Journal of Integrative Agriculture 2017, 16(4): 911–920

Available online at www.sciencedirect.com

ScienceDirect

RESEARCH ARTICLE

Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle

ZHU Bo*, ZHANG Jing-jing*, NIU Hong, GUAN Long, GUO Peng, XU Ling-yang, CHEN Yan, ZHANG Lu-pei, GAO Hui-jiang, GAO Xue, LI Jun-ya

Laboratory of Molecular Biology and Bovine Breeding, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, P.R.China

Abstract

Genomic selection has been demonstrated as a powerful technology to revolutionize animal breeding. However, marker density and minor allele frequency can affect the predictive ability of genomic estimated breeding values (GEBVs). To investigate the impact of marker density and minor allele frequency on predictive ability, we estimated GEBVs by constructing the different subsets of single nucleotide polymorphisms (SNPs) based on varying markers densities and minor allele frequency (MAF) for average daily gain (ADG), live weight (LW) and carcass weight (CW) in 1 059 Chinese Simmental beef cattle. Two strategies were proposed for SNP selection to construct different marker densities: 1) select evenly-spaced SNPs (Strategy 1), and 2) select SNPs with large effects estimated from BayesB (Strategy 2). Furthermore, predictive ability was assessed in terms of the correlation between predicted genomic values and corrected phenotypes from 10-fold cross-validation. Predictive ability for ADG, LW and CW using autosomal SNPs were 0.13±0.002, 0.21±0.003 and 0.25±0.003, respectively. In our study, the predictive ability increased dramatically as more SNPs were included in analysis until 200K for Strategy 1. Under Strategy 2, we found the predictive ability slightly increased when marker densities increased from 5K to 20K, which indicated the predictive ability of 20K (3% of 770K) SNPs with large effects was equal to the predictive ability of using all SNPs. For different MAF bins, we obtained the highest predictive ability for three traits with MAF bin 0.01–0.1. Our result suggested that designing a low-density chip by selecting low frequency markers with large SNP effects sizes should be helpful for commercial application in Chinese Simmental cattle.

Keywords: genomic prediction, cross-validation, Chinese Simmental beef cattle, marker density, minor allele frequency (MAF)

1. Introduction

Recent advances of high-throughput genotyping technology have promoted the application of genomic selection in farm animals (Meuwissen *et al.* 2001). Genomic selection (GS) which is a form of marker-assisted selection refers to the use of whole genome single nucleotide polymorphisms (SNPs) for breeding value estimation and subsequent selection of individuals based on genomic estimated breeding values

Received 18 April, 2016 Accepted 26 August, 2016
ZHU Bo, E-mail: zhubo525@126.com; ZHANG Jing-jing, E-mail: zhang_jingjing89@163.com; Correspondence LI Jun-ya, E-mail: jl1@iascaas.net.cn
* These authors contributed equally to this study.

doi: 10.1016/S2095-3119(16)61474-0

(Goddard and Hayes 2007, 2009). Genetic improvement of growth traits has been considered as a great indicator in the beef industry. With the application of GS, genomic estimated breeding values (GEBVs) can be obtained for selecting candidates at a young age before their phenotypic information are available, which could increase the accuracy of selection and shorten generation intervals (Lund *et al.* 2011; Saatchi *et al.* 2011; Wolc *et al.* 2011).

GEBVs are calculated as the sum of the effects of dense genetic markers across genome. Various methods have been proposed to estimate GEBVs, such as genomic best linear unbiased prediction approaches (GBLUP) with genomic relationship matrices (VanRaden 2008), random regression BLUP (RRBLUP) (Endelman 2011), Bayesian linear regression methods (Meuwissen *et al.* 2001; Gianola *et al.* 2009) and fully non-parametric approaches (Gianola and van Kaam 2008; de Los Campos *et al.* 2009; Long *et al.* 2010; Ober *et al.* 2011). GBLUP is a simple method with low computational requirements, and this method has been widely used for genomic prediction in cattle (Habier *et al.* 2010). Furthermore, the ability of genomic prediction depends on many effects, such as the heritability of the trait, the number of animals with phenotypes and genotypes (VanRaden *et al.* 2009), distribution of allele frequencies, linkage disequilibrium (LD), marker density and population genetic architecture (Calus and Veerkamp 2007; Habier *et al.* 2007, 2010; Makowsky *et al.* 2011; de Los Campos *et al.* 2013).

Marker density is an important factor that affects the ability of genomic prediction. However, high-density (HD) panel genotyped for a large population is costly and requires a large amount of computational time. Previous studies proposed using low-density SNP panels to predict GEBVs in animal breeding, and their results suggested selecting SNPs with large effects can produce almost as effective evaluation for GEBVs as HD panels (Weigel *et al.* 2010; Boichard *et al.* 2012; Hayes *et al.* 2012; Mullen *et al.* 2013). Moreover, a simulation study revealed that 95% of accuracy could be achieved using a small proportion of markers extracted from HD panel (Zhang *et al.* 2011).

Minor allele frequency (MAF) as another factor can affect the predictive ability and various studies have reported the impact of MAF on predictive ability. For example, Yang *et al.* (2010) found that most SNPs in genome may only explain a small fraction of the genetic variance of complex traits, and the MAF of SNPs in a panel should be different from that of causal variants with a lower MAF (Yang *et al.* 2010). Abdollahi-Arpanahi *et al.* (2014) revealed the low-frequency SNPs with large effect sizes may deliver better predictive ability for quantitative traits including body weight, ultra-sound measurement of breast muscle and hen house egg production in chickens. However, several studies suggested low alleles frequency markers with potential larger effects may not be feasible for genomic prediction due to the estimated effect may be low precision in a finite size training population (Lettre 2011; Park *et al.* 2011).

Chinese Simmental cattle is an important imported breed in China, and this breed is particularly renowned for the rapid growth rate and become a popular cattle breed since they were imported to China from 1970's onward. To our knowledge, previous studies have been widely discussed the marker density can influence the genomic predictive accuracy on cattle including Holstein, Brown Swiss and Japanese Black beef cattle (Zhang *et al.* 2011; Erbe *et al.* 2013; Ogawa *et al.* 2014), however, the impacts of MAF and marker density has not been explored for growth traits in Chinese Simmental beef cattle. The objectives of this study were to: (1) evaluate the impact of marker densities on predictive ability for growth traits in 1 059 Chinese Simmental beef cattle; (2) investigate the impact of MAF on predictive ability using different MAF bins; (3) assess the predictive ability using low-density SNP panels selected from BovineHD Beadchip with different selection strategies (select a subset of evenly-spaced SNPs or select SNPs with large effects) with 10-fold cross-validation.

2. Materials and methods

2.1. Ethics statement

All animals were treated following the guidelines for the experimental animals established by the Council of China. Animal experiments were approved by the Science Research Department of the Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (CAAS) (Beijing, China).

2.2. Phenotype data

Experimental animals consisted of 1 087 Chinese Simmental beef cattle, born between 2008 and 2012, originated from Ulgai, Xilingol League, Inner Mongolia, China. After weaning, the cattle were introduced to Beijing Jinweifuren Cattle Farm, China, and were fattened under the same feeding and management. More detailed descriptions of breeding and management of the cattle have been described previously (Zhu *et al.* 2016). Average daily gain (ADG) was the rate of weight gain per day over the specified period of time in farm. Live weight (LW) was measured before slaughter with fasting 24 h. Carcass weight (CW) was measured after slaughter and removal of most internal organs. In this study, ADG, body weight and CW were utilized to predict GEBVs by the information of genotypic variation. Systematic environment factors including farm, year of measurement and age at slaughter (seasons) effects were adjusted in the mixed linear model. Genetic parameters were calculated using residual

maximum likelihood method in animal model using **G** matrix.

2.3. Genotype data and population structure

The DNA for each animal was obtained from blood using the routine procedures. 1 087 samples were genotyped with the Illumina BovineHD Genotyping BeadChip (Illumina Inc., San Diego, CA). The BovineHD Beadchip contains 777 962 SNPs with an average probe spacing of 3.43 kb and a median spacing of 2.68 kb. These datasets are available from the Dryad Digital Repository (doi: 10.5061/dryad.4qc06). Before statistical analysis, SNPs were excluded as following: call rates<0.95, MAF<0.01, Hardy-Weinberg equilibrium test $P<10^{-6}$. Individual was also removed if genotype missing rate>0.1. The finial data consisted of 1 059 cattle and 667 954 SNPs on autosomes. The mean MAF in our study was 0.24. Quality control was performed using PLINK v1.07 Software (Purcell *et al.* 2007). Characteristics of the SNPs quality control were displayed in Table 1. To explore the genetic structure of Chinese Simmental beef cattle populations, principal component analysis (PCA) on genotypic data was used to correct for population stratification in our study.

2.4. Statistic models

Each trait was analyzed independently using GBLUP method as follows:

$$y=\mathbf{X}b+\mathbf{Z}g+e$$

Where y is a vector of original phenotypes for all cattle, **X** is the incidence matrix of b, b is the vector of fixed effects, **Z** is the incidence matrix for the random marker effects, g is a vector of breeding values of genotyped individuals, e is a vector of residuals. $g\sim N(0, \mathbf{G}\sigma_g^2)$, where σ_g^2 is the additive genetic variance, and **G** is the marker-based genomic relationship matrix (VanRaden 2008). Random residuals were assumed such that $e\sim N(0, I\sigma_e^2)$, where σ_e^2 is residual variance.

Using the SNP genotype data, the **G** matrix was calculated as:

$$\mathbf{G}=\frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2\sum_{i=1}^{m}p_i(1-p_i)}$$

Where, **M** is the matrix with values 0, 1 and 2 that specifies which marker alleles each individual inherited. Dimensions of **M** are the number of individuals (n) by the number of loci (m). p_i is the frequency of the second allele at locus i. Let **P** contain allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that the ith column of **P** is $(2(p_i-0.5), ..., 2(p_i-0.5))'$.

Genetic parameters (heritability) was obtained using **G** matrix to replace numerator relationship matrix (**A** matrix) in animal model by MTDFREML Software (Boldman *et al.* 1995). The simplex algorithm was stopped when the variance of the function values (i.e., –2logL with L=likelihood given y) in the simplex was less than 1×10^{-6}. Standard errors were obtained directly from MTDFREML program for the two-trait analyses for ADG, LW and CW, because they had the same number of observations using the average information matrix.

Table 1 Summary statistics of SNPs quality control

Total number of SNPs[1)]	777 962
SNPs unknown position	2 078
SNPs on chromosome Y	1 224
SNPs on chromosome X	39 367
SNPs with MAF<0.01	40 125
SNPs not in Hardy-Weinberg equilibrium ($P<10^{-6}$)	6 071
SNPs with missing rates>0.05	21 143
After QC	667 954
Individuals with genotype missing rate>0.1	28

[1)] SNPs, single nucleotide polymorphisms; MAF, minor allele frequency; QC, quality control.

2.5. Linkage disequilibrium

D´ and r^2 (Hill 1974) were widely used in practice to measure the extent of linkage disequilibrium. We used r^2 to estimate LD, as r^2 was more robust and no sensitive to the changing of gene frequency and effective population size compared to D´ (Niu *et al.* 2016).

Assuming two loci A and B, each locus had two alleles (denoted A_1, A_2 and B_1, B_2, respectively). P_{A1}, P_{A2}, P_{B1} and P_{B2} were the frequency of each of the alleles. P_{11}, P_{12}, P_{21} and P_{22} showed the frequency of haplotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Thus, r^2 could be expressed as:

$$r^2=\frac{(P_{11}P_{22}-P_{12}P_{21})^2}{P_{A1}P_{A2}P_{B1}P_{B2}}$$

2.6. Maker densities and MAF bins

To investigate the impacts of different markers densities on predictive abilities, we used two strategies to construct SNP subsets. Under Strategy 1, we selected evenly-spaced SNPs for subsets of 0.5K, 1K, 5K, 10K, 20K, 40K, 100K, 200K, 300K and 400K. For each group, the extent of LD was estimated for all pairs between two adjacent SNPs on autosomes. Under Strategy 2, we selected SNPs based on large SNP effects estimated from BayesB method, the ranked SNPs for each individual trait were separated to subsets with different sizes including 0.5K, 1K, 5K, 10K, 20K, 40K, 100K, 200K, 300K and 400K. In addition, to study the relationship between allele frequencies and predictive abilities, we divided SNPs into 5 groups with different MAF bins (0.01–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5), and

the number of SNPs in each group was nearly equal. A total of 24 sets of SNPs at autosomes were finally used to construct **G** matrix respectively.

2.7. Cross-validation procedure

In order to assess the ability of genomic prediction for each SNP subset, we utilized 10-fold cross-validation method. 1 059 samples were evenly divided into 10 subsets. One of the 10 subsets was used as the validation set where the phenotypic values were assumed to be unknown, and the remaining 9 subsets were used as the training set. GEBVs in the validation set were estimated using GBLUP method. The procedure was repeated 20 times, and each time subsets were randomly sampled. Furthermore, we used the other two types (3- and 5-fold) of cross-validation to analysis the impacts of varying the training set size on the predictive ability. The predictive ability were assessed by calculating the correlation between GEBVs and corrected phenotypic values (y_c), where y_c was defined as the original phenotypic values corrected for fixed effects.

3. Results

3.1. Summary statistics and population structure

The number of individuals, the mean and standard deviation, coefficient of variation, and the minimum and maximum of values of ADG, LW and CW traits were shown in Table 2. The heritability of the trait was 0.37, 0.38 and 0.42 for ADG, LW and CW, respectively, and all these three traits were moderate heritabilities. The genetic correlations among these traits were high, and the highest genetic correlation (0.94) existed between LW and CW (Table 3). To explore the genetic population structure, we performed PCA using the genotyping dataset in Chinese Simmental beef cattle population (Fig. 1). We found our population can be clustered into three groups, and these cluster information were used to correct population stratification.

Table 2 Summary statistics of the three growth traits in Chinese Simmental beef cattle

Trait[1)]	N[2)]	Mean±SD (kg)[3)]	CV (%)[4)]	Minimum	Maximum
ADG	1 059	0.70±0.2	28.69	0.42	1.73
LW	1 059	508.54±69.86	13.74	318	776
CW	1 059	273.62±44.62	16.31	163	436

[1)] ADG, average daily gain; LW, live weight; CW, carcass weight.
[2)] N, number of individuals.
[3)] SD, standard deviation.
[4)] CV, coefficient of variation.

3.2. Predictive ability with cross-validation

In this study, we performed genomic prediction for ADG, LW and CW in 1 059 Chinese Simmental beef cattle. For the 10-fold cross-validation, predictive abilities were 0.13±0.002 for ADG, 0.21±0.003 for LW and 0.25±0.003 for CW using all autosomal SNPs. We observed the predictive ability decreased with the training set animals decreasing (Fig. 2). Specifically, predictive ability using 10-fold cross-validation was higher than that of 3- and 5-fold cross-validation for three growth traits.

3.3. Impact of marker densities on predictive ability

Under Strategy 1, the distribution of LD (r^2) for marker densities from 0.5K to 400K evenly-spaced SNPs and all autosomal SNPs were shown in Fig. 3. The average of LD (r^2) was 0.30 for all autosomal SNPs. We found that the average r^2 values dropped dramatically with the decrease of marker densities until 200K. We observed that the predictive ability increased obviously as more SNPs were included in analysis until 200K (Fig. 4). The increasing tendency of predictive ability on LW was more robust than that for ADG and CW. Furthermore, under Strategy 2, we found the predictive ability slightly increased when marker density increased from 0.5K to 20K, 20K and 10K for LW, CW and ADG, respectively, and we found no additional improvement of the predictive ability as increasing the marker densities (Fig. 5).

3.4. Impact of MAF on predictive ability

To investigate the impact of MAF on predictive ability, we firstly characterized the frequency distribution of SNPs using

Table 3 Estimates (and standard errors) of heritability, genetic and environmental correlations for the three growth traits in Chinese Simmental beef cattle[1)]

Trait 1[2)]	Trait 2[3)]	σ^2_{p1}	σ^2_{p2}	h^2_1	h^2_2	r_g	r_e
ADG	LW	0.041	4 880.42	0.37±0.03	0.43±0.04	0.89±0.02	0.86±0.04
	CW	0.042	1 990.94	0.37±0.02	0.38±0.05	0.82±0.03	0.78±0.04
LW	CW	4 880.53	1 990.82	0.42±0.04	0.37±0.06	0.94±0.05	0.95±0.06

[1)] σ^2_p, the phenotypic variance; h^2, heritability; r_g, genetic correlation; r_e, environmental correlation.
[2)] ADG, average daily gain; LW, live weight.
[3)] CW, carcass weight.

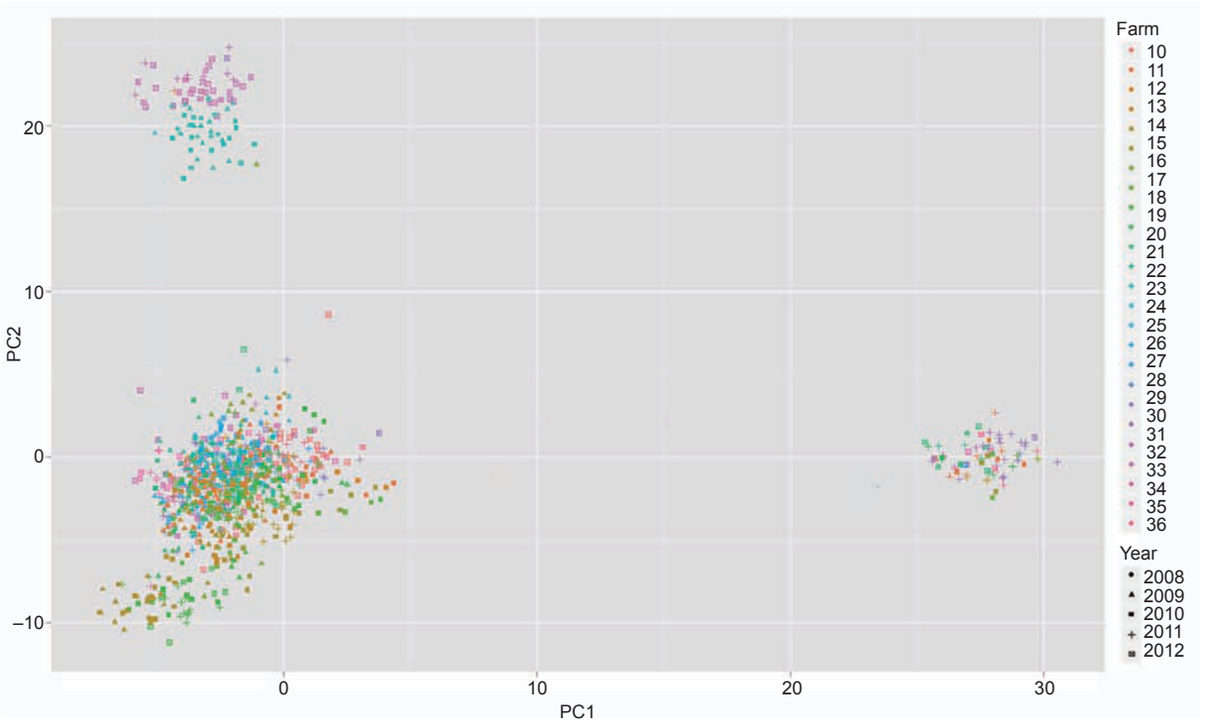

Fig. 1 Population genetic structure for Chinese Simmental beef cattle. PC, principal component.

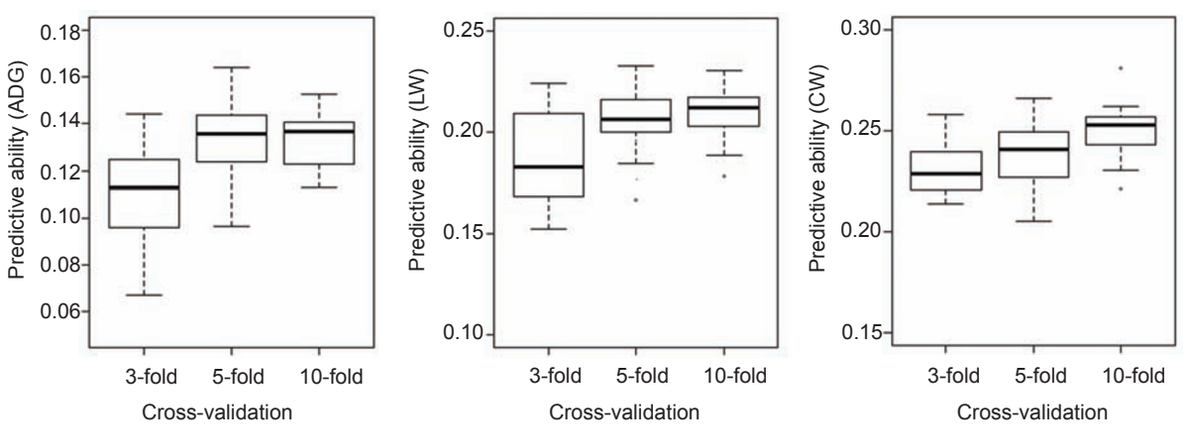

Fig. 2 Genomic predictive ability of genomic best linear unbiased prediction approaches (GBLUP) for three types of cross-validation. Each box plot illustrates the average predictive abilities for 20 replicates of cross-validation procedures. The number of individuals in training set for each cross-validation type was: 3-fold, *n*=706; 5-fold, *n*=847; 10-fold, *n*=953. ADG, average daily gain (kg); LW, live weight (kg); CW, carcass weight (kg). Error bar means standard deviation.

five MAF bins on 29 autosomes (Fig. 6). For different MAF bins, each chromosome showed similar trend that these five MAF bins shared the similar frequencies except for chromosomes 8, 13 and 14. Generally, MAF bin with 0.01–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5 contained 128 851, 118 919, 132 504, 140 569, and 147 111 SNPs, respectively. We found the predictive ability for ADG, LW and CW declined until the MAF bin was up to 0.2–0.3 (Fig. 7). Notably, the predictive ability using low MAF bin (0.01–0.1) was higher than these of high MAF bins for the three traits.

4. Discussion

To investigate the effects of marker densities and minor allele frequency on genomic predictive ability in Chinese Simmental cattle, we estimated GEBVs by constructing the different subsets of SNPs based on varying markers densities and minor allele frequency for ADG, LW and CW. Marker den-

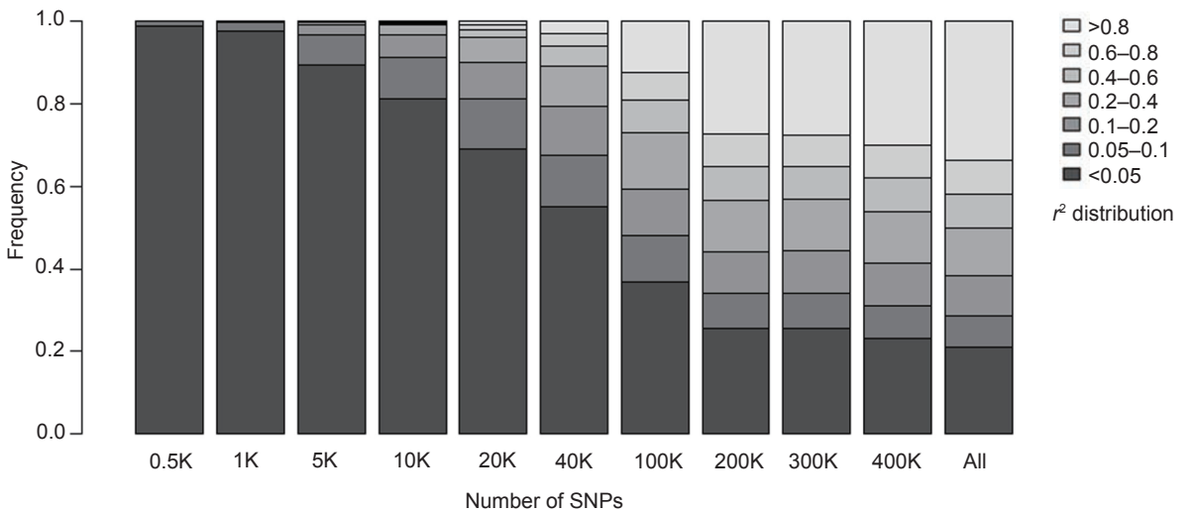

Fig. 3 The distribution of r^2 between adjacent single nucleotide polymorphisms (SNPs) under different marker densities. All, the whole autosomal SNPs.

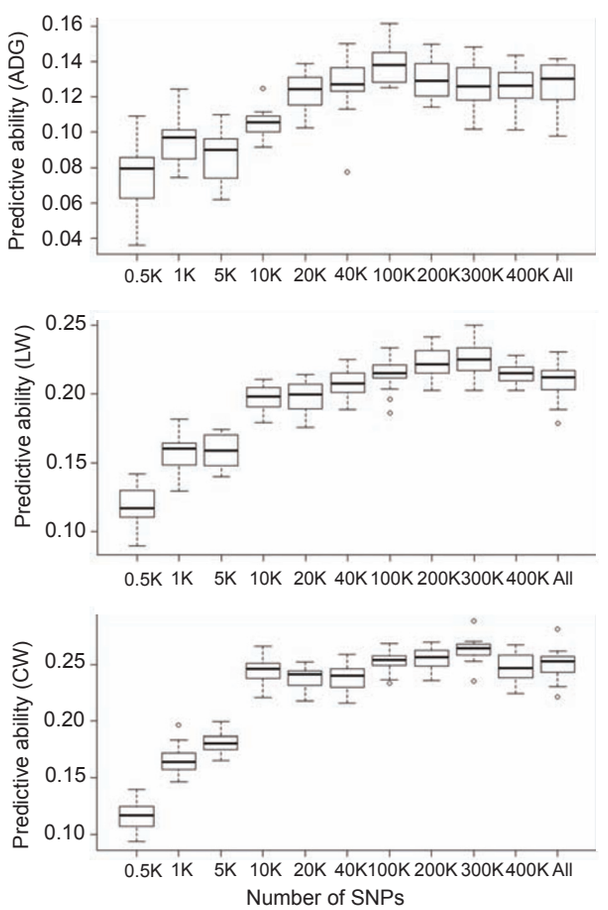

Fig. 4 Predictive abilities of 10-fold cross-validation with genomic best linear unbiased prediction approaches (GBLUP) for average daily gain (ADG, kg), live weight (LW, kg) and carcass weight (CW, kg) using different marker densities that were selected based on evenly-spaced strategy (Strategy 1). All, the whole autosomal SNPs. Error bar means standard deviation.

sity subsets Error bar means SD. were constructed using two strategies (Strategy 1, select evenly-spaced SNPs and Strategy 2, select large SNP effects estimated from BayesB method). The predictive abilities were evaluated through cross-validation. Our studies provide some valuable insights for applying genomic selection with low-density markers in Chinese Simmental cattle.

In the recent years, GBLUP has been widely used for genomic evaluation in dairy cattle (Ding *et al.* 2013; Gao *et al.* 2015; Mikshowsky *et al.* 2016). Essentially, predictive ability of GEBVs estimated by GBLUP can be affected by the characteristics of **G** matrix, which was significantly influenced by the number of markers, minor allele frequency (MAF) and marker effects size (Daetwyler *et al.* 2010; Ober *et al.* 2012; Edriss *et al.* 2013; Gianola 2013). Given different marker densities for the two selection strategies, we found the predictive ability increased with marker density for genomic prediction of ADG, LW and CW traits in our study. Notably, we found the increasing trends for two strategies were obviously different (Figs. 4 and 5), for example, the predictive ability under Strategy 1 dramatically increased with marker density, compared to Strategy 2. Previous studies have evaluated the effect of marker density and MAF on the predictive ability of GEBVs, i.e., Edriss *et al.* (2013) conducted genomic prediction in the Nordic Holstein and Jersey populations using Bovine50K Beadchip, and they investigated different marker-editing criteria on predictive accuracy and found the selected SNP subset with restriction of MAF and Hardy-Weinberg proportions (HWP) can achieve the highest predictive accuracy. Abdollahi-Arpanahi *et al.* (2014) assessed predictive ability based on different strategies for SNPs selection using 600K Affymetrix array in 1 352 chickens, their results provided strong evidences that designing low-density chip using low

MAF with large SNP effects was useful for commercial usage, and our findings suggested that it is feasible to design low-density SNP chip with large effect for genomic prediction in cattle, which were consistent with their results.

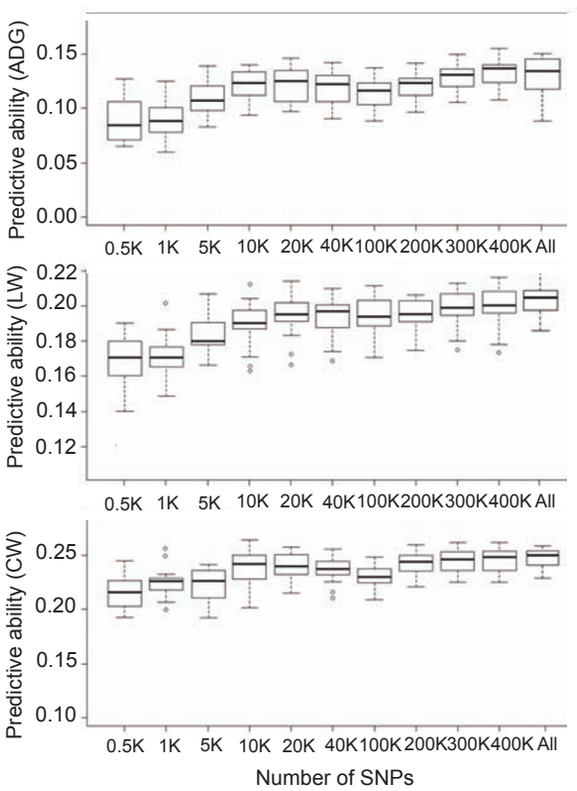

Fig. 5 Predictive abilities of 10-fold cross-validation for for average daily gain (ADG, kg), live weight (LW, kg) and carcass weight (CW, kg) using different marker densities that were selected based on large single nucleotide polymorphism (SNP) effects estimated from BayesB (Strategy 2). All, the whole autosomal SNPs. Error bar means standard deviation.

For Strategy 1, we found the predictive ability increased dramatically with marker density until the marker density increased up to 200K, and there was nearly no additional improvement with the marker density increase. Predictive ability was mainly affected by SNPs in the extend of LD with causal polymorphisms (Habier *et al.* 2007). With the decrease of marker density, the average LD for neighbor SNPs across the whole genome decreased significantly (Zhang *et al.* 2011) which can be discovered in our study (Fig. 3), therefore, reducing SNP density (Strategy 1) lead to a significant decay of predictive ability under Strategy 1 in our study (Fig. 4). Specially, for three growth traits, the predictive ability displayed dramatically increase with marker density until 200K. It is assumed that each marker has equal effect and all markers share the same variance in GBLUP model (VanRaden 2008), however, for complex trait, the genetic variance of traits should be mainly explained by a few regions in genome with several quantitative trait locus (QTL) with large effects. Overall, the genetic architecture of three traits may be different and delivered different predictive abilities in our analysis. In addition, our study suggested selecting markers with large effects the predictive ability was almost constant for all markers. For Strategy 2, the predictive ability slightly increased with the number of makers until 20K, which was consistent with quasi-infinitesimal genetic architecture (Ober *et al.* 2012). Our results indicated that LW, CW and ADG were quantitative traits with a highly polygenic genetic architecture rather than being driven by a few major causal QTL. Previous studies have evaluated the influence of the two strategies (based on the absolute values of SNP effects or using evenly-spaced SNP across the genome) for selecting subsets of SNP on genomic predictive ability in dairy cattle (Weigel *et al.* 2009; Vazquez *et al.* 2010), and their results suggested designing

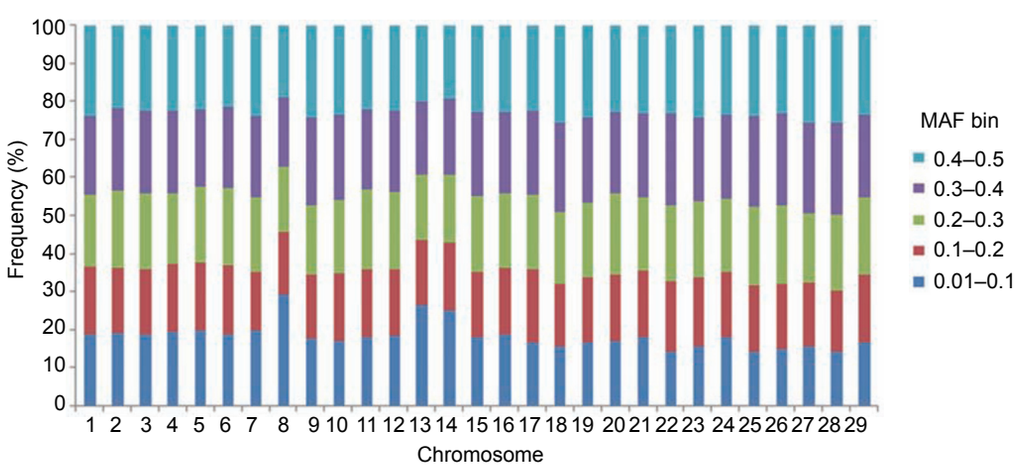

Fig. 6 Minor allele frequency distributions for autosomal single nucleotide polymorphisms (SNPs) across genome. MAF, minor allele frequency.

a low-density SNP panel with large effects outperformed the strategy of using evenly-spaced SNP. Thus, to design a low-density panel, our study indicated that the advantage of using Strategy 2 compared with Strategy 1 was clear (Figs. 4 and 5). Moreover, under Strategy 2, these selected markers are trait specific. To design a low-density panel, three target growth traits (ADG, LW and CW) needs to be selected simultaneously with the markers selected for all traits. Therefore, proper strategy exists between the number of traits in a breeding program and the number of trait-specific markers for each trait included in this low-density panel.

For different MAF bins, we found the predictive ability with low MAF bin was better than high MAF bin for three growth traits, especially for LW. The highest predictive abilities were obtained with MAF bin 0.01–0.1 for three traits (Fig. 7), and this is likely to be explained that the frequency of minor alleles can influence the property of **G** matrix. Normally, we tended to delete the SNP makers with the extremely low MAF. However, these rare mutations could largely contributed to the missing heritability, and causal variants had a lower minor allele frequency than common SNPs (Yang *et al.* 2010). The recently published 1 000 human genome project paper (Abecasis *et al.* 2012) reported an excess of rare variants (MAF<0.5%) located at functional sites. Under the assumption that most mutations are deleterious, there would be a relatively higher selection pressure on genetic variants at functional sites than other sites. This may be used to explain why better predictions were obtained with low MAF bin markers in our study, however, whether such markers were located in functional loci requires further study. In other aspects, the heritability of traits (Calus and Veerkamp 2007; Habier *et al.* 2007; Zhu *et al.* 2016), the genetic structure of inference population (Habier *et al.* 2010) also have a great impact on the genomic predictive ability.

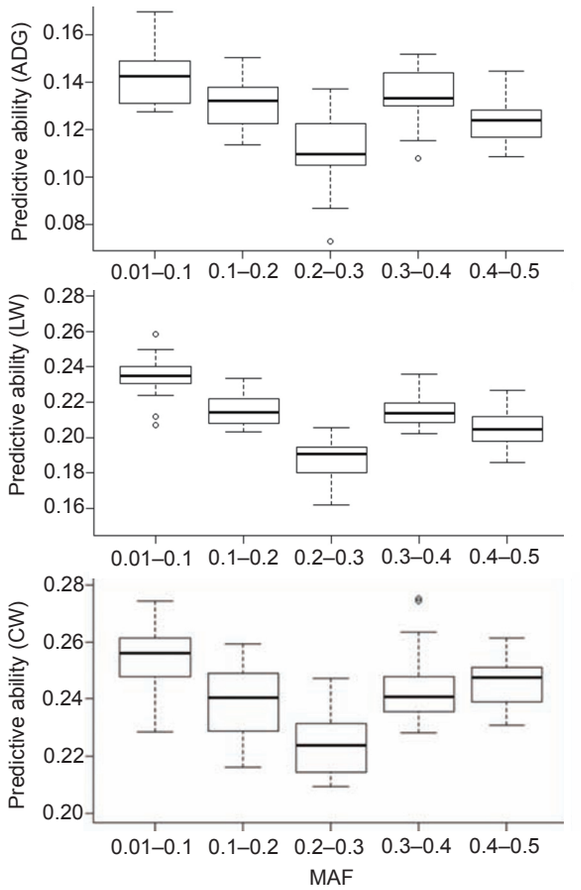

Fig. 7 Predictive abilities of 10-fold cross-validation for average daily gain (ADG, kg), live weight (LW, kg) and carcass weight (CW, kg) using different minor allele frequency (MAF) bins (0.01–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5). Error bar means standard deviation.

5. Conclusion

In our study, we found markers with low MAF showed better predictive ability than these with high MAF for three traits (ADG, LW, and CW) and selecting large SNP effects (Strategy 2) could obtain the predictive ability as well as all autosomal SNPs using GBLUP with less SNPs (20K) than that of selection based on evenly-spaced markers (Strategy 1) for the three growth traits. In conclusion, our findings could provide theoretical foundations for designing the SNP selection strategy and offer some useful insights to design a low-density panel with largest effect SNPs that are common for three traits in Chinese Simmental cattle.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31201782, 31672384 and 31372294), the Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences (ASTIP-IAS03), the Cattle Breeding Innovative Research Team of Chinese Academy of Agricultural Sciences (cxgc-ias-03), the Key Technology R&D Program of China during the 12th Five-Year Plan period (2011BAD28B04), the National High Technology Research and Development Program of China (863 Program 2013AA102505-4), and the Beijing Natural Science Foundation, China (6154032).

References

Abdollahi-Arpanahi R, Nejati-Javaremi A, Pakdel A, Moradi-Shahrbabak M, Morota G, Valente B D, Kranis A, Rosa G J, Gianola D. 2014. Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens. *Journal of Animal Breeding and Genetics*, **131**, 123–133.

Abecasis G R, Auton A, Brooks L D, DePristo M A, Durbin R M,

Handsaker R E, Kang H M, Marth G T, McVean G A. 2012. An integrated map of genetic variation from 1092 human genomes. *Nature*, **491**, 56–65.

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen K J, Hayes B J, Lawley C T, Sonstegard T S. 2012. Design of a bovine low-density SNP array optimized for imputation. *PLOS ONE*, **7**, e34130.

Boldman K G, Kriese L A, Van-Vleck L A, Van-Tassell C P, Kachman S D. 1995. *A Manual for Use of MTDFREML*. USDA, ARS, Clay Center, NE.

Calus M P, Veerkamp R F. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics*, **124**, 362–368.

Daetwyler H D, Pong-Wong R, Villanueva B, Woolliams J A. 2010. The impact of genetic architecture on genome-wide evaluation Methods. *Genetics*, **185**, 1021–1031.

Ding X, Zhang Z, Li X, Wang S, Wu X, Sun D, Yu Y, Liu J, Wang Y, Zhang Y, Zhang S, Zhang Y, Zhang Q. 2013. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. *Journal of Dairy Science*, **96**, 5315–5323.

Edriss V, Guldbrandtsen B, Lund M S, Su G. 2013. Effect of marker-data editing on the accuracy of genomic prediction. *Journal of Animal Breeding and Genetics*, **130**, 128–135.

Endelman J B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genetics*, **4**, 250–255.

Erbe M, Gredler B, Seefried F R, Bapst B, Simianer H. 2013. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLOS ONE*, **8**, e81046.

Gao H, Madsen P, Nielsen U S, Aamand G P, Su G, Byskov K, Jensen J. 2015. Including different groups of genotyped females for genomic prediction in a Nordic Jersey population. *Journal of Dairy Science*, **98**, 9051–9059.

Gianola D. 2013. Priors in whole-genome regression: The bayesian alphabet returns. *Genetics*, **194**, 573–596.

Gianola D, de los Campos G, Hill W G, Manfredi E, Fernando R. 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**, 347–363.

Gianola D, van Kaam J B. 2008. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, **178**, 2289–2303.

Goddard M E, Hayes B J. 2007. Genomic selection. *Journal of Animal Breeding and Genetics*, **124**, 323–330.

Goddard M E, Hayes B J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, **10**, 381–391.

de Los Campos G, Gianola D, Rosa G J. 2009. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. *Journal of Animal Science*, **87**, 1883–1887.

de Los Campos G, Vazquez A I, Fernando R, Klimentidis Y C, Sorensen D. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLOS Genetics*, **9**, e1003608.

Habier D, Fernando R L, Dekkers J C. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, **177**, 2389–2397.

Habier D, Tetens J, Seefried F R, Lichtner P, Thaller G. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, **42**, doi: 10.1186/1297-9686-42-5

Hayes B J, Bowman P J, Daetwyler H D, Kijas J W, van der Werf J H. 2012. Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, **43**, 72–80.

Hill W G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.

Lettre G. 2011. Recent progress in the study of the genetics of height. *Human Genetics*, **129**, 465–472.

Long N, Gianola D, Rosa G J, Weigel K A, Kranis A, Gonzalez-Recio O. 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research*, **92**, 209–225.

Lund M, de Roos A, de Vries A, Druet T, Ducrocq V, Fritz S, Guillaume F, Guldbrandtsen B, Liu Z, Reents R, Schrooten C, Seefried F, Su G. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics Selection Evolution*, **43**, 43.

Makowsky R, Pajewski N M, Klimentidis Y C, Vazquez A I, Duarte C W, Allison D B, de los Campos G. 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genetics*, **7**, e1002051.

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

Mikshowsky A A, Gianola D, Weigel K A. 2016. Improving reliability of genomic predictions for Jersey sires using bootstrap aggregation sampling. *Journal of Dairy Science*, **99**, 3632–3645.

Mullen M P, McClure M C, Kearney J F, Waters S M, Weld R, Flynn P, Creevey C J, Cromie A R and Berry D P. 2013. Development of a custom SNP chip for dairy and beef cattle breeding, parentage and research. Interbull Bulletin.

Niu H, Zhu B, Guo P, Zhang W, Xue J, Chen Y, Zhang L, Gao H, Gao X, Xu L, Li J. 2016. Estimation of linkage disequilibrium levels and haplotype block structure in Chinese Simmental and Wagyu beef cattle using high-density genotypes. *Livestock Science*, **190**, 1–9.

Ober U, Ayroles J F, Stone E A, Richards S, Zhu D, Gibbs R A, Stricker C, Gianola D, Schlather M, Mackay T F, Simianer H. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002685.

Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H. 2011. Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*, **188**, 695–708.

Ogawa S, Matsuda H, Taniguchi Y, Watanabe T, Nishimura S,

Sugimoto Y, Iwaisaki H. 2014. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. *BMC Genetics*, **15**, 15.

Park J-H, Gail M H, Weinberg C R, Carroll R J, Chung C C, Wang Z, Chanock S J, Fraumeni J F, Chatterjee N. 2011. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 18026–18031.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A, Bender D, Maller J, Sklar P, De Bakker P I, Daly M J. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Saatchi M, McClure M C, McKay S D, Rolf M M, Kim J, Decker J E, Taxis T M, Chapple R H, Ramey H R, Northcutt S L, Bauck S, Woodward B, Dekkers J C, Fernando R L, Schnabel R D, Garrick D J, Taylor J F. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, **43**, 40.

VanRaden P M. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**, 4414–4423.

VanRaden P M, Van Tassell C P, Wiggans G R, Sonstegard T S, Schnabel R D, Taylor J F, Schenkel F S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, **92**, 16–24.

Vazquez A I, Rosa G J M, Weigel K A, de los Campos G, Gianola D, Allison D B. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *Journal of Dairy Science*, **93**, 5942–5949.

Weigel K A, de los Campos G, Gonzalez-Recio O, Naya H, Wu X L, Long N, Rosa G J, Gianola D. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science*, **92**, 5248–5257.

Weigel K A, Van Tassell C P, O'Connell J R, VanRaden P M, Wiggans G R. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *Journal of Dairy Science*, **93**, 2229–2238.

Wolc A, Stricker C, Arango J, Settar P, Fulton J E, O'Sullivan N P, Preisinger R, Habier D, Fernando R, Garrick D J, Lamont S J, Dekkers J C. 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution*, **43**, doi: 10.1186/1297-9686-43-5

Yang J, Benyamin B, McEvoy B P, Gordon S, Henders A K, Nyholt D R, Madden P A, Heath A C, Martin N G, Montgomery G W, Goddard M E, Visscher P M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

Zhang Z, Ding X, Liu J, Zhang Q, de Koning D J. 2011. Accuracy of genomic prediction using low-density marker panels. *Journal of Dairy Science*, **94**, 3642–3650.

Zhu B, Zhu M, Jiang J, Niu H, Wang Y, Wu Y, Xu L, Chen Y, Zhang L, Gao X, Gao H, Liu J, Li J. 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.

(Managing editor ZHANG Juan)