file to simulate the DAG describing the phylogeny of simulated haplotypes, to simulate from the haplotype network model, and to fit the haplotype network model to the data.

## REFERENCES


Anisimova, M. (2012). *Evolutionary Genomics Statistical and Computational Methods*. New York, NY: Springer. doi: 10.1007/978-1-61779-585-5

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet*. 7:781. doi: 10.1038/nrg1916

Basseville, M., Benveniste, A., Chou, K. C., Golden, S. A., Nikoukhah, R., and Willsky, A. S. (1992). Modeling and estimation of multiresolution stochastic processes. *IEEE Trans. Inform. Theory* 38, 766–784. doi: 10.1109/18.119735

Bastide, P., Ho, L. S. T., Baele, G., Lemey, P., and Suchard, M. A. (2020). Efficient bayesian inference of general gaussian models on large phylogenetic trees. *arXiv [Preprint]* arXiv:2003.10336.

Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. C. (2012). Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evol. Int. J. Organ. Evol.* 66, 2369–2383. doi: 10.1111/j.1558-5646.2012.01619.x

Begum, R. (2019). A decade of genome medicine: toward precision medicine. *Genome Med*. 11. doi: 10.1186/s13073-019-0624-z. [Epub ahead of print].

Blangiardo, M., and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Chichester: John Wiley and Sons. doi: 10.1002/9781118950203

Blomberg, S. P., Rathnayake, S. I., and Moreau, C. M. (2019). Beyond brownian motion and the Ornstein-Uhlenbeck process: stochastic diffusion models for the evolution of quantitative characters. *Am. Natural.* 195, 000–000. doi: 10.1086/706339

Brajković, V. (2019). *Utjecaj mitogenoma na svojstva mliječnosti goveda (Eng: Impact of mitogenome on milk traits in cattle)* (Ph.D. thesis). University of Zagreb, Faculty of Agriculture, Zagreb, Croatia.

Chandler, C. H., Chari, S., Kowalski, A., Choi, L., Tack, D., DeNieu, M., et al. (2017). How well do you know your mutation? complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. *PLoS Genet*. 13:e1007075. doi: 10.1371/journal.pgen.1007075

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genet. Epidemiol*. 27, 321–333. doi: 10.1002/gepi.20025

Datta, A., Banerjee, S., Hodges, J. S., and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (dagar) models. *Bayesian Anal*. 14, 1221–1244. doi: 10.1214/19-BA1177

de los Campos, G., Vazquez, A. I., Hsu, S., and Lello, L. (2018). Complex-trait prediction in the era of big data. *Trends Genet*. 34, 746–754. doi: 10.1016/j.tig.2018.07.004

Duchen, P., Leuenberger, C., Szilágyi, S. M., Harmon, L., Eastman, J., Schweizer, M., et al. (2017). Inference of evolutionary jumps in large phylogenies using Lévy processes. *Syst. Biol*. 66, 950–963. doi: 10.1093/sysbio/syx028

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Popul. Biol*. 3, 87–112. doi: 10.1016/0040-5809(72)90035-4

Ewens, W. J. (2004). *Mathematical Population Genetics 1, 2nd Edn*. New York, NY: Springer-Verlag. doi: 10.1007/978-0-387-21822-9

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst*. 19, 445–471. doi: 10.1146/annurev.es.19.110188.002305

Gardiner, C. (2009). *Stochastic Methods. A Handbook for the Natural and Social Sciences, 4th Edn*. Berlin; Heidelberg: Springer.

Gibson, G. (2018). Population genetics and gwas: a primer. *PLoS Biol*. 16:e2005485. doi: 10.1371/journal.pbio.2005485

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc*. 102, 359–378. doi: 10.1198/016214506000001437

Hadfield, J., and Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol*. 23, 494–508. doi: 10.1111/j.1420-9101.2009.01915.x

Hansen, T. F., and Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50, 1404–1417. doi: 10.1111/j.1558-5646.1996.tb03914.x

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32, 69–83. doi: 10.2307/2529339

Hickey, J., Kinghorn, B., Tier, B., Clark, S. A., van der Werf, J., and Gorjanc, G. (2013). Genomic evaluations using similarity between haplotypes. *J. Anim. Breed. Genet*. 130, 259–269. doi: 10.1111/jbg.12020

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet*. 49:1297. doi: 10.1038/ng.3920

Housworth, E. A., Martins, E. P., and Lynch, M. (2004). The phylogenetic mixed model. *Am. Natural*. 163, 84–96. doi: 10.1086/380570

Huey, R. B., Garland, T. Jr., and Turelli, M. (2019). Revisiting a key innovation in evolutionary biology: Felsenstein's "phylogenies and the comparative method". *Am. Natural*. 193, 755–772. doi: 10.1086/703055

Ibanez-Escriche, N., and Simianer, H. (2016). Animal breeding in the genomics era [Special issue]. *Anim. Front*. 6, 4–5. doi: 10.2527/af.2016-0001

Johnsson, M., Whalen, A., Ros-Freixedes, R., Gorjanc, G., Chen, C.-Y., Herring, W. O., et al. (2020). Genetics of recombination rate variation in the pig. *bioRxiv*. doi: 10.1101/2020.03.17.995969

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol*. 12:e1004842. doi: 10.1371/journal.pcbi.1004842

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nat. Genet*. 51, 1330–1338. doi: 10.1038/s41588-019-0483-y

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., et al. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9780429031892

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30, 314–334. doi: 10.1111/j.1558-5646.1976.tb00911.x

Landis, M. J., Schraiber, J. G., and Liang, M. (2012). Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. *Syst. Biol.* 62, 193–204. doi: 10.1093/sysbio/sys086

Leigh, J. W., and Bryant, D. (2015). Popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D. (2018). Accurate genomic prediction of human height. *Genetics* 210, 477–497. doi: 10.1534/genetics.118.301267

Lindgren, F., and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scand. J. Stat.* 35, 691–700. doi: 10.1111/j.1467-9469.2008.00610.x

Liu, F., Schmidt, R. H., Reif, J. C., and Jiang, Y. (2019). Selecting closely-linked snps based on local epistatic effects for haplotype construction improves power of association mapping. *Genes Genomes Genet.* 9, 4115–4126. doi: 10.1534/g3.119.400451

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45, 1065–1080. doi: 10.1111/j.1558-5646.1991.tb04375.x

Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., et al. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* 9:989. doi: 10.1038/s41467-017-02769-6

Marchini, J. (2019). "Haplotype estimation and genotype imputation," in *Handbook of Statistical Genomics*, eds D. Balding, I. Moltke, and J. Marioni (Oxford: John Wiley & Sons Ltd.), 87–114. doi: 10.1002/9781119487845.ch3

Martins, E. P., and Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. ist* 149, 646–667. doi: 10.1086/286013

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

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. Available online at: https://www.genetics.org/content/genetics/157/4/1819

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202, 401–409. doi: 10.1534/genetics.115.182089

Morris, A. P., and Cardon, L. R. (2019). "Chapter 21: Genome-wide association studies," in *Handbook of Statistical Genomics: Two Volume Set, 4th Edn.* eds D. Balding, I. Moltke, and J. Marioni (Oxford: John Wiley and Sons, Ltd.), 597–550. doi: 10.1002/9781119487845.ch21

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401:877. doi: 10.1038/44766

Paradis, E. (2014). "Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice," in *Simulation of Phylogenetic Data*, ed L. Z. Garamszegi (Berlin; Heidelberg: Springer), 335–350. doi: 10.1007/978-3-662-43550-2_13

Puigbó, P., Wolf, Y. I., and Koonin, E. V. (2013). Seeing the tree of life behind the phylogenetic forest. *BMC Biol.* 11:46. doi: 10.1186/1741-7007-11-46

Quaas, R. (1988). Additive genetic model with groups and relationships. *J. Dairy Sci.* 71, 1338–1345.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rue, H., and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9780203492024

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* 71, 319–392. doi: 10.1111/j.1467-9868.2008.00700.x

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annu. Rev. Stat. Appl.* 4, 395–421. doi: 10.1146/annurev-statistics-060116-054045

Schliep, K., Potts, A. A., Morrison, D. A., and Grimm, G. W. (2017). Intertwining phylogenetic trees and networks. *Methods Ecol. Evol.* 8, 1212–1220. doi: 10.1111/2041-210X.12760

Schraiber, J. G., and Landis, M. J. (2015). Sensitivity of quantitative traits to mutational effects and number of loci. *Theoret. Popul. Biol.* 102, 85–93. doi: 10.1016/j.tpb.2015.03.005

Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A population genetic interpretation of gwas findings for human quantitative traits. *PLoS Biol.* 16:e2002985. doi: 10.1371/journal.pbio.2002985

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: a principled, practical approach to constructing priors. *Stat. Sci.* 32, 1–28. doi: 10.1214/16-STS576

Sørbye, S. H., and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *J. Time Ser. Anal.* 38, 923–935. doi: 10.1111/jtsa.12242

Steyn, Y., Lourenco, D. A. L., and Misztal, I. (2019). Genomic predictions in purebreds with a multi-breed genomic relationship matrix. *J. Anim. Sci.* 97, 4418–4427. doi: 10.1093/jas/skz258.099

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Templeton, A. R., Boerwinkle, E., and Sing, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* 117, 343–351.

Thompson, K. L. (2013). *Using ancestral information to search for quantitative trait loci in genome-wide association studies* (Ph.D. thesis). The Ohio State University. Columbus, OH, United States. doi: 10.1186/1471-2105-14-200

Uricchio, L. H. (2019). Evolutionary perspectives on polygenic selection, missing heritability, and gwas. *Hum. Genet.* 139, 5–21. doi: 10.1007/s00439-019-02040-6

Uyeda, J. C., Zenil-Ferguson, R., and Pennell, M. W. (2018). Rethinking phylogenetic comparative methods. *Syst. Biol.* 67, 1091–1109. doi: 10.1093/sysbio/syy031

Villumsen, T. M., Janss, L., and Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126, 3–13. doi: 10.1111/j.1439-0388.2008.00747.x

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198830870.001.0001

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 570, 514–518. doi: 10.1038/s41586-019-1310-4

Wu, P., Hou, L., Zhang, Y., and Zhang, L. (2020). Phylogenetic tree inference: a top-down approach to track tumor evolution. *Front. Genet.* 10:1371. doi: 10.3389/fgene.2019.01371

**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.

# 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