IJAS Iranian Journal of Applied Animal Science

Effect of Selection on the Accuracy of Genomic Selection and Genome-Wide Association Analysis

Research Article

A.A. Shadparvar[1], N. Ghavi Hossein-Zadeh[1], Z. Lotfi[1] and A. Safari[1*]

[1] Department of Animal Science, Faculty of Agricultural Sciences, University of Guilan, Rasht, Iran

Received on: 1 Nov 2024
Revised on: 27 May 2025
Accepted on: 11 Jun 2025
Online Published on: Jun 2025

*Correspondence E-mail: ab.safari64@gmail.com

Online version is available on: www.ijas.ir
https://doi.org/10.71798/ijas.2025.1214879

ABSTRACT

Several factors, such as trait heritability, marker density, distance between individuals in the reference population and selection candidates, as well as the number of phenotypic records in the reference population dataset, significantly affect the accuracy of genomic evaluation. The objective of the present study was to evaluate the effect of selection on the accuracy of genomic selection and genome-wide association analysis. Four selection schemes were considered: 1) no selection in both reference and validation populations. 2) selection in the reference population based on the estimated accuracy of estimated breeding values of 0.7, while no selection was applied in the validation population. 3) selection in both the reference and validation populations, comparable to scheme 2. 4) selection in both populations similar to that of scheme 3, however, the selection was done in the validation population with an accuracy of 0.5. In each scenario, a reference population and a validation population were randomly simulated using QMSim. The results indicated that the comparison of accuracy levels between the reference and validation populations showed higher accuracy in the reference population for all schemes, heritabilities, numbers of markers, and QTL numbers. The highest genomic evaluation accuracy for both populations was achieved in scheme 1. Results suggested that the correlation coefficient between genomic evaluation accuracy in the reference and validation populations was related to the number of common single nucleotide polymorphisms (SNPs) between the reference and validation populations. Selection in the reference population led to a significant reduction in the number of markers.

KEY WORDS accuracy, genome-wide association, genomic selection, quantitative trait locus, reference population, simulation, validation population.

INTRODUCTION

In previous years, the application of genetic markers in breeding programs faced technical limitations despite the simplicity of the concept. Access to a high density of single nucleotide polymorphism (SNP) markers has provided new opportunities for this endeavor. The use of high-density genetic markers allows for the prediction of genetic value across the entire genome, including quantitative trait loci (QTL), and their applications for selection purposes. SNP genotyping has made it possible to determine genotypes for a large number of animals at thousands of marker loci in a single analysis, resulting in a low cost per marker (Williams, 2005).

Genomic selection (GS) predicts the total genetic value using high-density marker maps, especially whole-genome SNPs, which are often in linkage disequilibrium with their neighboring QTLs. GS estimates the effects of thousands of DNA markers simultaneously (Meuwissen *et al.* 2013). GS allows for the development of new breeding strategies aimed at accelerating genetic progress while reducing costs and optimizing different breeding programs (Bouquet and

Juga, 2013). Several factors, such as trait heritability, marker density, distance between individuals in the reference and validation population, as well as the number of phenotypic records in the reference population dataset, significantly affect the accuracy of genomic evaluation (Meuwissen *et al.* 2001).

Accuracy refers to the correlation between true genetic value and genomic estimated breeding value (GEBV). The most reliable method for determining accuracy involves predicting the breeding values of candidate selections and then calculating the correlation between their actual genetic values and large offspring test (LOD) scores (Meuwissen *et al.* 2013). Genome-wide Association Studies (GWAS) have been used to identify genomic regions associated with variations in production and fertility traits. It has been observed that the accuracy of trait mapping improves with haplotype length, as a significant number of valid haplotypes have been identified across multiple breeds (Pryce *et al.* 2010). Interest in GWAS studies related to dairy cattle breeding stems from the discovery of markers that can enhance the accuracy of genetic values and improve our understanding of economically important traits. A notable characteristic of dairy cattle populations is their small effective population size, primarily due to the widespread use of artificial insemination (AI). This has impacted the pattern of Linkage Disequilibrium (LD) in dairy cattle breeds. Given that GWAS relies on LD, it should be capable of identifying significant associations in dairy cattle with markers located approximately every 100 kb (De Roos *et al.* 2008).

Nowadays, artificial insemination of dairy cows in many countries is performed using sperm obtained from advanced countries to enhance the genetic quality of dairy herds. With the widespread adoption of genomic selection in most countries that export improved genetic material as the reference population, the use of genomic selected males or their sperm has become common in the dairy herds of recipient countries as the validation population. Since the selection experience in reference countries may differ from that in validation countries, it is expected that the predictive accuracy of genomic breeding values in the validation population may vary due to selection effects.

This difference could lead to the expected genetic progress not occurring in dairy herds of countries that consume genetic material produced in reference countries. To date, no study has been conducted on this issue. Therefore, this study investigated the effect of differences or similarities in selection history between reference and validation populations on the accuracy of selection through stochastic simulation. Additionally, to broaden the perspective on results and enhance their interpretation, various levels of heritability for the trait under investigation were considered. Moreover, since marker density and the number of effective QTLs influencing the trait can also impact the accuracy of genomic evaluation, different levels were considered for these factors. Considering that selection can have effects on gene flow as a selection footprint, the results of Genome-Wide Association Study (GWAS) were also examined in each simulated scenario to achieve a better understanding of the impact of selection on the accuracy of selection in the validation population.

MATERIALS AND METHODS

Population simulation

This study utilized simulations conducted with QMSim (version 1.10) (Sargolzaei and Schenkel, 2009). Initially, a base population was simulated to establish linkage disequilibrium between markers and QTLs, as well as to find a balance between mutation and drift. Mutations occurred randomly at markers and QTLs, with a mutation rate assumed to be 2.5 *10-5 per locus per generation. This population, known as the Historical Population, began with 5000 individuals in generation zero, with an equal gender distribution. The size of the historical population decreased linearly in each generation, reaching 100 individuals by the 1000^{th} generation, remaining constant until the 1500^{th} generation. Subsequently, the population size gradually increased linearly, with the population reaching 3000 individuals by the 2000^{th} generation.

In the second stage, a reference population and a validation population, each consisting of 1000 females and 50 males, were randomly selected as founder individuals from the last generation of the historical population. In these populations, both males and females were selected for 10 consecutive generations based on one of the four described schemes. Subsequently, random mating was established among the selected individuals to generate the next generation. For each female animal, one offspring was simulated in each generation. Generations were simulated separately.
In the third stage, four different schemes for selection in the reference and validation populations were simulated. These schemes were as follows: Scheme 1: no selection in both reference and validation populations. Scheme 2: selection in the reference population was based on the estimated accuracy of estimated breeding values 0.7, while no selection was applied in the validation population. The estimation of individuals' breeding values in the reference population was carried out using the true additive genetic variance of the population, utilizing the "true_av" function in QMSim. Scheme 3: selection in both the reference and validation populations, comparable to scheme 2. Scheme 4: selection in both populations was similar to that of scheme 3, however the selection was done in the validation population

with an accuracy of 0.5.

Genome simulation

To simulate and investigate the effect of trait heritability, marker density, and the number of QTLs on the accuracy of GEBVs, a 100 cM genome was simulated. This genome consisted of one chromosome. The number of SNP markers was considered at three levels (5000, 7500, 10000), the number of QTLs was assumed at three levels (1000, 750, 500), and trait heritability was assessed at three levels (0.5, 0.3, 0.1). All markers and QTLs had two alleles. The effects of QTLs were sampled with a gamma distribution with shape and scaling parameters of b= 0.40 and a= 1.66, respectively (McHugh *et al.* 2011). The phenotypic variance of the trait was 1, and in all levels of trait heritability, QTL heritability was equal to trait heritability (Table 1).

Simulation scenarios and estimation of marker effects

In this study, as regards the number of heritability levels, the count of markers, the number of QTLs, and four selection schemes, 108 different scenarios were simulated. For each model, 20 repeats were conducted, and for each repeat, information regarding pedigree, generation, gender, phenotype and true hereditary value of individuals were included.

Marker effects

The effects of markers were estimated using BGLR (Bayesian Generalized Linear Regression) software (version 1.1.0) (De los Campos and Pérez-Rodríguez, 2014). This estimation was performed using information from the 9th generation of the reference population. Subsequently, with these estimates and available genetic information, genomic breeding values for individuals in the 9th and 10th generations were predicted for both the reference and validation populations. To predict the marker effects, Bayesian Ridge Regression (BRR) was employed using the following equation:

$$Y = Xb + Zg + e \quad \text{(Equation 1)}$$

Where:
Y: phenotypic values of a trait.
X: incidence matrix for the correlation of observations with fixed effects of the model.
b: vector of fixed effects of the model including the population mean, generation effect, and gender effect.
Z: incidence matrix that correlates the phenotype and genotype of animals for different marker loci.
g: vector of random marker effects.
e: vector of random residual errors.

Then, the genomic breeding value was estimated using Equation 2:

$$GEBV = Zg\hat{} \quad \text{(Equation 2)}$$

Where:
$g\hat{}$: vector of estimated marker effects.

The estimation of genetic parameters and marker effects was conducted through Gibbs sampling with 10000 samples, where the first 2000 samples were discarded as burn-in. To investigate the effect of different selection schemes on the number of significant markers for a trait and the allelic frequency of these markers, a GWAS analysis was performed for each scenario. In this analysis, the "glm()" function in the R programming was utilized to estimate the regression coefficient of the study trait phenotype on the corrected genotype of each SNP marker. The significance of the coefficient was determined by variance decomposition. Subsequently, this probability was adjusted using the "Benferroni" method and the "p.adjust()" function in the R program. Prior to performing the regression analysis, genotype markers were adjusted using the two principal components obtained from principal analysis conducted with the "dudi.pca()" function in R.

RESULTS AND DISCUSSION

Figure 1 depicts the change in the mean accuracy of genomic evaluation in the reference and validation populations for varying heritability levels. Scheme 1, where no selection occurs in either population, shows higher accuracy levels compared to other schemes. In schemes with selection, a decrease in genetic diversity of the trait in the reference population results in reduced accuracy. The accuracy levels of selection are similar in the reference population for all schemes but differ in the validation population. Scheme 2, with selection in the reference but not in the validation population, shows the lowest accuracy level, possibly due to variations in marker-phenotype relationships between populations. Therefore, using marker effects from reference populations with high accuracy for predicting genetic values in populations without selection may not be advisable.

The accuracy changes in the validation population for schemes 3 and 4 were similar, with slightly higher values in Scheme 4 indicating a lesser reduction in genetic diversity from past selection compared to scheme 3. Accuracy levels in all schemes increased with higher heritability of the trait. The rate of accuracy increase was greater from heritability 1.0 to 3.0 than from 3.0 to 5.0, suggesting a non-linear relationship between accuracy and heritability.

Table 1 Population structure and simulated parameters in the study

Population structure	Parameter	Number	Genome feature	Value
Historical population	Number of individuals (Gen)	5000 (0)	Number of chromosome	1
	Number of individuals (Gen)	100 (1000)	Chromosome length	100 cM
	Number of individuals (Gen)	100 (1500)	Number of markers	10000, 7500, 5000
	Number of individuals (Gen)	3000 (2000)	Type of marker	Biallelic SNP
Reference population	Number of males	50	Number of QTLs	10000, 7500, 5000
	Number of females	1000	QTL effects	0.4 Gamma distribution
	Number of generations	10	QTL distribution	Random
	Number of offspring per dam	1	Sex ratio	0.5
	Sex ratio	0.5	Mating system	Random
Validation population	Heritability	0.5, 0.3, 0.1		
Phenotypic variance		1		
QTL variance		0.5, 0.3, 0.1		

Gen: generation number; cM: Centimorgan; QTL: quantitative trait loci and SNP: single nucleotide polymorphism.

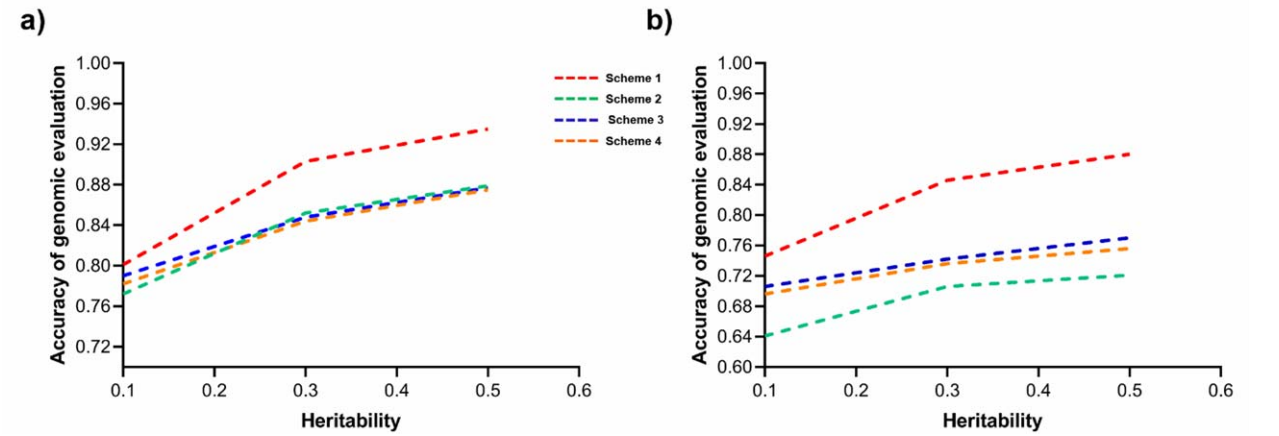

Figure 1 The accuracy of genomic evaluation in a) reference population and b) validation population at different levels of heritability of four selection schemes

Figure 2 illustrates that the level of genomic evaluation accuracy in both the reference and validation populations is higher in scheme 1 compared to other schemes. In the validation population, the lowest accuracy level is associated with scheme 2, where selection occurs in the reference population with an accuracy of 0.7, but no selection takes place in the validation population. Increasing the number of markers in both populations was accompanied by an increase in genomic evaluation accuracy, although this rise did not occur uniformly.

As shown in Figure 3, the results indicated that the accuracy level in scheme 1 was the highest, where no selection occurs in both populations. The validation population of scheme 2 exhibited the lowest accuracy level. An increase in the number of QTLs from 500 to 1000 results in a negligible change in the accuracy of genomic breeding values.

As shown in Table 2, the mean accuracy in the reference population ranged from 83.4% to 88.0%, while in the validation population, it varied between 69.0% and 82.4% for different selection schemes. Additionally, the correlation coefficient of the accuracy of genomic evaluation for the two populations in scheme 1 was higher compared to the other schemes (mean 78.1%). In terms of selection, the correlation coefficient decreased, but there was no significant difference among schemes 2, 3, and 4 in this regard. The difference among these three schemes was related to the status of the validation population. In scheme 2, the validation population was not subjected to selection, while in the other two schemes, selection occurred with 70% or 50% accuracy. The average number of common significant loci between the two populations is presented in Table 3. It was observed that the average number of common significant SNPs was higher in scheme 1 compared to the other schemes (approximately 24 loci *vs*. 14 loci).

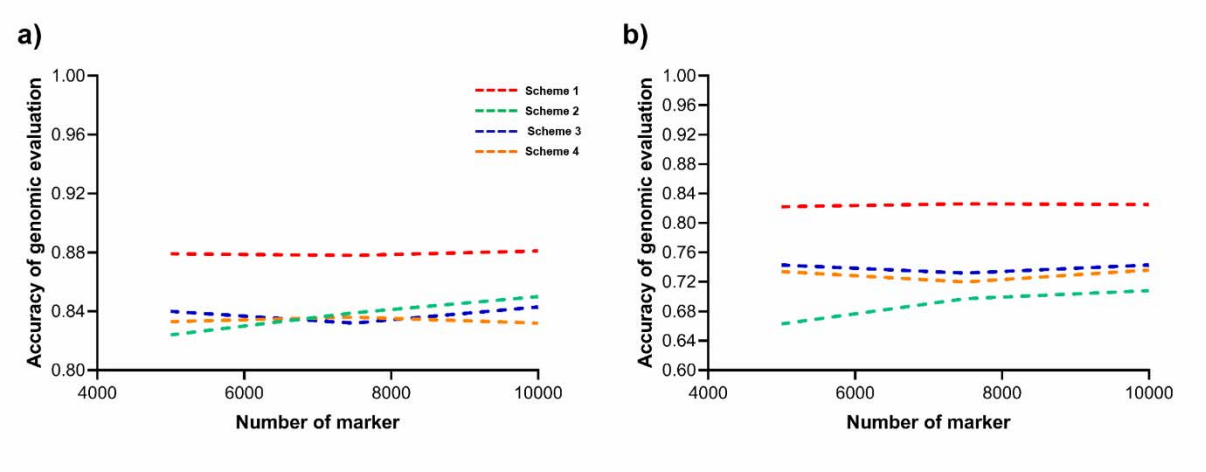

Figure 2 The accuracy of genomic evaluation in a) reference population and b) validation population at different levels of marker number for different selection schemes

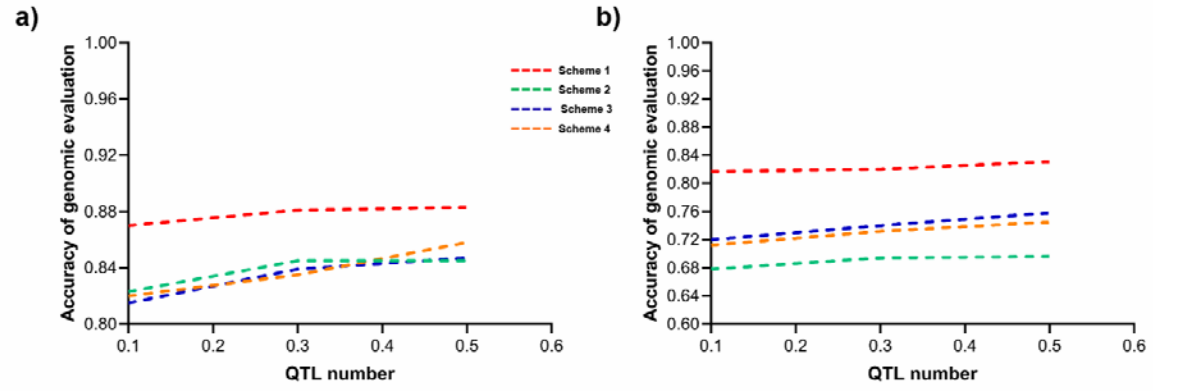

Figure 3 The accuracy of genomic evaluation in a) reference population and b) validation population at different numbers of QTLs for selection schemes

Table 2 Accuracy of genomic evaluation in the reference and validation populations in different schemes their correlation coefficient

Scheme	Mean ± SD		Correlation coefficient ± SD
	Reference	Validation	
1	0.880±0.059	0.824±0.060	0.781±0.043
2	0.848±0.048	0.690±0.000	0.652±0.093
3	0.834±0.047	0.730±0.040	0.610±0.092
4	0.838±0.043	0.739±0.039	0.615±0.060

SD: standard deviation.

The key distinction of scheme 1 from the others lies in the absence of selection in both the reference and validation populations. Selection and the consideration of specific phenotypes by breeders lead to increases in the frequency of certain variants and the elimination of others from the population.

The phenomenon known as a selective sweep leads to a reduction in heterozygosity and subsequently a decrease in the number of variants with significant effects. The higher accuracy of genomic evaluation in scheme 1 compared to other schemes may be associated with this phenomenon. As shown in Table 3, the highest correlation between the allelic frequency of significant gene loci common between two populations was related to scheme 1 (0.9), while the lowest correlation coefficient was related to scheme 2 (0.02). This indicates that when the validation population is not under selection, but the reference population undergoes selection, the minimum coordination in terms of the presence of significant gene loci occurs between the two populations.

The data presented in Table 4 shows that varying levels of heritability and marker numbers did not demonstrate a clear trend in the accuracy of genomic evaluation.

Based on the data obtained, the accuracy level in the validation population was found to be lower than that in the reference population.

This discrepancy is not influenced by the presence of selection or the heritability of the trait, as well as factors such as marker density and the number of QTLs affecting the trait.

Table 3 Mean and standard deviation of the number of common significant SNPs between the reference and validation populations in different selection schemes and their correlation coefficients of the frequency of these alleles in two populations

Scheme	Number of common significant SNPs	Correlation coefficient of the frequency of common significant SNPs
1	23.97±15.94	0.903±0.089
2	14.42±9.67	0.022±0.278
3	13.06±8.25	0.851±0.115
4	13.39±8.31	0.841±0.097

Table 4 Correlation coefficient of accuracy of genomic evaluation of reference and validation populations in Scheme 1 at different levels of heritability and number of markers

Marker	Heritability	Correlation coefficient
5000	0.1	-0.045
	0.3	-0.101
	0.5	-0.026
7500	0.1	0.09
	0.3	0.224
	0.5	-0.142
10000	0.1	0.392
	0.3	-0.023
	0.5	0.141

This result was expected because marker effects are estimated based on the phenotypic and genotypic data of the reference population (Meuwissen *et al.* 2013). In the reference population, there was no significant difference observed in the genomic evaluation accuracy between schemes 2, 3, and 4. This is because these schemes were associated with a selection accuracy of 0.7, resulting in results that were significantly different from scheme 1, where no selection occurred. However, in the validation population, selection accuracy depends on the selection status in the reference population and the degree of similarity between the reference and validation populations. Selection leads to changes in allele frequencies, either towards fixation or elimination (Charlesworth *et al.* 1993). Therefore, in scheme 2, significant gene loci in the two populations had completely different allelic frequencies. As a result, the accuracy in the validation population was lower compared to other schemes. Hence, marker effect estimates obtained from a reference population under selection may not be reliably applicable to a validation population that is not subjected to selection.

In the current study, the number of markers used in different scenarios was sufficient to reveal the impact of increasing the number of markers. It is possible that if 2500 or 1250 markers were also tested, we would observe a trend of increasing accuracy as the number of markers increases. One of the factors affecting the accuracy of genomic evaluation is the level of linkage disequilibrium between markers and QTLs (Hayes *et al.* 2009). The more markers per unit length of chromosome, the higher the expected level of linkage disequilibrium between markers and QTLs.

However, increasing the number of markers can lead to a linear correlation between their effects, which in turn can negatively affect genomic evaluation accuracy (Brito *et al.* 2011). It is expected that a higher number of markers along the chromosome results in greater linkage disequilibrium between markers and QTLs due to a denser marker distribution. Conversely, increasing the number of markers can lead to a linear correlation between their effects, which negatively impacts the accuracy of genomic evaluation.

According to our data, increasing the number of QTLs from 500 to 1000 resulted in a slight change in the accuracy of the genomic heritability value. Hayes *et al.* (2009) indicated that the accuracy of genomic evaluation depends on the distribution of QTLs influencing the trait, as well as the level of linkage disequilibrium between markers and QTLs, the number of individuals in the reference population, and the heritability of the trait. When there is a large number of QTLs with small effects contributing to trait variation, the number of individuals in the reference population should be increased to achieve a reliable level of accuracy. It can be concluded that if the number of individuals in the reference population remains constant, as in this study, with the increase in the number of QTLs, the accuracy level is expected to decrease slightly.

The correlation coefficient of allelic frequency of the significant marker loci shared between the reference and validation populations in scheme 2 was notably lower than in other schemes. Due to selection effects, it is expected that a divergence in the allelic frequency of gene loci will occur between populations that vary in terms of selection. This difference reaches its maximum level when no selective pressure is applied in the validation population. Therefore, the genomic selection accuracy in the validation population was at its lowest level in scheme 2.

CONCLUSION

The results of this study suggest that the accuracy of genomic evaluation is highest when there is no selection in the reference and validation populations, compared to scenarios where selection is present in either or both populations. The GWAS results indicate that this may be due to a similar number of significant SNPs in both populations. When selection occurs in the reference population, the number of these markers decreases, leading to decreased accuracy of genomic evaluation and a reduction in the correlation between accuracies in the two populations. Additionally, this correlation is more influenced by selection in the reference population rather than the validation population. The study also highlights the impact of trait heritability, marker density, and the number of QTLs on the accuracy of genomic breeding value estimation.

ACKNOWLEDGEMENT

Authors would like to extend their heartfelt gratitude to university of Guilan for providing academic environment and the essential resources that have been instrumental in the completion of this research.

REFERENCES

Bouquet A. and Juga J. (2013). Integrating genomic selection into dairy cattle breeding programs: A review. *Animal*. **7(5),** 705-713.

Brito F.V., Neto J.B., Sargolzaei M., Cobuci J.A. and Schenkel F.S. (2011). Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genet.* **12,** 1-10.

Charlesworth B., Morgan M. and Charlesworth D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*. **134(4),** 1289-1303.

De los Campos G. and Pérez-Rodríguez P. (2014). Bayesian generalized linear regression. R package version 1.0.4. Available at: http://CRAN.R-project.org/package=BGLR. Accessed Apr. 2016.

De Roos A., Hayes B.J., Spelman R. and Goddard M. (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey, and Angus cattle. *Genetics*. **179(3),** 1503-1512.

Hayes B.J., Bowman P.J., Chamberlain A.J. and Goddard M.E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci*. **92(2),** 433-443.

Mc Hugh N., Meuwissen T., Cromie A. and Sonesson A. (2011). Use of female information in dairy cattle genomic breeding programs. *J. Dairy Sci*. **94(8),** 4109-4118.

Meuwissen T.H., Hayes B.J. and Goddard M. (2013). Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci*. **1(1),** 221-237.

Meuwissen T.H., Hayes B.J. and Goddard M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157(4),** 1819-1829.

Pryce J., Bolormaa S., Chamberlain A., Bowman P., Savin K., Goddard M. and Hayes B.J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *J. Dairy Sci.* **93(7),** 3331-3345.

Sargolzaei M. and Schenkel F.S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*. **25(5),** 680-681.

Williams J. (2005). The use of marker-assisted selection in animal breeding and biotechnology. *Rev. Sci. Technol.* **24(1),** 379-391.