Journal of Genetics (2019)98:102
https://doi.org/10.1007/s12041-019-1149-3

© Indian Academy of Sciences

RESEARCH ARTICLE

Comparison of parametric, semiparametric and nonparametric methods in genomic evaluation

HAMID SAHEBALAM, MOHSEN GHOLIZADEH*, HASAN HAFEZIAN and AYOUB FARHADI

Faculty of Animal and Aquatic Science, Department of Animal Science, Sari Agricultural Sciences and Natural Resources University, P.O. Box -578, Sari, Iran
*For correspondence. E-mail: m.gholizadeh@sanru.ac.ir.

Received 14 March 2019; revised 20 July 2019; accepted 19 August 2019

Abstract. Access to dense panels of molecular markers has facilitated genomic selection in animal breeding. The purpose of this study was to compare the nonparametric (random forest and support vector machine), semiparametric reproducing kernel Hilbert spaces (RKHS), and parametric methods (ridge regression and Bayes A) in prediction of genomic breeding values for traits with different genetic architecture. The predictive performance of different methods was compared in different combinations of distribution of QTL effects (normal and uniform), two levels of QTL numbers (50 and 200), three levels of heritability (0.1, 0.3 and 0.5), and two levels of training set individuals (1000 and 2000). To do this, a genome containing four chromosomes each 100-cM long was simulated on which 500, 1000 and 2000 evenly spaced single-nucleotide markers were distributed. With an increase in heritability and the number of markers, all the methods showed an increase in prediction accuracy ($P < 0.05$). By increasing the number of QTLs from 50 to 200, we found a significant decrease in the prediction accuracy of breeding value in all methods ($P < 0.05$). Also, with the increase in the number of training set individuals, the prediction accuracy increased significantly in all statistical methods ($P < 0.05$). In all the various simulation scenarios, parametric methods showed higher prediction accuracy than semiparametric and nonparametric methods. This superior mean value of prediction accuracy for parametric methods was not statistically significant compared to the semiparametric method, but it was statistically significant compared to the nonparametric method. Bayes A had the highest accuracy of prediction among all the tested methods and, is therefore, recommended for genomic evaluation.

Keywords. accuracy; genomic selection; quantitative trait loci; single-nucleotide polymorphism.

Introduction

In recent years, the identification of thousands of single nucleotide polymorphisms (SNPs) dispersed at the genome level has resulted in predicted breeding values based on marker information (genomic breeding value) (Meuwissen *et al*. 2001). Genomic selection, in comparison with traditional methods (selection based on phenotypic records), leads to an increase in the genetic progress due to the reduction of generation interval and lack of an animal's need for a specific age (Schrooten *et al*. 2005). Genomic selection is a form of marker-assisted selection (MAS), in which all genetic markers that cover the entire genome are used simultaneously (Meuwissen *et al*. 2001; Goddard 2009). To this end, the number of markers should be such that each quantitative trait loci (QTL) is in linkage disequilibrium (LD) with at least one marker (Toosi *et al*. 2010). The accuracy of genomic estimated breeding value (GEBV) is influenced by heritability, marker density, minor allele frequency (MAF), and genetic architecture of target trait. (De los Campos *et al*. 2013). In genomic selection, first the genotype of the training set animals with phenotypic records is determined by a large number of markers and the effects of all markers are estimated simultaneously by statistical models. Then, the estimated marker effects are used to predict the genomic breeding value of individuals without phenotypic records in the validation set (Meuwissen *et al*. 2001).

The appropriate training set is highly effective in accurate predicting of the breeding values of young individuals without phenotypic records, since it plays an important role in estimating the marker effects. Factors such as the number of individuals, the reliability of the individuals' phenotypic information, the genetic relationships within the training set,

Published online: 06 November 2019

and the relationships between the individuals of the training set and the individuals of the validation set play a key role in the accuracy of prediction in the training set (Samuel *et al*. 2012).

For many effects, Bayes B considers the genetic variance to be zero in many analytical cycles, and therefore it is not introduced in the equations. Bayes A requires more computational power than Bayes B does (Meuwissen *et al*. 2001). According to the simulation studies, for traits with a limited number of large-effect QTL, differential shrinkage of estimates of effects and variable selection methods (e.g. Bayes A and Bayes B) have predicted superiority and yield higher accuracy than genomic best linear unbiased prediction (GBLUP). However, the differences between methods shown by simulation studies have not always been reported by empirical studies using real data analysis (De los Campos *et al*. 2013).

Ghafouri-Kesbi *et al*. (2017) compared three machine learning algorithms (support vector machines, boosting and random forests), as well as GBLUP to predict genomic breeding values. GBLUP had better predictive accuracy than machine learning methods in particular in the scenarios of normal and uniform distributions of QTL effects and higher number of QTL. In the scenarios of small number of QTL and gamma distribution of QTL effects, boosting surpassed other methods.

Some data do not follow a particular statistical distribution (e.g. normal distribution). For this reason, it is not possible to estimate marker effects using conventional statistical methods such as frequency-oriented methods (GBLUP and ridge regression best linear unbiased prediction) and Bayesian methods. As a result, we have to use nonparametric methods to estimate marker effects.

The purpose of this study was to compare the accuracy of the parametric methods (Bayesian ridge regression and Bayes A), semiparametric method (reproducing kernel Hilbert spaces) and nonparametric methods (random forest and support vector machine), in predicting genomic breeding values for traits with different genetic architecture in terms of marker density, number of QTLs, heritability and number of training set individuals (number of observations) using simulated data.

Materials and methods

Population simulation

Programming to create populations was done in the software environment R under the hypred package (Technow 2013). The base population including 100 individuals (50 males and 50 females) was simulated. This demographic structure was conducted for 50 generations by random mating (historical population) to create recombination and drift, and linkage disequilibrium between the marker and the QTL. In the historical population, assuming that both parents (over 50 generations of random mating) had produced two progeny, the effective population size was fixed along the base generations. Progeny chromosomal components were obtained from random sampling of each parent's paternal and maternal chromosomes. In the 51st generation, the population size was increased by 1000 and 2000 individuals known as the training set. The members of this population had both genotypic and phenotypic information. Generations 52, 53 and 54 were recognized as the validation set, for which only genomic data was simulated and their genomic breeding values were predicted.

Genome simulation

In this study, a genome consisting of four chromosomes each with a length of 100 cM was simulated. On each chromosome 500, 1000 and 2000 markers were considered at identical marker intervals throughout the genome. Based on the different simulation scenarios, 50 and 200 QTLs were randomly distributed on chromosomes. The markers and QTLs were considered as bi-allelic and with an initial allele frequency of 0.5. In the 51st generation, the substitute effect for each QTL was considered by using standard normal distribution (mean 0 and variance 1) and in three levels of heritability (0.1, 0.3 and 0.5). The whole genetic variance of the trait was covered by QTLs and the true breeding value of each individual was calculated from the following relation according to each individual's genotype from the total effect of QTLs:

$$\mathrm{TBV_i} = \sum_{j=1}^{n} x_{ij} b_j,$$

where $\mathrm{TBV_i}$ is the true breeding value of the individual *i*, *n* is the number of effective QTLs on the trait, x_{ij} is the QTL genotype at position j, and b_j is the additive effect of the *j* QTL.

The following equation was used to simulate the phenotype:

$$y_i = \mathrm{TBV_i} + e_i$$

where y_i is the phenotype of individual *i* and e_i is the residual effect.

LD estimation

LD value in the training set was measured by r^2 statistic (Hill and Robertson 1968):

$$r^2 = D^2/\mathrm{freq}(A_1) * \mathrm{freq}(A_2) * \mathrm{freq}(B_1) * \mathrm{freq}(B_2)$$

Freq (A_1) is the frequency of A_1 allele in the population likewise for other alleles in the population. D is the deviation of parental genotypes from the recombinant genotypes estimated as:

$$D = \text{freq}(A_1\_B_1) * \text{freq}(A_2\_B_2) - \text{freq}(A_1\_B_2) * \text{freq}(A_2\_B_1).$$

Evaluation methods-parametric methods

***Bayesian ridge regression (BRR)*:** In ridge regression (Hoerl and Kennard 1970), the distribution of marker effects is normal and are assumed as nonzero and partial.

Ridge regression is the same as the ordinary least squares, with the difference that, if the number of effects is more than the number of observations, it has no restrictions and also, when the markers are correlated, it has numerical stability and is calculated as the following:

$$\hat{\beta} = \left\{ \sum_i \left[y_i - \sum_j x_{ij}\beta j \right]^2 + \lambda \sum_{j \in s} \beta_j^2 \right\},$$

where, $\lambda \geq 0$ is a moderator for the controlling parameter to make balance between fitness (measured by the sum of error squares) and model complexity (which can be measurable by the sum of marker effect squares). The lambda is added to the diameter of the coefficient matrix and drives the estimates to zero. Although it stimulates the bias, it reduces the variance of estimates. If the lambda tends to infinity, β will equal zero. On the other hand, if the lambda is zero, the estimates of this method will be similar to the estimates of the least ordinary square method.

***Bayes A*:** In the method of Bayes A (Meuwissen *et al*. 2001), the prior assumption is that a large number of positions have minor effects and a small number of them have major effects and the conditional distribution considered for marker effects is *t* distribution. The prior distribution of the variance is a scale inverted chi-square distribution with a degree of freedom *v* and a parameter of scale *s*. The posterior distribution combines the former distribution information and data information together. Therefore, the posterior distribution will also be categorized as a scale inverted chi-square distribution.

Semiparametric method

***Reproducing kernel Hilbert space (RKHS) method*:** The RKHS method (Gianola *et al*. 2006) is a semiparametric method for genomic estimated breeding values in which the regression function is a linear combination of the basic function created by the *RK*. Thus, selection of *RK* is one of the central elements of model specification. The *RK* is a function that maps from pairs of points in the input space into the real line and must be positive semidefinite.

$$k(x_i, x_j) : \{(x_i, x_j) \rightarrow R\}$$

The K-kernel matrix inputs are as follows:

$$K(x_i, x_j) = \exp\{-h \times d(x_i, x_j)\},$$

$d(x_i, x_j)$ is the squared-Euclidean distance between i and j according to the genotype of their markers:

$$d(x_i, x_j) = (x_i - x_j)^2,$$

h is a bandwidth parameter that controls how the covariance (kernel) function velocity drops as the distance between pairs of vector genotypes increases. If the value of h is too small (for example, 0.001), it will create a very big kernel, and if the value of h is too big (for example, 50), it will create a very small kernel. This parameter plays an important role. In this research, a normal kernel was used. After optimization, the value of the parameter h was considered 0.1.

Nonparametric methods

***Random forest method*:** An RF regression was created using an accumulation of decision trees. Each decision tree uses a bootstrap sample of training data including genotypic and phenotypic information. The model is trained in the training set and is applied on the validation set. One of the *n* samples enters each split of each tree (mtry), and this sample of marker information is used to categorize animals in a way that the animals are classified for the selected marker according to their genotypic information. This is done in sequential splits, until we finally reach the nodes in which there is maximum uniformity (animals with phenotypic information accumulate with similar genotypes for different SNPs in a node). The RF prediction for the training set, $f_{rf}^{B}(x)$ is performed through the averaging of the B trees $\{T(x\ , \Psi_b)\}_1^B$ as follows (Hastie *et al*. 2009):

$$f_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x\ , \Psi_b),$$

where Ψ_b represents the *b*th tree in the RF. The most important parameters in the RF are the number of variables selected in each tree split (mtry), the number of trees (ntree), and the minimum size or the minimum number of observations in final nodes (nodesize) whose appropriate values must be defined prior to the analyses. For continuous data, the proposed value for the number of randomly sampled variables is equal to p/3 in each split (mtry) (p is the number of markers). In this study, the values of mtry were half, equal and twice the default value. After optimizing the parameters, the value of mtry, ntree and nodesize were obtained as 4000, 1000 and 5, respectively.

Support vector machine (SVM) method

The SVM method is a computer algorithm that learns through training information to classify observations. The

purpose of this method is to identify and distinguish complex patterns in the data and to categorize them. It is the best method to solve the linear separable two-class problems. If the data are linear separable, it will create a hyperplane with a maximum margin to separate the categories. SVM regression in corrective programmes can implement the relationship between marker genotype and phenotype with a linear or nonlinear function that employs samples from predictor spaces to feature spaces (Hastie *et al*. 2009). The statistical model is as follows:

$$f(x) = b + wx.$$

Where b is the constant effect and w is the unknown value vector.

The function f(x) is obtained by minimizing the function $\lambda \sum_{i=1}^{n} L(y_i - f(x_i)) + 1/2\|w\|^2$. L(.) denotes the loss function that measures the quality of the estimates. Lambda is the regulating parameter between model dispersion and complexity. Due to the high penalty applied, few errors in the classification of information will be acceptable. But if very large values of λ are used, since almost no errors are acceptable in the classification of information, there will be 'overfitting' and hence the generalization of the model will be reduced. Also, when the value of this parameter is small, the fine imposed will be small, and further mistakes will be acceptable in the classification of information. And when very small amounts of λ are used, 'underfitting' occurs and the model is wrongly trained and, as a result, classification of new data will be done with a high error rate. $\|w\|$ has an inverse relation with model complexity. By selecting w to minimize $\|w\|$, model complexity can be reduced. There are many loss functions that are used for SVM regression, such as squared loss, absolute loss, and ε-insensitive loss, which is as follows: (i) the squared loss function is $L(y - f(x)) = (y - f(x))^2$, which indicates that outliers have quadratic values that must be confronted with preregression analysis outliers. (ii) The absolute loss function is $L(y - f(x)) = |y - f(x)|$. This function evaluates linear loss through error size, which solves the problem of using the entire data with the outliers. (iii) The ε-insensitive loss function is as follows:

$$L(y - f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| < \varepsilon \\ |y - f(x)| - \varepsilon & \text{if } |y - f(x)| \geq \varepsilon \end{cases}.$$

Where ε determines the number of support vectors (SVs) used in the regression function. The increase of ε indicates fewer support vectors in the fitting. The ε-insensitive loss function ignores the existing errors in the model that are smaller than ε, and when the error is greater than ε, the loss function is $|y - f(x)| - \varepsilon$. Accordingly, the solving function is as follows:

$$\hat{f}(x) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b.$$

Where a_i and a_i* are positive weights given to each observation and are estimated from the data. The internal multiplication of the kernel K (x, xi) is a definite positive $n*n$ matrix. In this study, Epsilon regression and Gaussian kernel function were used. Also, the value of cost parameter (the fining parameter, λ) was considered 3 after optimization.

The accuracy of genomic estimated breeding values (GEBV) was obtained from the correlation between true breeding values and predicted breeding values. In this study, each scenario was repeated 10 times due to the use of a randomized model. The R package BGLR (Perez and De los Campos 2014) was used to run ridge regression, Bayes A and RKHS methods. To perform the RF method, the random Forest package (Liaw 2013) was used. To implement the SVM method, the R package e1071 (Meyer *et al*. 2013) was used. Also we used R to investigate the effect of factors affecting the accuracy of genomic breeding values. To compare different statistical methods, the Tukey's test was used at a significance level of 0.05.

Results and discussion

The value of r^2 as a measure of LD for marker density and different QTL numbers is in table 1. As shown, the r^2 values increases with marker density increased and its highest value was 0.21 for the marker density of 2000. In GWAS studies and genomic selection, the minimum value of r^2 between markers and QTL should be 0.2 for tracking the average effect (Hayes 2007). The results of variance analysis of prediction accuracy are presented in table 2. The main factors included marker density, number of QTLs, heritability and the size of training set, as well as the interactive effects among these factors. Among the effects, heritability, the size of training set, and statistical methods respectively had the greatest effect on the accuracy of prediction of genomic breeding values.

The prediction accuracy of five methods studied for four generations (the first generation is the training set and second to fourth generations are the validation set) and in different combinations of marker densities (500, 1000 and 2000), levels of heritabilities (0.1, 0.3 and 0.5), two levels of the size of training set (1000 and 2000) are shown in figures 1–5, respectively.

By increasing the generation interval between the training set and validation set, the accuracy of genomic breeding values decreased significantly (see figure 1) mainly due to the change in the marker or haplotype structure and the decrease of LD between markers and QTLs due to recombination (Hayes *et al*. 2009). As presented in figure 2, increasing the

Table 1. Values of r^2 for marker density and different QLT numbers.

N_SNP	500		1000		2000	
N_QTL	50	200	50	200	50	200
r^2 value	0.195	0.194	0.205	0.204	0.199	0.2068

Table 2. Variance analysis output for prediction accuracy.

Source of variation	Degree of freedom	Sum square	Mean square	*F* value	*P* value
Method	4	5.496	1.374	869.549	< 0.0001
N_SNP	2	0.154	0.077	48.591	< 0.0001
N_QTL	1	0.015	0.015	9.554	< 0.05
h^2	2	12.777	6.388	4042.81	< 0.0001
N_p	1	2.385	2.385	1509.219	< 0.0001
Method*N_SNP	8	0.014	0.002	1.109	0.354
Method*N_QTL	4	0.03	0.008	4.811	< 0.0001
Method*h^2	8	0.501	0.063	39.619	< 0.0001
Method*N_p	4	0.07	0.018	11.128	< 0.0001
N_SNP*N_QTL	2	0.012	0.006	3.762	< 0.05
N_SNP*h^2	4	0.100	0.025	15.81	< 0.0001
N_SNP*N_p	2	0.005	0.002	1.55	0.2116
N_QTL*h^2	2	0.013	0.006	4.023	< 0.05
N_QTL*N_p	1	0.000	0.000	0.151	0.6972
h^2*N_p	2	0.063	0.032	19.948	< 0.0001
Method*N_SNP*N_QTL	8	0.011	0.001	0.834	0.57239
Method*N_SNP*h^2	16	0.008	0.000	0.312	0.9956
Method*N_SNP*N_p	8	0.002	0.000	0.126	0.99819
Method*N_QTL*h^2	8	0.004	0.001	0.347	0.94727
Method*N_QTL*N_p	4	0.003	0.001	0.466	0.76108
Method*h^2*N_p	8	0.001	0.000	0.040	0.999975
N_SNP*N_QTL*h^2	4	0.040	0.010	6.297	< 0.0001
N_SNP*N_QTL*N_p	2	0.017	0.008	5.330	< 0.001
N_SNP*h^2*N_p	4	0.067	0.017	10.600	< 0.0001
N_QTL*h^2*N_p	2	0.004	0.002	1.392	0.248891
Method*N_SNP*N_QTL*h^2	16	0.011	0.001	0.432	0.974502
Method*N_SNP*N_QTL*N_p	8	0.009	0.001	0.677	0.712054
Method*N_SNP*h^2*N_p	16	0.007	0.000	0.288	0.997350
Method*N_QTL*h^2*N_p	8	0.006	0.001	0.453	0.888931
N_SNP*N_QTL*h^2*N_p	4	0.023	0.006	3.713	< 0.001
Method*N_SNP*N_QTL*h^2*N_p	16	0.005	0.000	0.212	0.999608

H^2, heritability; N_SNP, number of marker; N_QTL, number of QTL; N_p, size of reference population.

marker density resulted in increase in the predictive accuracy of genomic breeding value ($P < 0.05$). Parametric and semiparametric methods showed higher accuracy than nonparametric methods ($P < 0.05$). Among the three parametric and semiparametric methods, Bayes A showed the highest prediction accuracy which was not statistically significant ($P > 0.05$). Among nonparametric methods, SVM method showed higher accuracy than random forest method but it was not statistically significant ($P > 0.05$).

In a simulation study, the doubling in the number of markers resulted in an accuracy increase from 0.63 to 0.73 (Piyasation and Dekkers 2013).

Gianola *et al*. (2006) reported that in comparing RKHS and multiple linear regression (MLR), when the effective gene was additive, both methods showed the same accuracy (MLR), but when the effective gene was nonadditive (additive effect interaction) the parametric MLR was obviously superior to the RKHS method. In a simulation study, the accuracy of Bayes A and Bayes L were the same and higher than RKHS (Howard *et al*. 2014).

In all methods, by increasing the number of QTLs from 50 to 200, the accuracy of genomic breeding values was reduced, which is consistent with the results of other studies of this field (Daetwyler *et al*. 2010). In addition, Abdollahi-Arpanahi *et al*. (2013) simulated a trait controlled by 50, 100, 267 and 200 QTLs, and observed that by increasing the number of QTLs the accuracy of prediction decreased, this is due to the fact that by increasing the number of QTLs, due to the limited amount of genetic variance versus a large number of QTLs, the proportion of each QTL decreases in total genetic value, thereby, reducing the accuracy of genomic breeding values as well as the power of models in estimating the effects. Also, by increasing the number of QTLs, the number of markers should also increase so that the effects of all QTLs can be captured (Habier *et al*. 2009). An increase in the number of QTLs can increase the accuracy of genomic breeding values if the number of markers increases as QTLs increase.

The results of comparing predictive performance of different statistical methods for different levels of heritabilities (0.1, 0.3 and 0.5) are presented in figure 4. In all the methods, the accuracy of estimating breeding values increased significantly as the heritability increased. According to the studies reported, it is proposed that by

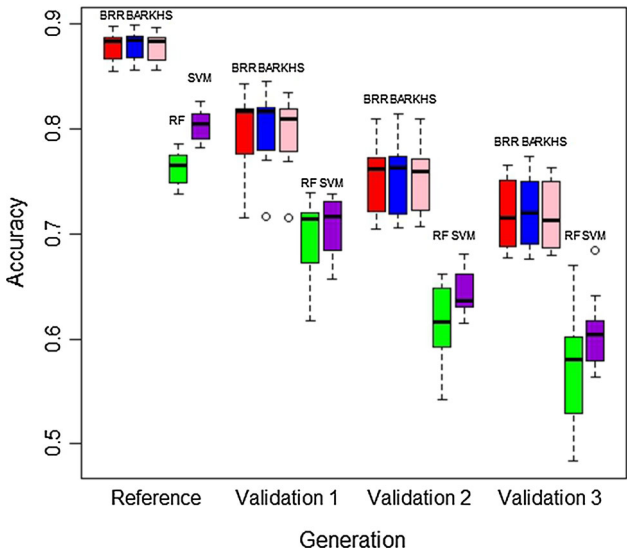

Figure 1. The accuracy of genomic breeding values prediction in different methods of ridge regression, Bayes A, RKHS, random forest and SVM over four generations.

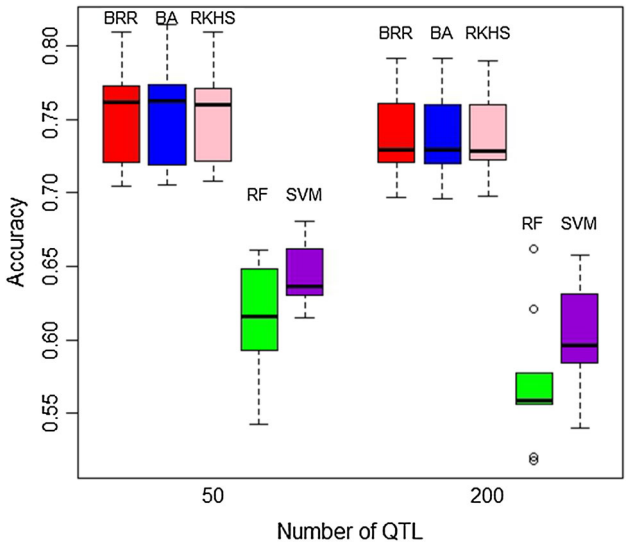

Figure 3. The accuracy of genomic breeding values prediction in different methods of ridge regression, Bayes A, RKHS, random forest and SVM in two different QLT number levels (50 and 200).

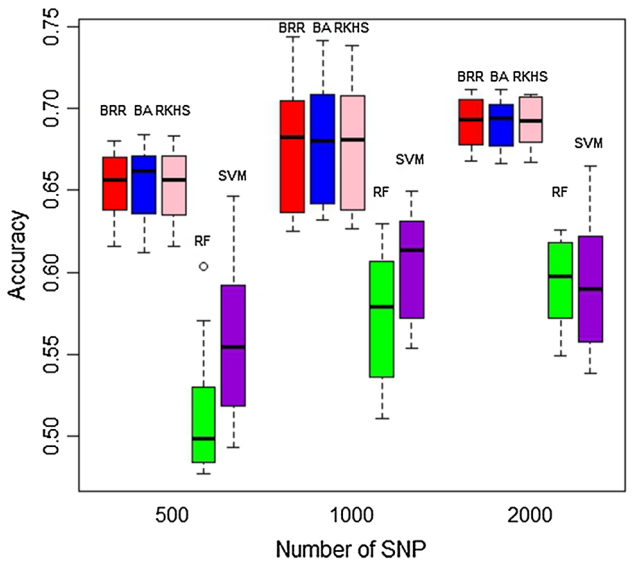

Figure 2. The accuracy of genomic breeding values prediction in different methods of ridge regression, Bayes A, RKHS, random forest and SVM in three different marker density levels.

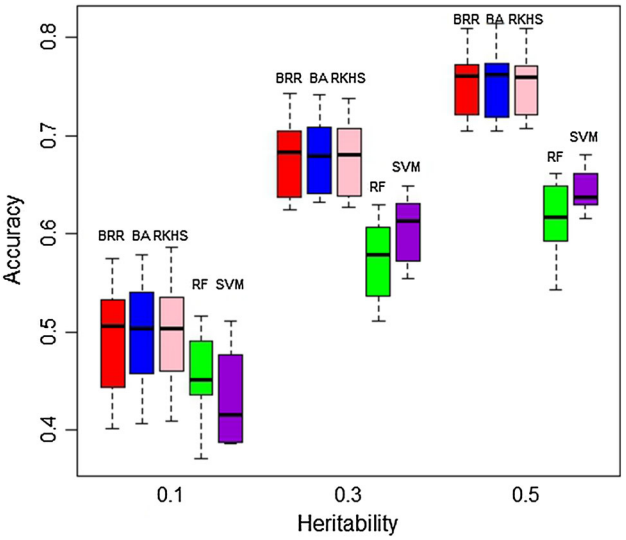

Figure 4. The accuracy of genomic breeding values prediction in different methods of ridge regression, Bayes A, RKHS, random forest and SVM in three different levels of heritabilities (0.1, 0.3, and 0.5).

increasing the heritability from 0.1 to 0.9, the predictive accuracy of genomic breeding values increased from 0.3 to 0.7 (Hayes *et al*. 2010). It has also been reported that by increasing heritability from 0.25 to 1, the accuracy of the prediction in terms of genetic architecture of the trait increased from 0.05 to about 1 (Combs and Bernardo 2012). The high value of heritability of a trait indicates that environmental factors have a less important role than genetic factors in the development of diversity. Reducing the role of environmental factors in the phenotypic value of the trait reduces the variance of model error and, consequently, increases the predictive accuracy of genomic breeding values (Meuwissen 2013). According to equation $r = \sqrt{N_p h^2 \left[N_p h^2 + M_e\right]^{-1}}$ (Deatwyler *et al*. 2013), predictive accuracy of genomic breeding value (r) has a direct relationship with the number of individuals with genotypic and phenotypic information in the training set (N_p) and trait heritability (h^2), as well as an inverse relationship with the number of independent chromosome segments (M_e). As a result, the maximum predictive accuracy of genomic breeding value for high-heritability traits and a high number of individuals in the training set would be expected (Hayes *et al*. 2009).

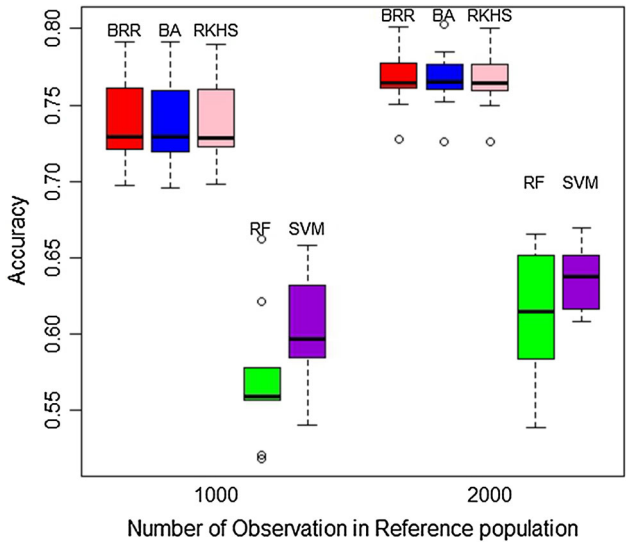

Figure 5. The accuracy of genomic breeding values prediction in different methods of ridge regression, Bayes A, RKHS, random forest and SVM for two different number of observations (1000 and 2000).

The results of comparing the predictive ability of different statistical methods in two levels of the size of training set (1000 and 2000) are presented in figure 5. By increasing the size of the training set from 1000 to 2000, the increase in accuracy was evident in all methods. As a result, there should be a direct relationship between the number of observations and the predictive accuracy. It has been indicated that if the number of individuals in the training set increases from 500 to 1000 and 2200, the estimation accuracy of breeding values in Bayes B will be increased from 0.708 to 0.787 and then to 0.848 (Meuwissen *et al*. 2001). It was also reported that in the heritability level of 0.2, with an increase in the number of animals in the training set from 1151 to 3576, the predictive accuracy of breeding values linearly increased from 0.35 to 0.53 (VanRaden *et al*. 2009). By increasing the size of the training set from 200 to 1600 bulls, the predictive accuracy of genomic breeding values increased from 0.3 to 0.6 (Hayes *et al*. 2010). Genomic studies that use real data are subject to biases such as genotypic and sampling errors. Simulation studies lack these biases, and thus, these differences can lead to a difference in the results of simulation studies compared to real studies.

According to the results, the predictive accuracy of the breeding values in genomic selection depends on heritability, marker density, QTL number, number of training individuals (number of observations), and statistical models used. In models with only gene additive effect, nonparametric methods such as random forest and SVM showed lower accuracy than parametric and semiparametric methods such as RKHS ($P < 0.05$). Also, parametric methods showed higher accuracy than semiparametric and this superior predictive accuracy was not statistically significant ($P > 0.05$). To succeed in genomic evaluation programmes, markers should be at an acceptable level of LD with QTL so that the marker can express QTL effect efficiently in the population. The accuracy of the estimates is directly related to the heritability of the trait, because if the heritability of the trait decreases, the ratio of environmental variance (residual) to genetic variance increases. As a result, the distributed environmental variance among all the recorded and genotype-determined animals increases; thus, the accuracy of predictions decreases. Although increasing the size of the training set increases the cost of genotype determination, it leads to an increase in the accuracy of the estimation of allelic effects and, as a result, increases genetic enhancement. Comparison of these methods for nonadditive models under different simulations as well as real data should is recommended.

References

Abdollahi-Arpanahi R., Pakdel A., Nejati-Javaremi A. and Moradi Shahre Babak M. 2013 Comparison of different methods of genomic evaluation in traits with different genetic architecture. *J. Anim. Prod.* **15**, 65–77 (in Persian with English abstract).

Combs E. and Bernardo R. 2012 Accuracy of genome wide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* **6**, 1–7.

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

Daetwyler H. D., Calus M. P. L., Pong-wong R., de los Campos G. and Hickey J. M. 2013 Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* **193**, 347–365.

De los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D. and Calus M. P. 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345.

Ghafouri-Kesbi F., Rahimi-Mianji G., Honarvar M. and Nejati-Javaremi A. 2017 Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Anim. Prod. Sci*. **57**, 229–236.

Gianola D., Fernando R. L. and Stella A. 2006 Genomic-assisted prediction of genetic value with semiparametric procudures. *Genetics* **173**, 1761–1776.

Goddard M. 2009 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetics* **136**, 245–257.

Habier D., Fernando R. L. and Dekkers J. C. M. 2009 Genomic selection using low-density marker panels. *Genetics* **182**, 343–353.

Hastie T. J., Tibshirani R. and Friedman J. 2009 *The elements of statistical learning*, 2nd edition. Springer-Verlag, New York.

Hayes B. 2007 QTL mapping, MAS, and genomic selection. A short-course, Animal Breeding and Genetics Department of Animal Science, Iowa State University **1**, 3–4.

Hayes B., Bowman P., Chamberlain A., Verbyla K. and Goddard, M. 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol*. **41**, 51.

Hayes B. J., Daetwyler H. D., Bowman P., Moser G., Tier B., Crump R. *et al*. 2010 Accuracy of genomic selection: comparing theory and results. *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18,** 34–37.

Hill W. and Robertson A. 1968 Linkage disequilibrium in finite populations. *Theor. Appl. Genet*. **38**, 226–231.

Hoerl A. E. and Kennard R. W. 1970 Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Howard R., Carriquiry A. L. and Beavis W. D. 2014 Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)* **4**, 1027–1046.

Liaw A. 2013 Breiman and Cutler's random forests for classification and regression. Available 403 at: http://cran.r-project.org/web/packages/randomForest/index.html.

Meuwissen T. H. 2013 The accuracy of genomic selection. Available at: http://www.umb.no/statisk/husdyrforsoksmoter/2013/1_1.pdf.

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

Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch K. 2013 Misc functions of the department of statistics (e1071), TU Wien. Available at: http://cran.rproject.org/web/packages/e1071/index.html.

Perez P. and De los Campos G. 2014 Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**, 483–495.

Piyasation N. and Dekkers J. 2013 Accuracy of genomic prediction when accounting for population structure and polygenic effects. *Anim. Industry Rep.* **659**, 68.

Samuel A. C., Hickey J. M., Daetwyler H. D. and van der Werf J. H. J. 2012 The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol*. **44**, 4–13.

Schrooten C., Bovenhuis H., Van Arendonk J. A. M. and Bijma P. 2005 Genetic progress in multistage dairy cattle breeding schemes using genetic markers. *J. Dairy Sci*. **88**, 1569–1581.

Technow F. 2013 hypred: Simulation of genomic data in applied genetics. Available at: 433 http://cran.r-project.org/web/packages/hypred/index.html.

Toosi A., Fernando R., Dekkers J. and Quaas R. 2010 Genomic selection inadmixed and crossbred populations. *J. Anim. Sci*. **88**, 32.

VanRaden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F. *et al*. 2009 Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci*. **92**, 16–24.

Corresponding editor: R. S. SANGWAN