



Factors affecting GEBV accuracy with single-step Bayesian models

Lei Zhou¹ · Raphael Mrode² · Shengli Zhang^{1,2,3} · Qin Zhang¹ · Bugao Li³ · Jian-Feng Liu¹

Received: 15 June 2017 / Revised: 4 September 2017 / Accepted: 14 September 2017 / Published online: 23 November 2017

© The Genetics Society 2018

Abstract

A single-step approach to obtain genomic prediction was first proposed in 2009. Many studies have investigated the components of GEBV accuracy in genomic selection. However, it is still unclear how the population structure and the relationships between training and validation populations influence GEBV accuracy in terms of single-step analysis. Here, we explored the components of GEBV accuracy in single-step Bayesian analysis with a simulation study. Three scenarios with various numbers of QTL (5, 50, and 500) were simulated. Three models were implemented to analyze the simulated data: single-step genomic best linear unbiased prediction (GBLUP; SSGBLUP), single-step BayesA (SS-BayesA), and single-step BayesB (SS-BayesB). According to our results, GEBV accuracy was influenced by the relationships between the training and validation populations more significantly for ungenotyped animals than for genotyped animals. SS-BayesA/BayesB showed an obvious advantage over SSGBLUP with the scenarios of 5 and 50 QTL. SS-BayesB model obtained the lowest accuracy with the 500 QTL in the simulation. SS-BayesA model was the most efficient and robust considering all QTL scenarios. Generally, both the relationships between training and validation populations and LD between markers and QTL contributed to GEBV accuracy in the single-step analysis, and the advantages of single-step Bayesian models were more apparent when the trait is controlled by fewer QTL.

Introduction

Meuwissen et al. (2001) first proposed the widely used genomic selection method using a dense marker panel for the genetic evaluation of animals and plants. This method achieves higher genetic evaluation accuracy and has the advantage of reducing generation intervals for some species such as dairy cattle with progeny testing schemes.

The accuracy of Genomic Best Linear Unbiased Prediction (GBLUP) was assumed to be mainly due to the linkage disequilibrium (LD) between markers and quantitative trait loci (QTL). However, Habier et al. (2007) demonstrated that GEBV accuracy depends not only on the LD between markers and QTL, but also on the genetic relationships among individuals captured by markers. According to their simulation study, GEBV accuracy decreases rapidly as the validation generation becomes distant from the generations of the training population, even when LD still exists between markers and QTL. Daetwyler et al. (2012) decomposed the components of GEBV accuracy by using a multi-breed sheep population. Surprisingly, they found that single-nucleotide polymorphism (SNP) markers from one single chromosome could achieve up to 86% of the accuracy of using all SNP markers, thus indicating that GEBV accuracy is not only due to LD between markers and QTL, but also due to population structure or genetic relationships among individuals. Habier et al. (2013) further demonstrated that the accuracy of GEBV within families depends largely on additive–genetic relationship information, and is also determined by the effective number of SNP markers and training data size.

Electronic supplementary material The online version of this article (<https://doi.org/10.1038/s41437-017-0010-9>) contains supplementary material, which is available to authorized users.

✉ Jian-Feng Liu
liujf@cau.edu.cn

- ¹ National Engineering Laboratory for Animal Breeding; Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture; College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
- ² Animal Biosciences, International Livestock Institute, Nairobi 00100, Kenya
- ³ Department of Animal Sciences and Veterinary Medicine, Shanxi Agricultural University, Taigu 030801, China

A single-step approach was proposed to overcome the limitation that not all animals are genotyped (Christensen and Lund 2010; Legarra et al. 2009). This approach has the merit of using all the genotyped and non-genotyped animals in one analysis, and it can estimate GEBV for all the animals in the analysis. It has been applied to the genetic evaluation of many livestock species, including pigs, chicken, and cattle (Aguilar et al. 2010; Chen et al. 2011; Christensen et al. 2012; Liu et al. 2014). Christensen et al. (2012) have shown that the single-step method provides improved accuracy for both genotyped and ungenotyped animals, whereas GBLUP can only be implemented for genotyped animals. The single-step method, compared with the GBLUP model, allows for less biased and more accurate GEBV predictions when the population is under strong selection (Vitezica et al. 2011). Moreover, Fernando et al. (2014) have presented single-step Bayesian regression models, which have the merit of modeling SNP effects with more flexible distributions (such as a *t*-distribution).

However, it is still unclear how different components such as LD between markers and QTL, in addition to population structure, contribute to GEBV accuracy in the single-step analysis. Furthermore, very few studies to date have investigated the relative performance of various single-step Bayesian models (Lee et al. 2017). Therefore, by using a simulation study, we investigated the contributions of GEBV accuracy in the single-step analysis in this study. Different numbers of generations between the validation and the training populations along with various numbers of QTL were simulated to show the contributions of these components to GEBV accuracy. We further investigated the performance of different single-step models (SSGBLUP, SS-BayesA, and SS-BayesB) in various scenarios with different number of QTL (5, 50, and 500) in the simulation.

Materials and methods

Models

SSGBLUP model

Legarra et al. (2009) and Christensen and Lund (2010) first proposed the single-step BLUP model, which has been further extended by Fernando et al. (2014) as follows:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} + \mathbf{e} \quad (1)$$

where \mathbf{y}_1 is the vector of phenotype for ungenotyped individuals and \mathbf{y}_2 is the vector of phenotype for genotyped

individuals. $\boldsymbol{\beta}$ is the vector of fixed effects and \mathbf{X}_1 and \mathbf{X}_2 are the incidence matrices for fixed effects of ungenotyped and genotyped individuals. \mathbf{Z}_1 and \mathbf{Z}_2 are the incidence matrices of ungenotyped and genotyped individuals, respectively. Here, \mathbf{g}_1 and \mathbf{g}_2 are GEBV of ungenotyped and genotyped individuals. Fernando et al. (2014) have further extended SSGBLUP model from the animal model to the marker effect model by defining

$$\begin{bmatrix} \hat{\mathbf{g}}_1 \\ \hat{\mathbf{g}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{T}}_1 \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\epsilon}} \\ \mathbf{T}_2 \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{T}_2 \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\epsilon}} \\ \mathbf{T}_2 \hat{\boldsymbol{\alpha}} \end{bmatrix} \quad (2)$$

where \mathbf{T}_2 is the centered and scaled observed genotype matrix of genotyped individuals ($\mathbf{T}_2 = \frac{(\mathbf{M}_j - 2p_j)}{\sqrt{\sum_{j=1}^m 2p_j(1-p_j)}}$, \mathbf{M}_j is

a vector of the genotype for all individuals of marker *j*, p_j is the minor allele frequency of marker *j*), $\hat{\boldsymbol{\alpha}}$ is the vector of estimated marker effects, and $\hat{\mathbf{T}}_1$ is the predicted or imputed genotype matrix for ungenotyped individuals with $\hat{\mathbf{T}}_1 = \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{T}_2$, where \mathbf{A}_{ij} is the partition of the pedigree relationship matrix \mathbf{A} that relates to $\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$. The variance and covariance matrix of $\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$ is $\text{cov}(\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2) = \mathbf{H}$, where \mathbf{H} was defined as (Legarra et al. 2009):

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12} \mathbf{A}_{22}^{-1} (\mathbf{G}_2 - \mathbf{A}_{22}) \mathbf{A}_{22}^{-1} \mathbf{A}_{12} & \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{G}_2 \\ \mathbf{G}_2 \mathbf{A}_{22}^{-1} \mathbf{A}_{12} & \mathbf{G}_2 \end{bmatrix} \quad (3)$$

where \mathbf{G}_2 is the genomic relationship matrix for genotyped individuals. The estimated marker effects are assumed to be normally distributed with $N(0, \text{I}\sigma_{\alpha}^2)$. The imputation residuals, $\boldsymbol{\epsilon}$, are assumed to be multivariate and normally distributed with $N(\mathbf{0}, (\mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{12}') \sigma_g^2)$. Here, σ_{α}^2 and σ_g^2 are the SNP variance and polygenic variance, respectively. The model further becomes

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{T}}_1 \boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \mathbf{T}_2 \boldsymbol{\alpha} \end{bmatrix} + \mathbf{e} \quad (4)$$

$$= \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\epsilon} + \mathbf{e}$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$, $\mathbf{U} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix}$, and $\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 \hat{\mathbf{T}}_1 \\ \mathbf{Z}_2 \mathbf{T}_2 \end{bmatrix}$.

The mixed model equation (MME) corresponding to Eq. (4) for the SSGBLUP marker effects model is

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} & \mathbf{X}'\mathbf{Z}_1 \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{\sigma_{\epsilon}^2}{\sigma_{\alpha}^2} & \mathbf{W}'\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{W} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_{\epsilon}^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y}_1 \end{bmatrix} \quad (5)$$

where $\mathbf{A}^{11} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{12}')^{-1}$, $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, and σ_e^2 is the residual variance.

SS-BayesA/B model

The BayesA/B model can be simply extended to the single-step analysis by using the predicted genotypes for the ungenotyped individuals (Fernando et al. 2014). In the BayesA model, marker variances are assumed to be different for different SNP markers, and marker variances are commonly handled with a scaled-inverse χ^2 prior (Fernando and Garrick 2013; Gianola et al. 2009; Meuwissen et al. 2001):

$$p(\sigma_{\alpha_j}^2 | \nu_{\alpha}, s_{\alpha}^2) = \frac{\left(\frac{\nu_{\alpha}s_{\alpha}^2}{2}\right)^{\frac{\nu_{\alpha}}{2}}}{\Gamma\left(\frac{\nu_{\alpha}}{2}\right)} (\sigma_{\alpha_j}^2)^{-(\frac{\nu_{\alpha}}{2}+1)} e^{-\frac{\nu_{\alpha}\sigma_{\alpha_j}^2}{2}} \tag{6}$$

where ν_{α} and s_{α}^2 are the degrees of freedom and scale of the scaled-inverse χ^2 prior, respectively, and j is the j th number of marker. The Mixed model equation (MME) for the single-step BayesA (SS-BayesA) model further become

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} & \mathbf{X}'\mathbf{Z}_1 \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{D}^{-1}\sigma_e^2 & \mathbf{W}'\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{W} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y}_1 \end{bmatrix} \tag{7}$$

where $\mathbf{D} = \text{Diag}(\sigma_{\alpha_j}^2)$.

For the single-step BayesB (SS-BayesB) model, the Mixed model equation (MME) is the same as Eq. (7), and marker effects are assumed to be independently distributed as follows:

$$\alpha_j | \pi_m, \sigma_{\alpha_j}^2 = \begin{cases} 0 & \text{with probability } (1 - \pi_m) \\ \sim N(0, \sigma_{\alpha_j}^2) & \text{with probability } \pi_m \end{cases} \tag{8}$$

where π_m is the proportion of markers that have non-zero effect. We can estimate π_m using a Beta($\alpha_{\pi}, \beta_{\pi}$) prior (Habier et al. 2011). π_m was fixed at 0.01 in this study.

The joint posterior densities of each single-step model and Markov Chain Monte Carlo (MCMC) sampling strategies for other parameters and hyper-parameters were illustrated in Supplementary File 1.

Data simulation

A simulation study was conducted with the program QMSim (Sargolzaei and Schenkel 2009). First, 5000 historical generations (generations 1–5000), each with 2000 animals, were simulated to generate LD between SNP markers (Fig. 1). Then, five recent generations (generations 5001–5005) were generated from the last historical generation (generation

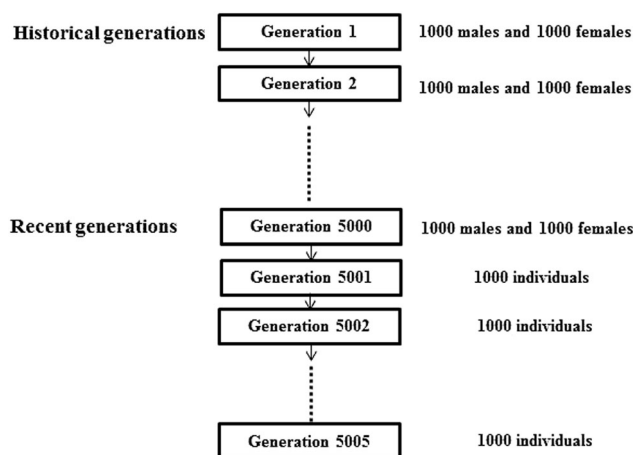


Fig. 1 The simulated historical and recent generations

5000) by random mating of 50 randomly selected males and 1000 females from the previous generation. There was no selection for the trait in each recent generation. For the recent population, each female had one offspring with an assuming male and female ratio of 1:1, and each recent generation had 1000 individuals. The dam’s culling rate was 0.5. Fifty percent of dams were from the last generation, and 50% were from the generation before last generation. All sires were from the last generation. For the genome, we simulated 20 chromosomes for each individual, and each chromosome had a length of one Morgan. On each chromosome, 2000 SNP markers were generated in generation 1. After data editing (MAF > 0.01 and $r^2 \leq 0.98$), the total number of SNPs retained for the analysis was close to 40,000 (range from 39,956 to 39,972 for each replicate). For the phenotype, heritability was set at 0.2. All individuals’ phenotypes were generated by summing true breeding values (QTL genotypes multiply by QTL effects) and residual effects (sampled from a normal distribution). Three scenarios with different number of QTL (5, 50, or 500) were considered, and QTL were randomly selected among all the SNPs. All other SNPs, except QTL, were assumed to have no effect on the trait. QTL effects were simulated from a normal distribution. The total number of replicates was 10 for each QTL scenario.

Data analysis

To investigate the influence of genetic relationships between the training and validation populations on the GEBV accuracy, we carried out single-step analysis by using all individuals from generations 5000–5002 as the training population, and all individuals from each generation 5003, 5004, and 5005 as a separate validation population. For each QTL scenario (5, 50, and 500 QTL), the design of the training and validation populations was shown

Table 1 The training and validation design for genomic prediction of the single-step analysis

Population specification		Number of individuals by gender		Origin of parents	
		Male	Female	Male	Female
Training	G_0^*	50	1000	G_{-1}^*	G_{-1}
	G_1	500	500	$50(G_0)$	$1000(G_0)$
	G_2	500	500	$50(G_1)$	$500(G_0)$ $500(G_1)$
Validation	G_3	500	500	$50(G_2)$	$500(G_1)$ $500(G_2)$
	G_4	500	500	$50(G_3)$	$500(G_2)$ $500(G_3)$
	G_5	500	500	$50(G_4)$	$500(G_3)$
					$500(G_4)$

Note: $*G_0$ indicates generation 0 of recent generations (generation 5000 from historical populations); G_{-1} indicates historical generation 4999, G_{1-5} indicates generations 5001–5005. This design was the same for all QTL scenarios (5, 50, and 500 QTL)

in Table 1. To mimic the single-step analysis, we set the genotyping rate at 50% for the training and individual validation populations. The genotyped individuals were randomly selected from the training and validation populations, and the remaining individuals were treated as ungenotyped individuals. To compare the prediction performance of different models, we computed GEBV accuracy as the correlation of GEBV and true breeding values (TBV). GEBV for the ungenotyped individuals were computed by $\hat{\mathbf{g}}_1 = \hat{\mathbf{T}}_1\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\varepsilon}}$, and GEBV for the genotyped individuals were computed by $\hat{\mathbf{g}}_2 = \mathbf{T}_2\hat{\boldsymbol{\alpha}}$.

To analyze the simulated data, we ran MCMC for 50,000 iterations with 5000 as the burn-in for all three models (SSGBLUP, SS-BayesA, and SS-BayesB) for each replicate within each QTL scenario. For the SS-BayesA model, we estimated hyper-parameters scale (s_{α}^2) and degree of freedom (ν_{α}). For the SS-BayesB model, to simplify the model structure, we fixed the proportion of non-zero effect SNPs (π_m) as 0.01. The data were analyzed by self-developed R and C codes, and they were available on request.

Results

Influence of the relationships between the training and validation populations

We investigated the GEBV accuracy by using different generations (5003, 5004, and 5005) as the validation population and generations 5000–5002 as the training

population for all three single-step models (SSGBLUP, SS-BayesA, and SS-BayesB). Figure 2 shows GEBV accuracy for both genotyped and ungenotyped individuals of generations 5003, 5004, and 5005 each as the validation population for all QTL scenarios. As the generation number of the validation population (measured as distances between validation and training populations) increased with respect to the training population, GEBV accuracy decreased significantly for both genotyped and ungenotyped individuals. For SSGBLUP, GEBV accuracy always decreased with validation generation for both ungenotyped and genotyped individuals. For SS-BayesA and SS-BayesB, the GEBV accuracy for genotyped individuals did not decrease dramatically, compared with SSGBLUP at scenarios of 5 or 50 QTL. For ungenotyped validation individuals, the accuracy of GEBV decreased with the increase of the generation of the validation population. However, for genotyped individuals, the influence of generations of the validation population on GEBV accuracy was more sensitive to both different single-step models and the number of QTL (Fig. 2).

Comparison of single-step models

We also compared GEBV accuracy for the three single-step models (SSGBLUP, SS-BayesA, and SS-BayesB). For the scenarios with 5 and 50 QTL, SS-BayesA and SS-BayesB always achieved higher accuracy than SSGBLUP, and SS-BayesB performed better than SS-BayesA at validation population generations 5003 and 5004 (Fig. 3). However, when the number of QTL was 500 in the simulation, no advantage of SS-BayesA and SS-BayesB was found, and SS-BayesB realized the lowest GEBV accuracy. These findings indicated that the single-step Bayesian-type models had an advantage over the SSGBLUP model when there were fewer QTL affecting the trait. Moreover, considering the scenario with 5 and 50 QTL, we observed that the single-step Bayesian models exceeded SSGBLUP by a larger margin for the genotyped animals than for the ungenotyped animals (Fig. 3). When there were only 5 QTL in the simulation (with $h^2 = 0.2$), the GEBV accuracy of SS-BayesA and SS-BayesB for genotyped animals exceeded 0.93, while it was below 0.5 with SSGBLUP.

Influence of different number of QTL

We further compared the effect of different numbers of QTL for each single-step model. According to Fig. 4, it was clear that the GEBV accuracy of SSGBLUP did not change significantly as the number of QTL increased. However, the GEBV accuracy of SS-BayesA and SS-BayesB for both ungenotyped and genotyped individuals decreased

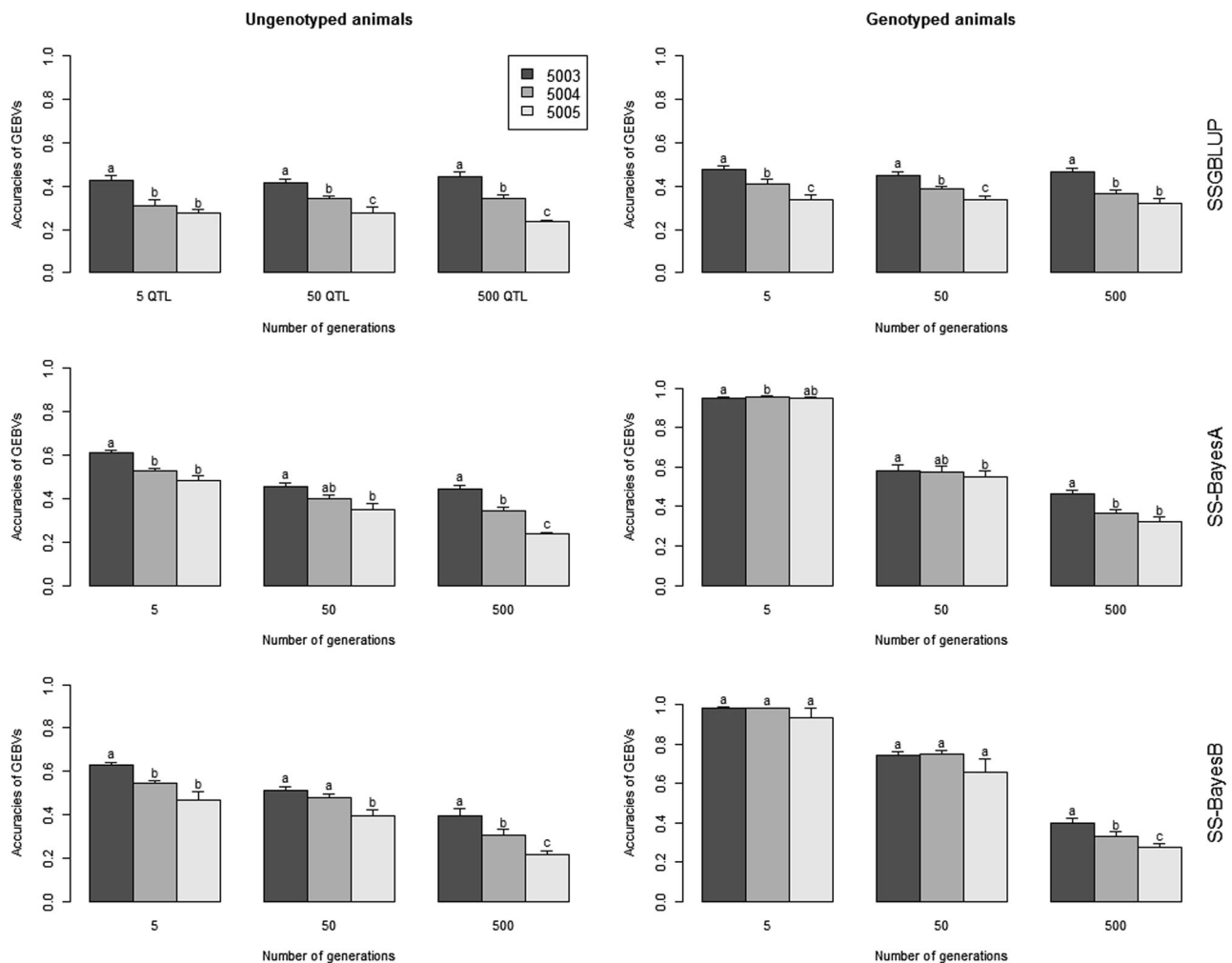


Fig. 2 Accuracy of GEBV for different validation population generations, using SSGBLUP, SS-BayesA, and SS-BayesB models at scenarios of 5, 50, and 500 QTL. These results are the means and

standard errors of 10 replications. Different letters indicate a significant difference at P value < 0.05

significantly when the number of QTL increased. These results indicate that single-step Bayesian models are more sensitive to the number of QTL affecting the trait, even when the relationship structure of the training and validation populations is almost the same for the various number of QTL scenarios. Meanwhile, SSGBLUP is a robust model to handle scenarios with different number of QTL.

Discussion

The objective of this study was to analyze the influence of relationships between the training and validation populations and of LD between markers and QTL on the GEBV accuracy with various single-step models. We further extended single-step Bayesian models of Fernando and Garrick (2013) to single-step BayesB model, and

investigate three single-step models (SSGBLUP, SS-BayesA, and SS-BayesB) with a simulation study. To investigate the influence of relationships between the training and validation populations, we used each one of three successive generations (5003, 5004, and 5005) as the validation population. Generally, GEBV accuracy decreased as the distance (measured as the number of generation gap between the training population and validation population) of validation population increased for different single-step models, which was in agreement with results of many previous studies (Habier et al. 2013; Habier et al. 2010; Kang et al. 2016; Wolc et al. 2011). The relationship between the training and validation populations influenced GEBV accuracy more substantially than LD between markers and QTL. Moreover, we observed that Bayesian-type single-step models (SS-BayesA and SS-BayesB) outperformed SSGBLUP in the scenarios with

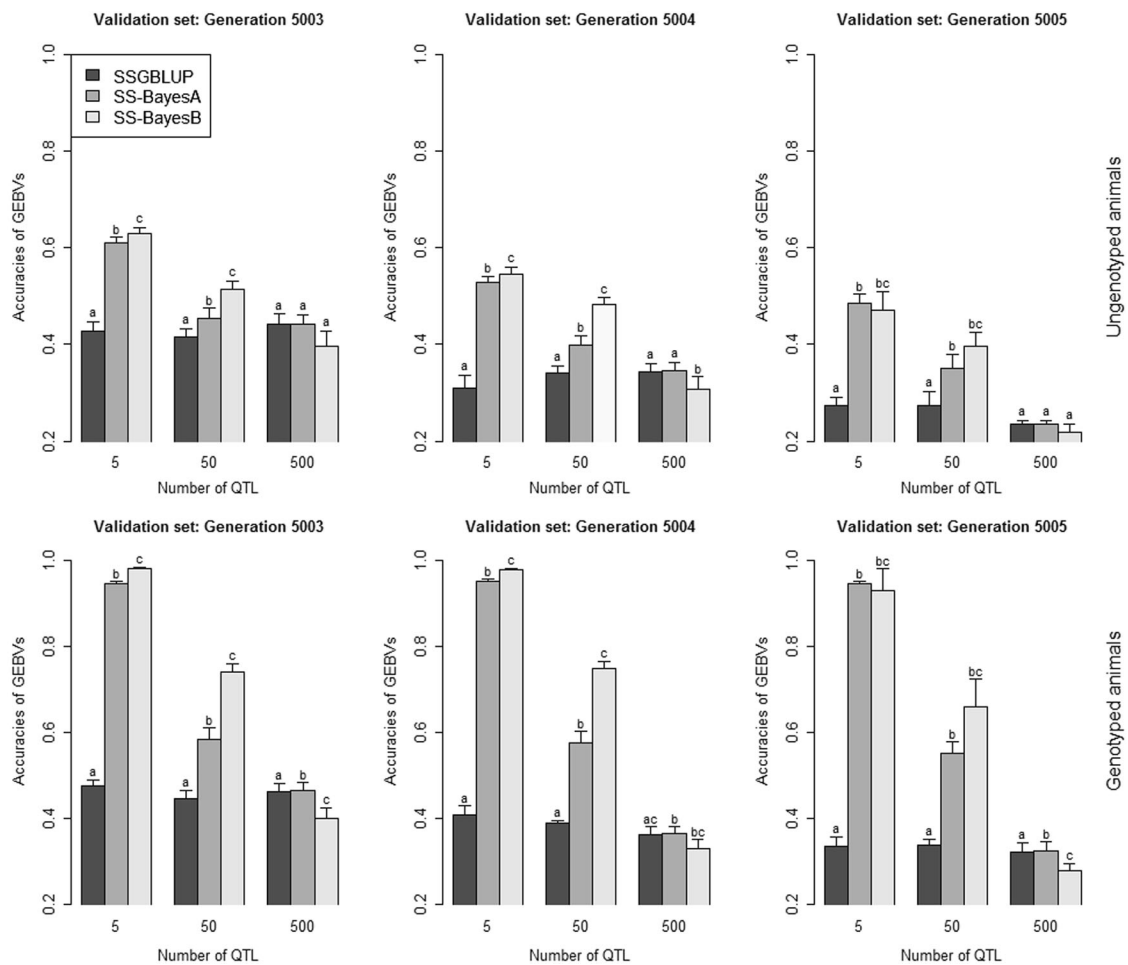


Fig. 3 Accuracy of GEBV for different single-step models at each validation generation. *Note:* The single-step models are SSGBLUP (single-step GBLUP), SS-BayesA (single-step BayesA), and SS-

BayesB (single-step BayesB). These results are the means and standard errors of 10 replications. Different letters indicate a significant difference at $P < 0.05$

fewer QTL (5 or 50 QTL), whereas SSGBLUP outperformed Bayesian models (SS-BayesB) when the number of QTL reached 500 in the simulation.

Influence of relationships between the training and validation populations

GEBV accuracy decreased as the generation of validation population increased, especially with SSGBLUP. To further investigate the reason for this, we computed the average linkage disequilibrium (r^2) of all adjacent SNP pairs for all individuals in each recent generation (generations 5001–5005), along with the means and standard deviations of pedigree-based genetic relationships (A_{12}) of the training (generations 5000–5002) and validation populations (each generation of 5003–5005), along with the means and standard deviations of pedigree-based genetic relationships (A_{12}) of the training (generations 5000–5002) and validation populations (each generation of 5003–5005). It can be

seen from Fig. 5 that the LD between adjacent SNP markers slightly increased with the number of generation. The average A_{12} between the training and validation populations was almost the same for validation generations 5003, 5004, and 5005. However, the standard deviations of A_{12} decreased by the number of generation. These results indicated that there were more individuals with closer genetic relationships with the training population for validation generation 5003 compared with that of generation 5005. These few animals that had close relationships with the training population caused the overall GEBV accuracy of generation 5003 to be higher than that of generation 5005 (results not shown).

Habier et al. (2010) have also found that the accuracy of GEBV for four traits (milk yield, fat yield, protein yield, and somatic cell score) decreased when the relationship between the training and validation populations decreased in German Holstein bulls’ data. Kang et al. (2016) have also found that the GEBV accuracy declined by generation in the

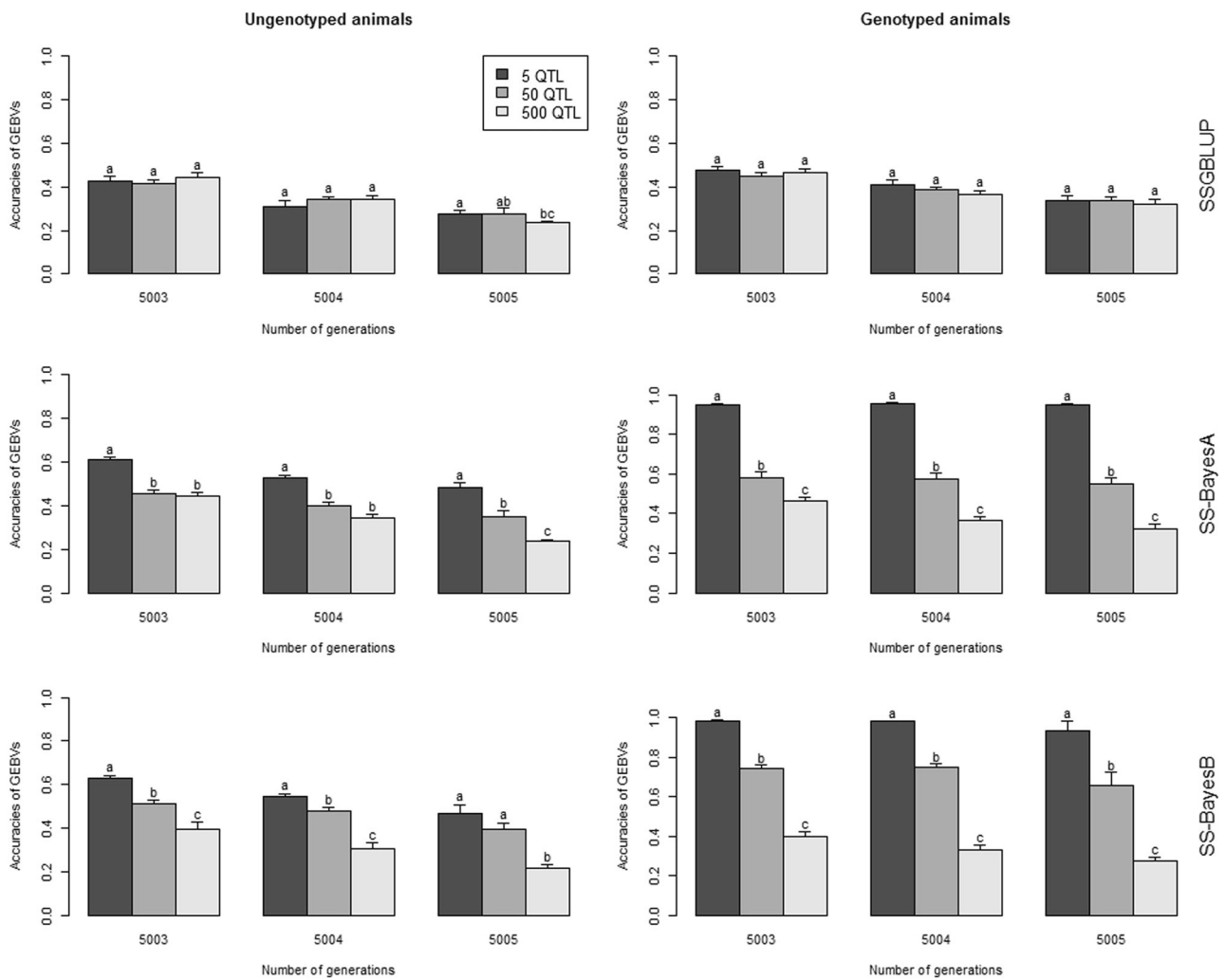


Fig. 4 Accuracy of GEBV for different numbers of QTL at each validation generation. Note: The single-step models are SSGBLUP (single-step GBLUP), SS-BayesA (single-step BayesA) and SS-

BayesB (single-step BayesB). These results are the means and standard errors of 10 replications. Different letters indicate a significant difference at P value <0.05

single-step analysis with a simulation study. Daetwyler et al. (2012) have conducted a genomic prediction analysis using a multiple-breed sheep population, and have also found that a large amount of GEBV accuracy was due to population structure or family relationships instead of LD between markers and QTL at current marker densities. Therefore, our results and those of Kang et al. (2016) indicate that the accuracy of GEBV for the single-step analysis decreases when the generation gap between the training and validation populations increases. In addition, our results indicate that this decrease in GEBV accuracy occurred even when the LD between markers increased marginally (with P value <0.001 for t -test of r^2) (Fig. 5). This finding indicates that the relationship between training and validation populations plays a more important role than the LD between markers and QTL in the GEBV accuracy for both genotyped and ungenotyped individuals, especially

with the single-step BLUP model. Habier et al. (2007) have also concluded that the GBLUP model (or RR-BLUP) was influenced mostly by genetic relationships.

For the ungenotyped individuals, a pedigree-based relationship is used for the prediction of their genotypes. According to the formula for the predicted genotype of ungenotyped individuals $\hat{\mathbf{T}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{T}_2$, it is obvious that a larger element in \mathbf{A}_{12} corresponds to a higher regression coefficient for the corresponding element of \mathbf{T}_2 (Here, \mathbf{T}_2 is the centered genotype matrix of genotyped individuals and $\hat{\mathbf{T}}_1$ is the predicted genotype matrix of ungenotyped individuals). Chen et al. (2014) have also reported that individuals with close relatives in the training population had higher genotype imputation accuracy and higher accuracy of genomic prediction. Our results further illustrate that the genetic relationships between the training and validation populations affect GEBV accuracy more strongly for

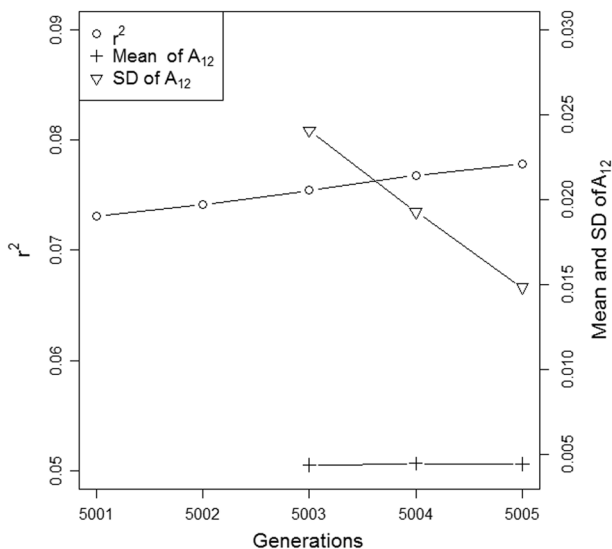


Fig. 5 The plot of averaged linkage disequilibrium (r^2) and the means and standard deviations of A_{12} by generation. *Note:* The r^2 was calculated using genotypes of all adjacent SNP markers, and A_{12} was the pedigree-based numeric relationship between the training and validation populations. All these statistics are the means of 10 replicates in the scenario of 50 QTL

ungenotyped individuals than genotyped individuals in single-step analysis (Fig. 2). This is so because the GEBV of ungenotyped individuals is composed of two parts: (1) estimated marker effects and (2) imputation residuals. Imputation residuals are estimated on the basis of a pedigree relationship matrix $\left(\boldsymbol{\varepsilon} \propto N\left(\mathbf{0}, \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{12}'\right)\sigma_g^2\right)\right)$, whereas the GEBV of genotyped individuals only depends on the estimated marker effects.

Single-step models comparison

In this study, we used three single-step models such as SSGBLUP, SS-BayesA, and SS-BayesB. The Bayesian-type models outperformed the SSGBLUP models when there were fewer QTL (5 or 50 QTL) in the simulation. These results were in agreement with both previous simulation and real data analyses of genomic selection (Habier et al. 2007; Hayes et al. 2009; Meuwissen et al. 2001). SS-BayesB model showed extremely optimistic prediction ability at the case of 5 QTL (Fig. 3). Here the GEBV accuracy values were 0.98 for SS-BayesB and 0.48 for SSGBLUP. The advantage of SS-BayesA and SS-BayesB was mainly due to both models assuming a non-normal distribution of marker effects, in which a t -distribution was assumed for SS-BayesA and a two components mixture distribution for SS-BayesB (Gianola et al. 2009; Habier et al. 2011). These marker effects' assumptions more closely matched to the QTL and marker structure in our simulation, especially with the 5 and 50 QTL scenarios. Meanwhile,

according to Figs. A1–A3, it could be seen that the estimated marker effects had much larger range (−600.0 to 200.0) for SS-BayesA/B compared to that of SSGBLUP (−10.0 to 10.0) at the 5 and 50 QTL scenarios. Zhang et al. (2016) found greater accuracies using weighted genomic relationships (vs. regular single-step GBLUP, BayesB, and BayesC) when few QTLs were simulated, and their weighted genomic relationships approaches (WssGBLUP) were more similar to the SS-BayesA/B models in this study.

However, when the number of QTL increased to 500, the Bayesian-type models had no advantage over SSGBLUP (Fig. 3). Interestingly, SS-BayesB obtained lower accuracy than SSGBLUP and SS-BayesA for the 500 QTL scenario. In this scenario, 500 SNPs were simulated as QTL in the phenotype simulation, while only ~400 SNPs (1% of 40,000 SNPs) were allowed to have non-zero effects in SS-BayesB model, as the non-zero proportion (π_m) of markers was fixed at 0.01. This fixation potentially limited the power of SS-BayesB model to capture all existing 500 QTL.

Generally, considering all different scenarios of QTL, SS-BayesA model was the most efficient and robust according to our simulation analysis. The SS-BayesB model with the freedom of estimating π_m would capture more LD between markers and QTL, and may obtain better GEBV prediction performance. Karaman et al. (2016) have reported that BayesB and BayesC have no advantage over GBLUP when the reference population is small (<6000 individuals). Therefore, given the findings from Karaman et al. (2016) and several other studies (Habier et al. 2013; Habier et al. 2010; Kang et al. 2016), the advantages of Bayesian models in genomic selection and single-step analysis depend on the training population size, number of QTL for the trait, and other potential factors.

Three single-step models also performed differently in terms of prediction bias. On the basis of regression coefficients of TBV on GEBV and means of deviation between TBV and GEBV, SSGBLUP model achieved the least prediction bias for ungenotyped individuals, and SS-BayesA and SS-BayesB models realized less prediction bias for genotyped individuals. These models need to be further investigated for prediction bias for the application in real data.

Influence of the number of QTL

The three single-step models performed differently as the number of QTL increased in the simulation. For the SSGBLUP model, the GEBV accuracy changed minimally as the number of QTL increased from 5 to 500 (Fig. 4). This could be explained by the fact that SSGBLUP mainly utilized the genomic relationship among the training and validation populations, instead of capturing the LD between

markers and QTL. When the number of QTL changed from 5 to 500, the genetic relationship between the training and validation populations did not vary, and thus the GEBV accuracy of SSGBLUP showed little change. However, for the SS-BayesA and SS-BayesB models, the GEBV accuracy decreased as the number of QTL increased. This was because the marker effects were more accurately estimated for the scenarios of 5 and 50 QTL with the single-step Bayesian models. From Figs. A2 and A3 in the Supplementary File 2 (the plots of estimated marker effects for one replicate), it can be seen that only a few SNPs adjacent to the true QTL were estimated with large non-zero effects by SS-BayesA and SS-BayesB at the 5 and 50 QTL scenarios. When the number of QTL was 500 in the simulation, SSGBLUP and SS-BayesA, which allowed all markers to have a non-zero effect, showed better agreement between estimated marker effects and true QTL effects (Figs. A1–A3 in Supplementary File 2), in addition to higher GEBV accuracy compared with SS-BayesB (Fig. 3). To further investigate the influence of QTL numbers, we have also simulated a scenario of 5000 QTL. The results (Fig. A4) also indicated that SSGBLUP and SS-BayesA had obvious advantages over SS-BayesB model, which was similar to the scenario of 500 QTL.

Generally, our results suggest that single-step Bayesian models have appealing advantages when the number of QTL controlling the trait is small (Zhang et al. 2016). Kang et al. (2016) have proposed a single-step random regression model (single-step random regression test-day model, SS RR-TDM) for longitudinal traits, and SS RR-TDM has been found to have an advantage over the pedigree-based RR-TDM and GBLUP. It will be meaningful to further extend single-step Bayesian models to longitudinal traits.

Currently, a new algorithm that uses recursion to compute the genomic relationship matrix has become commonly applied (Misztal 2016; Misztal et al. 2014). This algorithm is also called “algorithm for proven and young,” which splits genotyped animals into core (proven) animals and noncore (young) animals. This methodology can produce an inverse genomic relationship matrix of all genotyped animals by only computing the inverse of core animals (Misztal et al. 2014), thereby dramatically decreasing the computing cost compared with the traditional single-step GBLUP. Because the breeding values of non-core animals can be derived by recursions on the breeding values of core animals (Misztal 2016), these results indicate that phenotypes of core or proven animals are sufficient for estimating markers effects in the Bayesian-like model. Therefore, how to extend the core and noncore concept to the Bayesian and single-step Bayesian genomic models will be an interesting and valuable research topic. Fernando et al. (2016) further proposed a hybrid model for the single-step Bayesian models with an efficient new computing

algorithm, and they are easy to extend to multiple traits and multiple-breed analyses.

Data archiving

The simulated data analyzed in this study is available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.hk14j>.

Acknowledgments The authors appreciate the financial support provided by the National Major Development Program of Transgenic Breeding (2014ZX0800953B), the National High Technology Research and Development Program of China (863 Program 2013AA102503), the National Natural Science Foundations of China (31661143013), and the Program for Changjiang Scholar and Innovation Research Team in University (IRT_15R62).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

- Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score 1. *J Dairy Sci* 93(2):743–752
- Chen CY, Misztal I, Aguilar I, Tsuruta S, Meuwissen THE, Aggrey SE *et al* (2011) Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. *J Anim Sci* 89(1):23–28
- Chen L, Li C, Sargolzaei M, Schenkel F (2014) Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. *PLoS ONE* 9(7):e101544
- Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. *Genet Sel Evol* 42(1):1–8
- Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G (2012) Single-step methods for genomic evaluation in pigs. *Animal* 6(10):1565–1571
- Daetwyler HD, Kemper KE, Jh VDW, Hayes BJ (2012) Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci* 90(10):3375–3384
- Fernando RL, Dekkers JC, Garrick DJ (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet Sel Evol* 46(1):1–13
- Fernando RL, Garrick D (2013) Bayesian methods applied to GWAS. In: Gondro C, van der Werf J, Hayes B (eds) *Genome-wide association studies and genomic prediction*. Humana Press, Totowa, NJ, pp. 237–274
- Fernando RL, Hao C, Golden BL, Garrick DJ (2016) Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. *Genet Sel Evol* 48(1):96
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the bayesian alphabet. *Genetics* 183(1):347–363
- Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177(4):2389–2397

- Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194(3):597–607
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12(1):186
- Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol* 42(1):5
- Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* 91(1):47
- Kang H, Zhou L, Mrode R, Zhang Q, Liu JF (2016) Incorporating single-step strategy into random regression model to enhance genomic prediction of longitudinal trait. *Heredity* 119, 459–467
- Karaman E, Cheng H, Firat MZ, Garrick DJ, Fernando RL (2016) An upper bound for accuracy of prediction using GBLUP. *PLoS ONE* 11(8):e0161054
- Lee J, Hao C, Garrick D, Golden B, Dekkers J, Park K *et al* (2017) Comparison of alternative approaches to single-trait genomic prediction using genotyped and non-genotyped Hanwoo beef cattle. *Genet Sel Evol* 49(1):2
- Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. *J Dairy Sci* 92(9):4656–4663
- Liu Z, Goddard ME, Reinhardt F, Reents R (2014) A single-step genomic model with direct estimation of marker effects. *J Dairy Sci* 97(9):5833–5850
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829
- Misztal I (2016) Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202(2):401–409
- Misztal I, Legarra A, Aguilar I (2014) Using recursion to compute the inverse of the genomic relationship matrix. *J Dairy Sci* 97(6):3943–3952
- Sargolzaei M, Schenkel FS (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics* 25(5):680–681
- Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011) Bias in genomic predictions for populations under selection. *Genet Res* 93(5):357–366
- Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R *et al* (2011) Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet Sel Evol* 43(1):23
- Zhang X, Lourenco D, Aguilar I, Legarra A, Misztal I (2016) Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. *Front Genet* 7(134):151