frontiers in Genetics

ORIGINAL RESEARCH
published: 29 July 2021
doi: 10.3389/fgene.2021.665382

Genomic Prediction Using LD-Based Haplotypes Inferred From High-Density Chip and Imputed Sequence Variants in Chinese Simmental Beef Cattle

Hongwei Li[1]†, Bo Zhu[1,2]†, Ling Xu[1], Zezhao Wang[1], Lei Xu[1], Peinuo Zhou[1], Han Gao[1], Peng Guo[3], Yan Chen[1], Xue Gao[1], Lupei Zhang[1], Huijiang Gao[1,2], Wentao Cai[1], Lingyang Xu[1]* and Junya Li[1,2]*

[1] Laboratory of Molecular Biology and Bovine Breeding, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] National Centre of Beef Cattle Genetic Evaluation, Beijing, China, [3] College of Computer and Information Engineering, Tianjin Agricultural University, Tianjin, China

A haplotype is defined as a combination of alleles at adjacent loci belonging to the same chromosome that can be transmitted as a unit. In this study, we used both the Illumina BovineHD chip (HD chip) and imputed whole-genome sequence (WGS) data to explore haploblocks and assess haplotype effects, and the haploblocks were defined based on the different LD thresholds. The accuracies of genomic prediction (GP) for dressing percentage (DP), meat percentage (MP), and rib eye roll weight (RERW) based on haplotype were investigated and compared for both data sets in Chinese Simmental beef cattle. The accuracies of GP using the entire imputed WGS data were lower than those using the HD chip data in all cases. For DP and MP, the accuracy of GP using haploblock approaches outperformed the individual single nucleotide polymorphism (SNP) approach (GBLUP_In_Block) at specific LD levels. Hotelling's test confirmed that GP using LD-based haplotypes from WGS data can significantly increase the accuracies of GP for RERW, compared with the individual SNP approach (∼1.4 and 1.9% for G_HBLUP and G_HBLUP+GBLUP, respectively). We found that the accuracies using haploblock approach varied with different LD thresholds. The LD thresholds ($r^2 \geq 0.5$) were optimal for most scenarios. Our results suggested that LD-based haploblock approach can improve accuracy of genomic prediction for carcass traits using both HD chip and imputed WGS data under the optimal LD thresholds in Chinese Simmental beef cattle.

Keywords: genomic prediction, prediction accuracy, LD, haplotype, Chinese Simmental beef cattle

Edited by:
Luiz Brito,
Purdue University, United States

Reviewed by:
Li Ma,
University of Maryland, College Park, United States
Fernando Baldi,
São Paulo State University, Brazil
Beatriz Cuyabano,
Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), France

Correspondence:
Lingyang Xu
xulingyang@163.com
Junya Li
lijunya@caas.cn

†These authors have contributed equally to this work

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

Received: *08 February 2021*
Accepted: *30 June 2021*
Published: *29 July 2021*

Citation:
Li H, Zhu B, Xu L, Wang Z, Xu L, Zhou P, Gao H, Guo P, Chen Y, Gao X, Zhang L, Gao H, Cai W, Xu L and Li J (2021) Genomic Prediction Using LD-Based Haplotypes Inferred From High-Density Chip and Imputed Sequence Variants in Chinese Simmental Beef Cattle. Front. Genet. 12:665382. doi: 10.3389/fgene.2021.665382

INTRODUCTION

Genomic prediction (GP) has been widely used in the past decades (Meuwissen et al., 2001). Many approaches, including GBLUP (VanRaden, 2008), Bayes alphabet (Habier et al., 2011; Gianola, 2013), and machine learning (Li et al., 2018; Yin et al., 2020), have been proposed to improve prediction accuracy. Most of these approaches were developed based on single nucleotide polymorphisms (SNPs). Genomic prediction using haplotypes instead of SNPs can be more accurate (Zondervan and Cardon, 2004). A haplotype is defined as a combination of alleles at adjacent loci belonging to the same chromosome that are transmitted as a unit (Vormfelde and Brockmöller, 2007; Won et al., 2020) and a haplotype may contain the combined effects of causal variants with high linkage disequilibrium (LD) (Balding, 2006; Garnier et al., 2013), thus this approach can effectively identify the loci with small effects, which may not be captured by a single marker (Feitosa et al., 2020).

Many previous studies have shown that genomic selection using haplotypes is more reliable than that using individual SNPs for both simulated and real data, even when the marker density is low (Calus et al., 2008; De Roos et al., 2008). Cuyabano et al. (2014) compared the genomic predictions between the haplotype-based (constructed based on LD and using HD chip data) and the SNP-based approach for milk production and health traits in dairy cattle, suggesting the high prediction ability using the haplotype-based approach. Moreover, Hess et al. (2017) found that fitting covariates for haplotype alleles instead of SNPs can increase the prediction accuracy up to 5.5% (Hess et al., 2017). Recently, Xu et al. (2020) reported that the haplotype-based model using HD chip data can improve the accuracy by 5.4–9.8%, compared with the SNP-based approach for carcass and live weight traits.

Haploblocks can be constructed through multiple strategies including the fixed block length based on centimorgans (cM) (Boichard et al., 2012), base pairs (bp) (Sun, 2016), or a constant number of SNPs per block (Hayes et al., 2007; Calus et al., 2009; Villumsen et al., 2009) and not fixed length approach based on the LD pattern (Cuyabano et al., 2015). Many improved methods have been proposed to account for recombination hotspots and coldspots across the genome (Calus et al., 2008; Sandor et al., 2012; Weng et al., 2014; Cuyabano et al., 2015). Haploblock construction based on the LD is expected to achieve a high prediction accuracy by selecting the effective SNPs and reducing the amount of predictor variables in the model (Cuyabano et al., 2015).

The WGS data can provide more potential causative polymorphisms, thus imputation from low density marker panels to WGS for datasets with a large number of individuals may be an effective approach to increase the accuracy of GP (Marchini et al., 2007; Browning and Browning, 2009; Howie et al., 2009; Li et al., 2010). A recent study suggested that genomic prediction within-population using simulated WGS data can increase (∼31%) the accuracy of prediction for traits with low and moderate heritability (Iheshiulor et al., 2016). Similarly, Druet et al. (2014) suggested that the prediction accuracy using simulated sequence data can be improved (∼30%) when including causal mutations with low minor allele frequencies. A previous study suggested that the haploblock approach may play an important role in the genomic prediction involving genome sequences (Cuyabano et al., 2014). The haploblocks containing additional markers are likely to be generated from WGS, which may reduce the number of variables compared with SNP and keep all SNP information. The haplotype approach based on WGS is likely to improve the accuracy of GP. However, the evaluations of prediction accuracies on the economically important traits using this strategy are still yet to be explored in cattle.

The objectives of current study were to (1) evaluate the predictive performance of carcass traits using HD chip and WGS data in Chinese Simmental beef cattle; (2) compare the differences of predictive accuracies between haplotype-based prediction model (G_HBLUP), SNP-based prediction model (GBLUP), and the combination of haplotype and SNP prediction model (G_HBLUP+GBLUP); and (3) investigate the LD-based haplotypes with different thresholds on the prediction accuracies.

MATERIALS AND METHODS

Ethics Statement

All animals used in the study were treated following the guidelines established by the Council of China Animal Welfare. The procedure for collecting cattle blood samples and phenotypes was carried out in strict accordance with the protocol approved by the Science Research Department of the Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (CAAS) (Beijing, China).

Data

Data available comprised a total of 1,233 Simmental cattle born between 2008 and 2015 from Ulgai, Xilingol League, and Inner Mongolia, China. After weaning, cattle were moved to Jinweifuren Co., Ltd. (Beijing, China) for fattening under the same feeding and management conditions. A more detailed description of the management processes was reported in previous studies (Zhu et al., 2016, 2017). All individuals were slaughtered at an average age of 20 ± 2.2 months. Carcass and meat quality traits were measured in accordance with the guidelines proposed by the Institute of Meat Purchase Specifications established by the Agricultural Marketing Service of the USDA. From these traits, dressing percentage (DP), meat percentage (MP), and rib eye roll weight (RERW) were analyzed.

Genotyping and Imputation

The DNA samples from blood were genotyped with Illumina BovineHD BeadChip. Before statistical analysis, the original SNP dataset was filtered using PLINK (v1.07) (Purcell et al., 2007; Chang et al., 2015). Individuals and autosomal SNPs were filtered by the following criteria: SNP call rate (<0.90), minor allele frequency (MAF < 0.01), Hardy–Weinberg equilibrium ($p < 10^{-6}$), and individual call rate (<0.90). Missing genotypes were imputed using BEAGLE (v4.1) (Browning and Browning, 2016). Consequently, 1,233 individuals and 671,164 SNPs remained.

Forty-four unrelated individuals (according to the pedigree and PI-HAT value estimated using PLINK v1.07) were selected as the reference population for imputation. The whole genome sequencing of these individuals was performed using Illumina Hiseq2500 instruments (Illumina Inc., San Diego, CA, United States). All processes were performed according to the standard manufacturer's protocols.

The SNPs from the HD chip were imputed to the sequencing level using BEAGLE (v4.1) (Browning and Browning, 2016). The imputed WGS was filtered by removing SNPs with a MAF less than 0.05. After quality control, a total of 6,776,719 SNPs remained. The imputation accuracy was assessed by the allelic R-squared measure (AR^2), which is an estimate of the squared correlation between the most probable and the true reference dose. The average imputation accuracy was 0.83 when the MAF was larger than 0.05.

Heritability and Variance Component Estimation

Phenotypes were adjusted for the fixed effects, including sex, year, and the covariates of body weight upon entering the fattening farm, and the number of fattening days. Subsequently, the adjusted phenotypes were used for further analyses. Variance components were estimated using the following univariate animal model in ASREML (v4.1).

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Za} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the vector of the adjusted phenotypes, $\mathbf{1_n}$ is an n × 1 vector with entries equal to 1; μ is the overall mean; $\mathbf{a} \sim N\left(0, \sigma_a^2\mathbf{G}\right)$ is a vector of random additive genetic effect, where $\mathbf{G}$ is the additive genomic relationship matrix constructed using all SNPs and σ_e^2 is the additive genetic variance, $\mathbf{Z}$ is incidence matrix linking $\mathbf{a}$ to $\mathbf{y}$; and $\mathbf{e} \sim N\left(0, \sigma_e^2\mathbf{I}\right)$ is a vector of random residuals, where $\mathbf{I}$ is the identity matrix and σ_e^2 is the residual variance. The heritability estimates were calculated as $\boldsymbol{h}^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$.

Haplotype Construction

The LD-based haploblocks were generated separately for each chromosome. A group of SNPs was defined as a haploblock if the LD between every two SNPs in the group was greater than or equal to the threshold value (r^2). For two bi-allelic loci (A_1/A_2 *and* B_1/B_2), r^2 was calculated as,

$$\boldsymbol{r}^2 = \frac{\boldsymbol{D}^2}{(\boldsymbol{p}_{A_1}\boldsymbol{p}_{A_2}\boldsymbol{p}_{B_1}\boldsymbol{p}_{B_2})} \tag{2}$$

where $\boldsymbol{D} = \boldsymbol{p}_{A_1B_1}\boldsymbol{p}_{A_2B_2} - \boldsymbol{p}_{A_1B_2}\boldsymbol{p}_{A_2\,B_1}$.

Seven different LD levels (r^2) (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8) were set as the thresholds in this study.

Haplotype effects were modeled using numerical dosage coding strategies (Calus et al., 2008; Cuyabano et al., 2014, 2015; Meuwissen et al., 2014; Da, 2015). Numerical dosage coding of a haploblock is formed by two consecutive SNPs (**Table 1**). In the numerical dosage model, artificial SNPs were created for each haploblock, and these "SNPs" were coded as the number of copies.

Genomic Prediction Models

The genomic best linear unbiased prediction (GBLUP) model including the haplotype/SNP effect was used for DP, MP, and RERW as described in Eq. (1). Three approaches based on (a) the SNPs, (b) the haploblock only, and (c) the haploblock and the non-blocked SNPs were considered for predictions. Seven different r^2 thresholds were used for haploblock construction.

TABLE 1 | Numerical dosage coding of a haploblock formed by two consecutive single nucleotide polymorphisms (SNPs).

Haplotype allele 1	Haplotype allele 2	Numerical coding of haploblock			
		AB	Ab	aB	ab
AB	AB	2	0	0	0
AB	Ab	1	1	0	0
AB	aB	1	0	1	0
AB	ab	1	0	0	1
Ab	Ab	0	2	0	0
Ab	aB	0	1	1	0
Ab	ab	0	1	0	1
aB	aB	0	0	2	0
aB	ab	0	0	1	1
ab	ab	0	0	0	2

AB, Ab, aB, and ab are four haplotype alleles of the same haploblock.

We performed genomic prediction using GBLUP for all SNP markers, and the genomic relationship matrix was calculated as $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}\text{-}\mathbf{P})'}{2\sum_{i=1}^{m} p_i\left(1-p_i\right)}$, where $\mathbf{M}$ denotes the (0, 1, 2)-encoded genotype matrix, p_i is the MAF of marker i, m is the number of markers, and $\mathbf{P}$ is a matrix with columns equal to $2p_i$.

Genomic prediction using GBLUP for the SNP markers inside of the block in HD chip and WGS data were defined as GBLUP_770K_In_Block and GBLUP_WGS_In_Block, respectively.

The haplotype-based genomic best linear unbiased prediction (G$_\mathrm{H}$BLUP) was performed for all markers. The haplotype-based genomic relationship matrix in G$_\mathrm{H}$BLUP was constructed as the product of the haplotype allele matrix ($\mathbf{M_H}$) and expressed as $\mathbf{G_H} = \frac{\mathbf{M_H}\mathbf{M_H}'}{\mathrm{Q_H}}$, where $\mathbf{M_H}$ is the pseudo-markers matrix with entries 0, 1, and 2 representing the number of copies of each haplotype allele in a haploblock, and Q$_\mathrm{H}$ is the total number of haplotype alleles of whole genome. In the G$_\mathrm{H}$BLUP+GBLUP model:

$$\boldsymbol{y} = \mathbf{1}_{\boldsymbol{n}}\boldsymbol{\mu} + \boldsymbol{Za} + \boldsymbol{Z_u a_u} + \boldsymbol{e} \tag{3}$$

which included the haploblock effects and the SNP effects estimated from outside the haploblocks (non-blocked SNPs). $\boldsymbol{a} : N\left(0, \sigma_a^2\boldsymbol{G_H}\right)$ is a vector of random additive genetic effect, where $\boldsymbol{G_H}$ is the additive genetic relationship matrix constructed using haploblock and σ_a^2 is the additive genetic variance based on the haploblock, $\mathbf{Z}$ is incidence matrix associating $\boldsymbol{a}$; $\boldsymbol{a_u} : N\left(0, \sigma_{a_u}^2\boldsymbol{G}\right)$ is a vector of random additive genetic effect, where $\boldsymbol{G}$ is the additive genetic relationship matrix constructed using non-blocked SNPs and $\sigma_{a_u}^2$ is the additive genetic variance based on the haploblock, $\boldsymbol{Z_u}$ is incidence matrix associating $\boldsymbol{a_u}$; $\boldsymbol{a}$ is composed of haploblock effects and $\boldsymbol{a_u}$ is composed of SNP effects estimated from outside the haploblocks. Also, they are considered as uncorrelated effects.

Assessment of Prediction Accuracy

The accuracy of genomic prediction was assessed using fivefold cross-validation (CV). The CV procedure was applied by assigning animals randomly into five separate subsets. This procedure was randomly repeated 10 times.

The regression coefficient of the adjusted phenotype on GEBVs for individuals in the validation set was obtained to measure the degree of inflation/deflation of prediction, which was defined as follows:

$$\boldsymbol{b} = \frac{\boldsymbol{Cov(gebv, y^{*})}}{\boldsymbol{var(gebv)}} \tag{4}$$

The average Pearson correlation coefficient between the adjusted phenotypic values and genomic estimated breeding values (GEBVs) in the validation set divided by square root of heritability was used as a measurement of prediction accuracy. The prediction accuracy was calculated as (Bolormaa et al., 2013):

$$\boldsymbol{Prediction\ accuracy} \ = \frac{\boldsymbol{cor(y^{*}, gebv)}}{\sqrt{\boldsymbol{h}^{2}}} \tag{5}$$

where y^{*} is adjusted phenotypic values, *gebv* is the genomic estimated breeding values (GEBVs), and h^2 is the heritability.

To compare the differences of the accuracies of GP using three approaches (GBLUP, $\mathrm{G_H}$BLUP, and $\mathrm{G_H}$BLUP+GBLUP) and marker densities (HD chip and WGS), we used Hotelling's (1940) *t* statistic (Hotelling, 1940) to test the significance of the differences.

The test statistic $\boldsymbol{t}$ is given by,

$$\boldsymbol{t} \ \boldsymbol{=} \frac{(\boldsymbol{r_{jk}} - \boldsymbol{r_{jh}})\sqrt{(\boldsymbol{n}-3)(1+\boldsymbol{r_{kh}})}}{\sqrt{2\,|\boldsymbol{R}|}} \tag{6}$$

with $df \ = \ n-3$, where,

$$|R| \ = \ 1 + 2r_{jk}r_{jh}r_{kh} - r_{jk}^{2} - r_{jh}^{2} - r_{kh}^{2} \tag{7}$$

where $\boldsymbol{r}$ is the observed correlation and n is the number of observations. For instance, while comparing the differences of accuracy between the GBLUP and $\mathrm{G_H}$BLUP, the $\boldsymbol{r_{jk}}$ is the $\mathrm{cor}(y^{*}, gebv_{GBLUP})$, the $\boldsymbol{r_{jh}}$ is the $cor(y^{*}, gebv_{G_H BLUP})$, and the $\boldsymbol{r_{kh}}$ is the $cor(gebv_{GBLUP}, gebv_{G_H BLUP})$. If $\boldsymbol{P}(T \geq \ t) \ \leq \ \alpha(\alpha \ = \ 0.05)$, then the hypothesis $(H_0 : r_{jk} \ = \ r_{jh})$ is rejected. Hence, we can conclude whether correlations were significantly different.

RESULTS

Heritability Estimation and Haploblock Construction

Based on the HD chip data, the estimated heritabilities of DP, MP, and RERW using univariate animal model were 0.27, 0.17, and 0.23, respectively, and the statistical description is shown in **Table 2**. Notably, under threshold $r^2 > 0.2$, we observed 68,775 (362,710 SNPs) and 634,662 (3,536,404 SNPs) blocks from the HD chip and WGS data, while the number of SNPs out of blocks were 298,454 and 3,240,315 and haplotype allele counts were 840,676 and 3,370,157. Details about the total number of haplotype alleles (variables), haploblocks, and non-blocked SNPs with different r^2 are presented in **Table 3**. The number of haplotype alleles and haploblocks decreases with increasing r^2. The average number of SNPs per haploblock ranged from 3.3 to 5.3 for the HD chip data and from 4.5 to 5.6 for the WGS data. According to our results, we found that the method based on haploblock reduced the number of variables (haplotype alleles) for the WGS data. However, as for the HD chip data, the haploblock approach increased the number of variables compared with the SNP approach. This result mainly depends on the data type used for haploblock construction (HD or WGS).

We also evaluated the LD decay between 0 and 100 kb for BTA1 in the HD chip and WGS data, respectively. The average r^2 was calculated for each 1-kb window size. LD decay suggested that the HD chip data had a faster LD decay than WGS data (**Supplementary Figure 1**), thus prediction accuracies using the HD chip data among different LD thresholds displayed obvious changes compared with the WGS data. We observed r^2 decreased from 0.8 to 0.2 as the marker distances of the HD chip (from 0 to 35 kb) and WGS data (from 0 to 25 kb) increased. However, no obvious difference was found when $r^2 < 0.2$. Therefore, we chose the LD thresholds ($r^2 \geq 0.2$ to $r^2 \geq 0.8$) to construct the haploblocks in our study.

Genomic Prediction Accuracy

Comparison of Accuracies of GP Based on Three Different Approaches

Three different approaches, including (a) GBLUP, (b) $\mathrm{G_H}$BLUP, and (c) $\mathrm{G_H}$BLUP+GBLUP, were considered for the comparisons. As shown in **Figure 1**, $\mathrm{G_H}$BLUP+GBLUP had better performance for DP and MP than $\mathrm{G_H}$BLUP. We also found $\mathrm{G_H}$BLUP+GBLUP_770K yielded ~1.8% higher accuracy than $\mathrm{G_H}$BLUP_770K for DP on average. However, as for the RERW, $\mathrm{G_H}$BLUP_770K had a slight higher accuracy than $\mathrm{G_H}$BLUP+GBLUP_770K, and the $\mathrm{G_H}$BLUP+GBLUP_WGS had better performance than GBLUP_WGS_In_Block.

To evaluate whether the observed differences were statistically significant, we compared the correlation of the prediction accuracies using Hotelling's test. In the current study, we found no significant differences between the $\mathrm{G_H}$BLUP_770K and GBLUP_770K_In_Block for all scenarios (**Table 4**). However, $\mathrm{G_H}$BLUP_WGS using the WGS data had a significant improvement for RERW compared with GBLUP_WGS_In_Block (**Table 5**).

Accordingly, the slopes of the regression of the adjusted phenotype on GEBVs based on three approaches were presented in **Figure 2**. Our result showed that the regression coefficients of the HD chip were closer to 1 than those of the WGS data for most scenarios. The $\mathrm{G_H}$BLUP was near 1 for RERW when $r^2 \geq 0.5$. However, regression coefficients based on WGS data were almost stable for different LD levels.

TABLE 2 | Statistical description and heritability estimation of three traits in Chinese Simmental beef cattle.

Trait[1]	Number of phenotypes	Mean ± SD	Maximum	Minimum	h^2 ± SE
DP	1,221	0.535 ± 0.029	0.690	0.410	0.27 ± 0.07
MP	1,226	0.456 ± 0.031	0.616	0.325	0.17 ± 0.06
RERW	1,228	10.67 ± 2.20	18.32	5.03	0.23 ± 0.06

[1] *Trait: DP, dressing percentage; MP, meat percentage; RERW, rib eye roll weight.*

TABLE 3 | Total number of haplotype alleles, haploblocks, and the non_blocked SNPs from the 770K array and sequence data.

Data	r^2[1]	Haplotype alleles	Haploblocks (Blocked_SNPs)	Number of SNPs per haploblock	Non_blocked SNPs
770K	0.2	840,676	68,775 (362,710)	5.3	298,454
	0.3	599,270	66,027 (309,126)	4.7	352,038
	0.4	462,287	62,150 (268,055)	4.3	393,109
	0.5	371,074	58,009 (234,305)	4.0	426,859
	0.6	303,774	53,876 (204,725)	3.8	456,439
	0.7	249,195	49,320 (176,622)	3.6	484,542
	0.8	199,702	43,892 (147,008)	3.3	514,156
WGS	0.2	3,370,157	634,662 (3,536,404)	5.6	3,240,315
	0.3	2,701,350	601,761 (3,142,601)	5.2	3,634,118
	0.4	2,323,965	571,908 (2,877,612)	5.0	3,899,107
	0.5	2,067,572	545,069 (2,676,430)	4.9	4,100,289
	0.6	1,866,317	522,099 (2,502,764)	4.8	4,273,955
	0.7	1,684,138	500,294 (2,329,147)	4.7	4,447,572
	0.8	1,493,399	477,242 (2,123,952)	4.5	4,652,767

[1] *Seven different LD thresholds set from $r^2 \geq 0.2$ to $r^2 \geq 0.8$.*

Comparison of Accuracies of GP Based on Different Marker Densities

We found that genomic predictions using the HD chip were superior to the WGS data for all three traits (**Figure 1**); the accuracy of GBLUP_770K was 0.011, 0.01, and 0.015 higher than that of GBLUP_WGS for DP, MP, and RERW, respectively. However, no significant difference was found between the accuracies based on the two different densities according to Hotelling's test (**Table 6**). Moreover, significant differences between the two marker densities were observed for RERW using both GBLUP and G_HBLUP when SNPs within the blocks (divided by different LD thresholds) were selected. As for the G_HBLUP+GBLUP, no significant differences were found between HD chip and WGS. It should be noted that the accuracies of the three traits decreased obviously for GBLUP_770K_In_Block compared with the GBLUP_770K (**Figure 1**). However, no obvious change was found between GBLUP_WGS and GBLUP_WGS_In_Block.

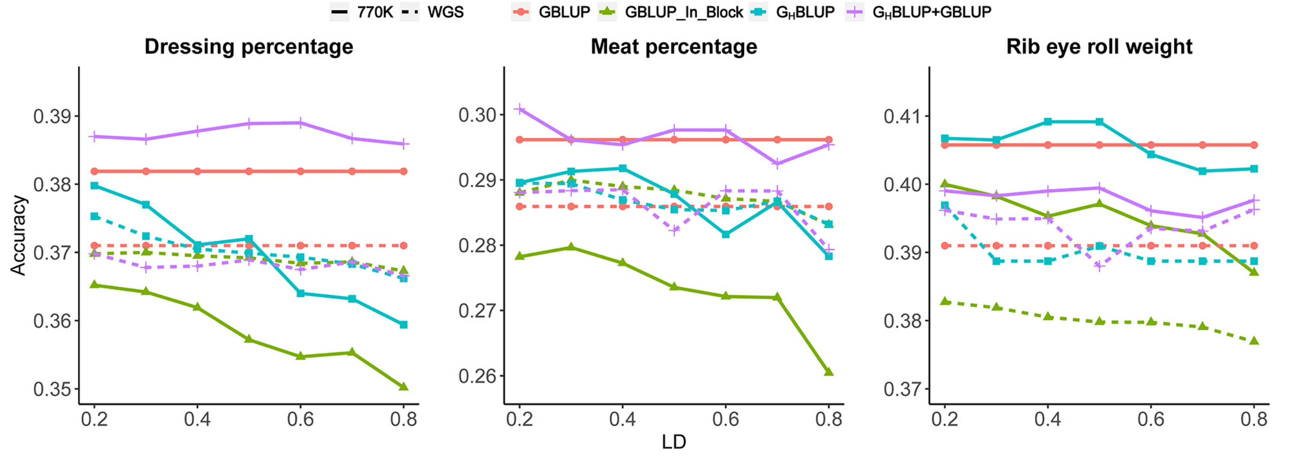

FIGURE 1 | Prediction accuracies of different r^2 thresholds for three traits based on the 770K data and WGS data.

TABLE 4 | *P*-values of Hotelling's *t*-test comparing the prediction accuracy obtained with the individual SNP and haploblock approaches using 770K data.

r^{2}[1]	G_HBLUP_770K			G_HBLUP+GBLUP_770K		
	DP	MP	RERW	DP	MP	RERW
0.2	0.102	0.315	0.462	0.097	0.211	0.931
0.3	0.147	0.295	0.352	0.089	0.342	0.994
0.4	0.299	0.198	0.113	0.052	0.292	0.719
0.5	0.101	0.213	0.161	0.019	0.167	0.811
0.6	0.293	0.400	0.221	0.013	0.150	0.864
0.7	0.367	0.188	0.274	0.026	0.240	0.837
0.8	0.288	0.102	0.065	0.015	0.059	0.391

[1] *Seven different LD thresholds set from $r^2 \geq 0.2$ to $r^2 \geq 0.8$.*

TABLE 5 | *P*-values of Hotelling's *t*-test comparing the prediction accuracy obtained with the individual SNP and haploblock approaches using WGS data.

r^{2}[1]	G_HBLUP_WGS			G_HBLUP+GBLUP_WGS		
	DP	MP	RERW	DP	MP	RERW
0.2	0.311	0.835	0.012	0.999	0.998	0.040
0.3	0.643	0.938	0.218	0.717	0.837	0.047
0.4	0.852	0.769	0.142	0.829	0.946	0.588
0.5	0.908	0.670	0.047	0.939	0.357	0.161
0.6	0.869	0.796	0.111	0.902	0.875	0.095
0.7	0.956	0.997	0.088	0.997	0.833	0.074
0.8	0.833	0.992	0.036	0.936	0.628	0.025

[1] *Seven different LD thresholds set from $r^2 \geq 0.2$ to $r^2 \geq 0.8$.*

Comparison of Accuracies of GP Among Different LD Levels

To investigate the influence of LD levels (r^2) on the prediction accuracy, we constructed haploblocks (from $r^2 \geq 0.2$ to $r^2 \geq 0.8$) using seven different levels. In our study, we found that the haploblock approach (including G_HBLUP and G_HBLUP+GBLUP) was better than the individual SNP approach (GBLUP_In_Block) at specific LD thresholds (r^2) (**Figure 1**). The accuracy of G_HBLUP_770K showed the highest accuracy for RERW when $r^2 \geq 0.5$, and the G_HBLUP_WGS outperformed GBLUP_WGS_In_Block. Under the strict LD threshold (r^2), the G_HBLUP+GBLUP_770K showed significant improvement compared with GBLUP_770K_In_Block for DP (**Table 4**).

Computation Time

In our study, the average computation time of GBLUP, G_HBLUP, and G_HBLUP+GBLUP were 4.42, 5.41, and 41.5 min using

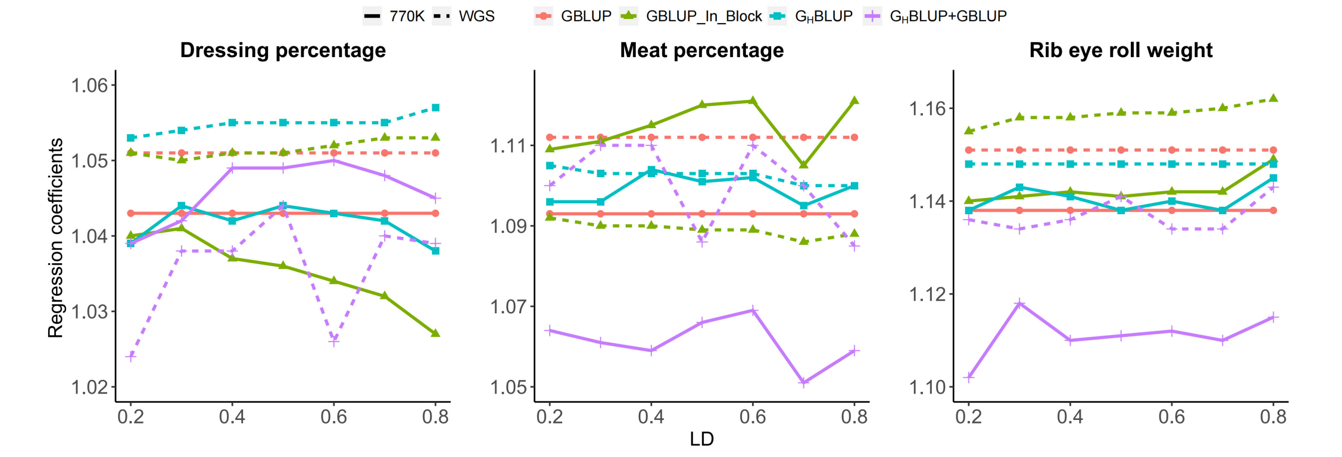

FIGURE 2 | Regression coefficients of pre-adjusted phenotypes on GEBVs for three traits in Chinese Simmental beef cattle.

TABLE 6 | *P*-values of Hotelling's *t*-test comparing the prediction accuracy obtained with the 770K data and the WGS data.

r^2 [1]	GBLUP			G_HBLUP			G_HBLUP+GBLUP		
	DP	**MP**	**RERW**	**DP**	**MP**	**RERW**	**DP**	**MP**	**RERW**
–	0.881	0.915	0.852	–	–	–	–	–	–
0.2	0.359	0.189	0.002	0.405	0.992	0.079	0.244	0.515	0.797
0.3	0.287	0.199	0.007	0.575	0.850	0.034	0.163	0.668	0.729
0.4	0.193	0.171	0.022	0.929	0.600	0.007	0.129	0.689	0.687
0.5	0.053	0.104	0.012	0.726	0.772	0.002	0.116	0.359	0.298
0.6	0.036	0.116	0.048	0.423	0.679	0.016	0.085	0.577	0.815
0.7	0.049	0.135	0.065	0.429	0.990	0.034	0.139	0.798	0.839
0.8	0.017	0.031	0.196	0.309	0.587	0.035	0.109	0.325	0.892

[1] *Seven different LD thresholds set from $r^2 \geq 0.2$ to $r^2 \geq 0.8$.*

fivefold cross-validation respectively (repeated 10 times). The time for constructing haplotype-based genomic relationship matrix ($\mathbf{G_H}$) were 8.22 and 203.97 h on average for the HD chip and WGS data, and 0.898 and 9.12 h for the SNP-based genomic relationship matrix (**G**).

DISCUSSION

Predictive Performance of Different Marker Density

In this study, we compared the accuracies of genomic prediction using both the HD chip and WGS data. A previous study suggested that prediction of breeding value was expected to be more accurate using the WGS data compared with the high-density chip because the causal mutations are assumed to be included in the WGS data (Druet et al., 2014). Genomic predictions based on sequence data can increase accuracy compared with predictions based on ∼30K SNP chips in simulation data (Meuwissen and Goddard, 2010; Clark et al., 2011; Druet et al., 2014; MacLeod et al., 2014). In contrast, for real data, a recent study found that no increases for prediction accuracy was observed using the imputed sequence data in Holstein Friesian cattle (Van Binsbergen et al., 2015). Our results presented the HD chip data had better performance than the WGS data using GBLUP approach. These findings can be explained by several factors including imputation accuracy, LD, MAF, genotyping errors, and population size (Iwata and Jannink, 2010; Zhang and Druet, 2010; Hayes et al., 2012; Ali et al., 2020). For instance, small reference population size and high imputation error rate from low-frequency SNPs may cause the decrease of accuracy for GP in WGS data (Heidaritabar et al., 2016). In addition, the strong LD between multiple true causal SNPs and potential QTLs segregating in long haplotypes in WGS data may make it difficult to pinpoint the truly causal SNP (Van Binsbergen et al., 2015).

Comparison of Methods of Genomic Prediction Based on SNP Chip

For the HD chip data, our results showed that G_HBLUP+GBLUP had the highest accuracy and G_HBLUP was better than GBLUP at different LD levels for DP and MP (**Figure 1**). In contrast, G_HBLUP showed the highest accuracy for RERW, which can be explained by the different genetic architectures of three traits. Moreover, DP and MP can be regarded as the compound traits, compared with RERW, which were determined by many genes with small effects. G_HBLUP+GBLUP for these two traits can include the non-blocked SNPs in the model, which should be more effective to increase the prediction accuracy. However, G_HBLUP+GBLUP approach may produce large prediction error variance and decrease the accuracy of GP for RERW due to the overestimation of the effects.

In addition, the G_HBLUP+GBLUP approach may reflect the real genetic architectures of these traits. For instance, a gene region contains many consecutive loci, which can be effectively modeled by the G_HBLUP approach. As for the gene regulatory region, the promoters or enhancers may be influenced by a single mutation, and this feature can be effectively integrated by the GBLUP approach.

Our findings were consistent with previous reports (Cuyabano et al., 2014; Teissier et al., 2020; Xu et al., 2020); they found that haplotype approach based on average LD threshold ($r^2 \geq 0.45$) can increase the prediction accuracies for milk production traits (up to 3.1%) compared with the individual SNP approach. Also, the accuracies of G_HBLUP+GBLUP and G_HBLUP can be influenced by the genetic architectures of different traits (Cuyabano et al., 2014).

The advantage of haplotype approach can be explained by the fact that SNPs are commonly bi-allelic, and SNP mutations in different loci tended to cause major changes in the haplotype frequencies (Curtis et al., 2001). Moreover, a QTL may be in complete LD with a multi-marker haplotype even if it is not in complete LD with any individual bi-allelic SNP marker (Cuyabano et al., 2014).

In our study, we found that the G_HBLUP+GBLUP_770K showed the highest accuracy for DP and MP. The LD level was set to $r^2 \geq 0.2$ in the haplotype prediction and several SNPs showing weak LD (r^2 from 0 to 0.2) with potential QTLs were not included in the model; therefore, adding the non-blocked SNPs may increase the prediction accuracy without loss of information. Also, we found that the regression coefficients using haplotype approach including G_HBLUP+GBLUP and

G_HBLUP is close to 1 for all three traits, compared with SNP approach (**Figure 2**), which were consistent with the prediction accuracy using the average Pearson correlation coefficient between the adjusted phenotypic values and genomic estimated breeding values.

Comparison of Methods of Genomic Prediction Based on WGS

As for the WGS data, our results suggested the haploblock approach based on LD can increase the accuracies of GP while reducing the number of variables. For RERW, G_HBLUP_WGS and G_HBLUP+GBLUP_WGS showed better performance than GBLUP_WGS_In_Block; however, no significant difference was found for DP and MP. The WGS data incorporating genotypes at causal variants into haplotypes allow effective estimation of haplotype effects. For DP and MP, we did not observe significant difference using both G_HBLUP_WGS and G_HBLUP+GBLUP_WGS. This result may be explained by the fact that the WGS data with high SNP density can produce the identified haplotype alleles (including some rare haplotype alleles); however, due to a large number of rare haplotype alleles with small effects or no effect were included in sequencing data (Gianola, 2013), the haplotype approach with them may not effectively improve prediction accuracy for DP and MP.

Predictive Performance of Different LD Levels

In this study, we found that the prediction accuracies using haplotype approach varied among different LD thresholds for three traits, especially for the HD chip data. One possible reason is that the size of haploblocks varies among different LD thresholds and the QTL effects can be accurately estimated at specific LD levels because the effective haploblocks were included. The HD chip data may cause the loss of effective haploblock effects for genomic prediction compared with WGS data. However, no obvious difference among different LD thresholds using G_HBLUP+GBLUP for three traits was observed. This result can be explained by the fact that G_HBLUP+GBLUP approach contains both the haploblock effects and SNP effects which were estimated from outside haploblocks.

Similar as previous studies (Cuyabano et al., 2014, 2015; Feitosa et al., 2020), our study also revealed that the optimum LD threshold should be considered in the haplotype approach. For DP, the optimum LD threshold was $r^2 \geq 0.2$ (**Figure 1**). For MP and RERW, the optimum LD threshold appeared at $r^2 \geq 0.5$. Cuyabano et al., reported that haploblocks built based on $D' \geq 0.45$ can produce an optimal set of variables for milk protein, fertility, and mastitis traits. Our results indicated that the optimal thresholds of different traits are different. Therefore, it is hard to determine the optimal haploblock length for all scenarios. For instance, Villumsen et al. (2009) evaluated the optimal haploblock length for the simulated traits with heritabilities ranging from 0.02 to 0.30, they found that haploblocks of 1 cM (0.8 Mb) can produce the highest accuracies across all traits in New Zealand dairy cattle. Previous studies found that the optimal haploblock length ranged from 0.4 to 0.8 Mb per haploblock (Villumsen and Janss, 2009; Villumsen et al., 2009). Hess et al. (2017) found the highest prediction accuracy using short haploblock (250 kb) in the admixed dairy cattle population. Our study suggested the setting of optimal LD threshold depends on the LD between SNPs and QTLs and the population structure. Thus, the optimal LD threshold was required to be evaluated for each dataset independently.

It should be noted that haplotype approach based on LD had less improvement on the prediction accuracies compared with the fixed block length approach, which was in agreement with a previous study (Cuyabano et al., 2014). Xu et al. (2020) constructed the haploblock using the constant number of SNPs, and their findings suggested that the extension from the SNP-based model to haplotype-based model can improve the accuracy by 5.4–9.8%. Moreover, Hess et al. (2017) reported that fitting covariates for fixed-length haplotype alleles can increase the accuracy of GP up to 5.5% compared with SNPs.

In our study, we found that LD-based haplotype approach cannot increase the accuracy to 5%, which was consistent with a previous report (Cuyabano et al., 2014). They performed genomic prediction for three important traits (milk protein, fertility, and mastitis) using LD-based haplotypes in the Nordic Holstein population, and their finding suggested Bayesian model can produce the highest accuracy for the milk protein trait. This difference can be explained by the fact that all SNP information was included in the fixed block length approach, while only a small set of SNPs was included in the LD-based haplotype approach. In our study, we found that computation times were much longer for G_HBLUP+GBLUP than GBLUP and G_HBLUP, while no obvious difference was found between the GBLUP and G_HBLUP approach. Two genomic relationship matrixes ($\boldsymbol{G_H}$and **G**) were estimated in the G_HBLUP+GBLUP model, thus the long time was required for this approach. In addition, our results suggested that the haplotype approach for WGS data requires more time to construct the genomic relationship matrixes than the SNP-based approach. It should be noted that the haplotype-based genomic relationship matrix need to be recoded using numerical dosage coding strategies for each haploblock (Calus et al., 2008).

CONCLUSION

Our study suggested that haploblock approach using both HD chip and WGS data can improve the prediction accuracy compared with the individual SNP approach. The prediction accuracies of haploblock approach varied in different LD thresholds. Therefore, it is important to determine the optimal r^2 threshold when constructing haploblocks for genomic prediction. The advent of whole-genome sequencing has made it possible to contemplate linking the diverse phenotypes to genetic variations at the genome level. Furthermore, haplotype strategy integrating biological information could be used to identify sequence variants which are likely to harbor mutations affecting complex traits.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Dryad: doi: 10.5061/dryad.4qc06.

ETHICS STATEMENT

The animal study was reviewed and approved by the Science Research Department of the Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (CAAS) (Beijing, China). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

LYX designed the experiments. HL analyzed the data and wrote the article. LiX, ZW, LeX, PZ, and HG collected the data. PG, YC, XG, LZ, HJG, WC, LYX, BZ, and JL discussed and improved the article. All authors read and approved the final article.

FUNDING

This work was supported by the National Key Research and Development Program of China (2018YFD0501802), the National Natural Science Foundation of China (31802049, 31372294, and 31201782), Chinese Academy of Agricultural Sciences of Technology Innovation Project (CAAS-XTCX2016010, CAASZDXT2018006, and ASTIP-IAS03), Cattle Breeding Innovative Research Team of Chinese Academy of Agricultural Sciences (cxgc-ias-03, Y2016PT17, and 2019-YWF-YTS-11), Beijing Natural Science Foundation (6154032), China Agriculture Research System of MOF and MARA, and the Science and Technology Project of Inner Mongolia Autonomous Region (2020GG0210).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665382/full#supplementary-material

REFERENCES

Ali, M., Zhang, Y., Rasheed, A., Wang, J., and Zhang, L. (2020). Genomic prediction for grain yield and yield-related traits in Chinese winter wheat. *Int. J. Mol. Sci.* 21:1342. doi: 10.3390/ijms21041342

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7, 781–791. doi: 10.1038/nrg1916

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., et al. (2012). Genomic selection in French dairy cattle. *Anim. Prod. Sci.* 52, 115–120.

Bolormaa, S., Pryce, J. E., Kemper, K., Savin, K., Hayes, B. J., Barendse, W., et al. (2013). Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle1. *J. Anim. Sci.* 91, 3088–3104. doi: 10.2527/jas.2012-5827

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

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

Calus, M. P. 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.

Calus, M. P. L., Meuwissen, T. H. E., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L. J., 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.

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

Clark, S. A., Hickey, J. M., and Van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol.* 43:18.

Curtis, D., North, B. V., and Sham, P. C. (2001). Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Ann. Hum. Genet.* 65, 95–107.

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, 61.

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

De Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503.

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.

Feitosa, F. L. B., Pereira, A. S. C., Amorim, S. T., Peripolli, E., Silva, R. M. O., Braz, C. U., et al. (2020). Comparison between haplotype-based and individual snp-based genomic predictions for beef fatty acid profile in Nelore cattle. *J. Anim. Breed. Genet.* 137, 468–476. doi: 10.1111/jbg.12463

Garnier, S., Truong, V., Brocheton, J., Zeller, T., Rovital, M., Wild, P. S., et al. (2013). Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS Genet.* 9:e1003240. doi: 10.1371/journal.pgen.1003240

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

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

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., and van der Werf, J. H. (2012). Accuracy of genotype imputation in sheep breeds. *Anim. Genet.* 43, 72–80. doi: 10.1111/j.1365-2052.2011.02208.x

Hayes, B. J., Chamberlain, A. J., Mcpartlan, H., Macleod, I. M., and Goddard, M. E. (2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet. Res.* 89, 215–220.

Heidaritabar, M., Calus, M. P., Megens, H. J., Vereijken, A., Groenen, M. A., and Bastiaansen, J. W. (2016). Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *J. Anim. Breed. Genet.* 133, 167–179. doi: 10.1111/jbg.12199

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.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Stat.* 11, 271–283.

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

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

Iwata, H., and Jannink, J. L. (2010). Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. *Crop Sci.* 50, 1269–1278.

Li, B., Zhang, N., Wang, Y. G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet*. 9:237. doi: 10.3389/fgene.2018.00237

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834. doi: 10.1002/gepi.20533

MacLeod, I. M., Hayes, B. J., and Goddard, M. E. (2014). The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics* 198, 1671–1684.

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). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185, 623–631.

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.

Meuwissen, T. H., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* 46:49. doi: 10.1186/1297-9686-46-49

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., and Georges, M. (2012). Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet.* 8:e1002854. doi: 10.1371/journal.pgen.1002854

Sun, X. (2016). "Haplotype-based genomic prediction of breeds not in training," in *Proceedings of the PLANT & Animal Genome Conference XXIV, January 08-13, 2016*, San Diego, CA.

Teissier, M., Larroque, H., Brito, L. F., Rupp, R., Schenkel, F. S., and Robert-Granié, C. (2020). Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. *J. Dairy Sci.* 103, 11559–11573. doi: 10.3168/jds.2020-18662

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

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

Villumsen, T. M., and Janss, L. (2009). Bayesian genomic selection: the effect of haplotype length and priors. *BMC Proc.* 3(Suppl 1):S11. doi: 10.1186/1753-6561-3-s1-s11

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

Vormfelde, S. V., and Brockmöller, J. (2007). On the value of haplotype-based genotype-phenotype analysis and on data transformation in pharmacogenetics and -genomics. *Nat. Rev. Genet.* 8, 1916. doi: 10.1038/nrg1916-c1

Weng, Z. Q., Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet. Sel. Evol.* 46:34.

Won, S., Park, J. E., Son, J. H., Lee, S. H., Park, B. H., Park, M., et al. (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

Xu, L., Gao, N., Wang, Z., Xu, L., Liu, Y., Chen, Y., et al. (2020). Incorporating genome annotation into genomic prediction for carcass traits in Chinese Simmental beef cattle. *Front. Genet.* 11:481. doi: 10.3389/fgene.2020.00481

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

Zhang, Z., and Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.* 93, 5487–5494. doi: 10.3168/jds.2010-3501

Zhu, B., Niu, H., Zhang, W., Wang, Z., Liang, Y., Guan, L., et al. (2017). Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. *BMC Genomics* 18:464. doi: 10.1186/s12864-017-3847-7

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

Zondervan, K. T., and Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* 5, 89–100. doi: 10.1038/nrg1270

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

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