Using Machine Learning To Improve the Accuracy of Genomic Prediction on Reproduction Traits in Pigs

Xue Wang
China Agricultural University
Shaolei Shi
China Agricultural University
Guijiang Wang
Hebei Province Animal Husbandry and Improved Breeds Work station
Wenxue Luo
Hebei Province Animal Husbandry and Improved Breeds Work Station
Xia Wei
Zhangjiakou Dahao Heshan New Agricultural Development Co., Ltd
Ao Qiu
China Agricultural University
Fei Luo
Hebei Province Animal Husbandry and Improved Breeds Work Station
Xiangdong Ding (xding@cau.edu.cn)
China Agricultural University

Research

Keywords: machine learning, genomic prediction, prediction accuracy, pig

Posted Date: November 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1083849/v1

Using machine learning to improve the accuracy of genomic prediction on reproduction traits in pigs

Xue Wang[a], Shaolei Shi[a], Guijiang Wang[b], Wenxue Luo[b], Xia Wei[c], Ao Qiu[a], Fei Luo[b], Xiangdong Ding[a*]

a Key Laboratory of Animal Genetics and Breeding of Ministry of Agriculture and Rural Affairs, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China

b Hebei Province Animal Husbandry and Improved Breeds Work Station, Shijiazhuang, Hebei, China

c Zhangjiakou Dahao Heshan New Agricultural Development Co., Ltd, Zhangjiakou, Hebei, China

* Corresponding author.

E-mail addresses:

xwangchnm@163.com

BS20183040372@cau.edu.cn

13933077085@126.com

13832350632@163.com

weixia16888@163.com

956482319@qq.com

43068883@qq.com

xding@cau.edu.cn

Abstract

Background: Recently, machine learning (ML) is becoming attractive in genomic prediction, while its superiority in genomic prediction and the choosing of optimal ML methods are needed investigation.

Results: In this study, 2566 Chinese Yorkshire pigs with reproduction traits records were used, they were genotyped with GenoBaits Porcine SNP 50K and PorcineSNP50 panel. Four ML methods, including support vector regression (SVR), kernel ridge regression (KRR), random forest (RF) and Adaboost.R2 were implemented. Through 20 replicates of five-fold cross-validation, the genomic prediction abilities of ML methods were explored. Compared with genomic BLUP(GBLUP), single-step GBLUP (ssGBLUP) and Bayesian method BayesHE, our results indicated that ML methods significantly outperformed. The prediction accuracy of ML methods was improved by 19.3%, 15.0% and 20.8% on average over GBLUP, ssGBLUP and BayesHE, ranging from 8.9% to 24.0%, 7.6% to 17.5% and 11.1% to 24.6%, respectively. In addition, ML methods yielded smaller mean squared error (MSE) and mean absolute error (MAE) in all scenarios. ssGBLUP yielded improvement of 3.7% on average compared to GBLUP, and the performance of BayesHE was close to GBLUP. Among four ML methods, SVR and KRR had the most robust prediction abilities, which yielded higher accuracies, lower bias, lower MSE and MAE, and comparable computing efficiency as GBLUP. RF demonstrated the lowest prediction ability and computational efficiency among ML methods.

Conclusion: Our findings demonstrated that ML methods are more efficient than

traditional genomic selection methods, and it could be new options for genomic prediction.

Key words: machine learning, genomic prediction, prediction accuracy, pig

Background

Genomic selection (GS) has been widely recognized and successfully implemented in animal and plant breeding programs [1-3]. It is reported that the breeding costs of dairy cattle using GS were 92% lower than that of tradition progeny testing [4]. At present, the genetic gain rate of the annual yield traits of US Holstein dairy cattle has increased from around 50% to 100%[5]. The accuracy of GS depends on methods of genomic breeding values estimation (GEBV), reference population size, marker density, and heritability, *etc*. Currently, parametric methods are most commonly used for livestock and poultry genomic selection, mainly based on either the genomic covariance between genotyped individuals *e.g.* genomic BLUP (GBLUP)[6] or single-step GBLUP (ssGBLUP)[7, 8]) or Bayesian regression models[9, 10], with differences mainly depends on the prior distribution of marker effects. Nevertheless, these linear models usually only take into account the additive inheritance and ignore the complex non-linear relationships that may exist between markers and phenotypes (*e.g.* epistasis, dominance, genotype-by-environment interactions). In addition, parametric methods usually provide limited flexibility for handling non-linear effects in high-dimensional genomic data, resulting in huge computational demands [11], while considering nonlinearity may enhance the predictive ability of complex traits [12]. Therefore, new strategy should be explored to

more accurately estimate the genomic breeding values.

Driven by applications in intelligent robots, self-driving cars, automatic translation, face recognition, artificial intelligence games and medical services, machine learning (ML) has gained considerable attention in the past decade. Some characteristics of the ML methods make it potentially attractive to deal with high-order non-linear relationships in high-dimensional genomic data, *e.g.* allowing the number of variables larger than the sample size [13], capable of capturing the hidden relationship between genotype and phenotype in an adaptive manner, and imposing little or no specific distribution assumptions about the predictor variables as GBLUP and Bayesian methods [14, 15] .

Studies have shown that random forest (RF), support vector regression (SVR), kernel ridge regression (KRR) and other machine learning methods gained advantage over GBLUP and Bayes B, *etc.* [16-18]. Ornella et al. compared the performance of support vector regression, random forest regression, Reproducing Kernel Hilbert space (RKHS), ridge regression, and Bayesian Lasso in genomic prediction, and found that RKHS and random forest regression were the best [19]. González-Camacho et al. reported the support vector machine (SVM) with linear kernel performed the best in comparison with other ML methods and linear models in the genomic prediction of the rust resistance of wheat [18]. Additionally, ML algorithms have also been widely used in the fields of gene screening, genotype imputation, and protein structure and function prediction, *etc.* [20-23], demonstrating its superiority as well. However, one challenge for the ML is choosing the optimum ML method as a series of ML algorithms have been

proposed and each has its own characteristics and shows different prediction abilities in different datasets and traits.

Therefore, the objective of this study was to assess the performance of machine learning methods in genomic prediction through the comparison with existing prevail GBLUP and Bayesian methods, and on the other hand, the efficiency of different ML methods were compared as well in order to explore the ideal ML algorithm for genomic prediction.

Materials and Methods

Ethics Statement

The whole procedure for blood sample collection was carried out in strict accordance with the protocol approved by the Animal Care and Use Committee of China Agricultural University (Permit Number: DK996).

Population and Phenotypes

A Yorkshire pig population from DHHS, a breeding farm in Hebei province, China, was studied. A total 2566 animals born between 2016 and 2020 were sampled and 4274 reproductive records of total number of piglets born (TNB) and number of piglets born alive (NBA) were available, and 3893 animals were traced back to construct pedigree relationship matrix (A matrix). A single-trait repeatability model was used to estimate heritabilities. The fixed effects included herd-year-season, and the random effects included additive genetic effects, random residuals, and permanent effects. The information of the animals, phenotypes and genetic components, as well as the estimated heritabilities were listed in Table 1. The estimated heritabilities of TNB and

NBA were both 0.12.

Derivation of corrected phenotypes

In order to avoid double counting of parental information, the corrected phenotypes (y_c) derived from the estimated breeding values (EBV) were used as response variable in genomic prediction. The pedigree-based BLUP and single-trait repeatability model were performed to estimate the breeding values for each trait separately.

$$y = Xb + Z_a a + Z_{pe} pe + e, \qquad (1)$$

where y is the vector of phenotypic values; b is the vector of fixed effects including herd-year-season; a represent additive genetic effects, following a norm distribution $N(0, A\sigma^2_a)$, where A is the pedigree-based relationship matrix, σ^2_a is the additive genetic variance. pe is permanent environment effects with norm distribution $N(0, I\sigma^2_{pe})$, where σ^2_{pe} is permanent environment variance. e is the vector of random error, following a norm $N(0, I\sigma^2_e)$, where σ^2_e represents residual variance. X, Z_a, and Z_{pe} are incidence matrices linked b, a and pe to y. A total of 3893 individuals were traced to construct A matrix. Their EBVs were calculated using the DMUAI procedure of the DMU software [24] . The y_c were calculated as EBV plus the average estimated residuals for multiple parties of a sow following Guo et al. [25] .

Genotype data and imputation

Two kinds of 50K SNP panels, PorcineSNP50 BeadChip (Illumina, CA, USA) and GenoBaits Porcine SNP 50K (Molbreeding, China) were used for the genotyping. A total of 1189 sows were genotyped with PorcineSNP50 BeadChip, which includes 50,697 SNPs across the genome, and 1978 individuals were genotyped using GenoBaits

Porcine SNP 50K with 52,000 SNPs. There are 30,998 common SNPs between these two SNP panels, and 601 individuals were genotyped with both SNP panels and, 2566 genotyped individuals were therefore finally used for further analysis including 1189 animals with PorcineSNP50 BeadChip and 1377 pigs with GenoBaits Porcine SNP 50K. The animals genotyped with GenoBaits Porcine SNP 50K were imputed to PorcineSNP50 BeadChip using Beagle 5.0 [26]. The reference population size for genotype imputation was 3720. Imputation accuracy was assessed by the dosage R-squared measure (DR2), which is the estimated squared correlation between the estimated allele dose and the true allele dose. The genotype correlation (COR) and the genotype concordance rate (CR) were also calculated based on the 601 overlap animals to evaluate the imputation accuracy. After imputation, the quality control on genotype were carried out using PLINK software [27], SNPs with minor allele frequency (MAF) lower than 0.01 and call rate lower than 0.90 were removed, and individuals with call rate lower than 0.90 were excluded. Finally, all animals and 44,922 SNPs on autosomes were remained for further analysis.

Statistical models

GBLUP, ssGBLUP, Bayesian Horseshoe (BayesHE) and four ML regression methods, support vector regression (SVR), Kernel ridge regression (KRR), Random forest (RF), and Adaboost.R2 were used to predict GEBV. For ssGBLUP, in order to prevent the problem that singular matrix cannot be inverted, $G_w = (1-w)G_a + wA_{22}$, and w was equal to 0.05 [28]. BayesHE was developed by Shi. et al [29], it is based on Global-local priors to increase the flexibility and adaptability of the Bayesian model. In this study, the first

form of BayesHE (BayesHE1) was used [29], and the Markov chain Monte Carlo (MCMC) chain was run for 50,000 cycles, with the first 20,000 cycles being discarded as burn-in and every 50 sample of the remaining 30,000 iterations were saved to infer posterior statistics. In-house scripts written in Fortran 95 were used for BayesHE analyses, and the DMUAI procedure implemented in DMU software was used for GBLUP and ssGBLUP analyses. Meanwhile, the four ML regression methods are introduced as follows.

Support vector regression

Support vector machine (SVM) was proposed by Vapnik [30] for binary classification. SVR is the application of SVM in regression for dealing with quantitative responses, which uses a linear or non-linear kernel function to map the input space (the marker dataset) to a higher dimensional feature space, and performed modeling and prediction on the feature space [31]. In other words, we can build a linear model in the feature space to deal with regression problems. The model formulation of SVR can be expressed as:

$$f(x) = \beta_0 + h(x)^T\beta, \qquad (2)$$

in which $h(x)^T\beta$ is the kernel function, β is the vector of weights, and β_0 is the bias. Generally, the formalized SVR is given by minimizing the following restricted loss function:

$$\min_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} V\left(y_i - f(x_i)\right), \qquad (3)$$

in which

$$V_\varepsilon(r) = \begin{cases} 0, & if|r| < \varepsilon \\ |r| - \varepsilon, & otherwise \end{cases}. \qquad (4)$$

$V_\varepsilon(r)$ is the ε-insensitive loss and C ("cost parameter") is the regularization constant

that controls the trade-off between prediction error and model complexity. y is a quantitative response and $\|\cdot\|$ is the norm in the Hilbert space. After optimization, the final form of SVR can be written as:

$$f(x) = \sum_{i=1}^{m}(\hat{a}_i - a_i)k(x, x_i), \qquad (5)$$

in which $k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. In this research, grid search was used to find the best kernel function and the best hyper-parameters of C and gamma. An internal five-fold cross validation (5-fold CV) strategy was performed to adjust the hyper-parameters when performing a grid search.

Kernel ridge regression

Kernel ridge regression (KRR) is a non-linear regression method, which can effectively discover the non-linear structure of the data[32]. KRR uses a non-linear kernel function to map the data to a higher dimensional kernel space, and then builds a ridge regression model to make the data linearly separable in this kernel space. The linear function in the kernel space is selected according to the mean squared error loss of ridge regularization [32]. The final KRR prediction model can be written as:

$$y(x_i) = k'(K + \lambda I)^{-1}\hat{y}, \qquad (6)$$

where λ is the regularization constant; K is the Gram matrix with entries $K_{ij} = K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)^T$, thus, for n training samples, the obtained kernel matrix is:

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \Lambda & K(x_1, x_n) \\ K(x_2, x_1) & K(x_2, x_2) & \Lambda & K(x_2, x_n) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ K(x_n, x_1) & K(x_n, x_2) & \Lambda & K(x_n, x_n) \end{bmatrix}_{n\times n}. \qquad (7)$$

I is the identity matrix; $k' = K(x_i, x_j)$ with $j = 1,2,3,\ldots,n$, n is the number of training samples, and x_i is the test sample. In the expanded form,

$$k' = \begin{bmatrix} K(x_i, x_1) \\ K(x_i, x_2) \\ \mathrm{M} \\ K(x_i, x_n) \end{bmatrix}. \qquad (8)$$

The grid search was used to find the most suitable kernel function and λ in this study, and an internal 5-fold CV strategy was used for tuning the hyper-parameters.

Random forest

Random forest (RF) is a machine learning method that uses voting or the average of multiple decision trees to determine the classification or predicted values of new instances [33]. Random forest is essentially a collection of decision trees, and each decision tree is slightly different from other trees [34]. Random forest reduces the risk of overfitting by averaging the prediction results of many decision trees [18]. Random forest regression can be written in the following form:

$$y = \frac{1}{M}\sum_{m=1}^{M} t_m\left(\psi_m(y{:}X)\right), \qquad (9)$$

in which y is the predicted value of random forest regression, $t_m\left(\psi_m(y{:}X)\right)$is an individual regression tree, and M is the number of decision trees in the forest. The prediction is obtained by passing down the predictor variables in the flowchart of each tree, and the corresponding estimated value at the terminal node is used as the predicted value. Finally, the predictions of each tree in RF are averaged to calculate the final prediction of unobserved data. The grid search was used to find the most suitable hyper-parameter M and the maximum depth of the tree, and the inner 5-fold CV was performed to tune the hyper-parameters.

Adaboost.R2

Adaboost.R2 [35] is an ad hoc modification of Adaboost.R and is an extension of

Adaboost.M2 to deal with regression problems, which repeatedly uses a regression tree as a weak learner followed by increasing the weights of incorrectly predicted samples and decreasing the weights of correctly predicted samples. It builds a "committee" by integrating multiple weak learners, making its prediction effect better than those of weak learners [36]. Adaboost.R2 regression model can be written as:

$$y = inf\left[y \in Y: \sum_{t:f_t(x)\leq y} log\frac{1}{\varepsilon_t} \geq \frac{1}{2}\sum_t log\frac{1}{\varepsilon_t}\right], \qquad (10)$$

in which y is the predicted GEBV, $f_t(x)$ is predicted value of the t-th weak learner, and ε_t is the error rate of $f_t(x)$, $\varepsilon_t = \bar{L}_t / (1 - \bar{L}_t)$ and the average loss $\bar{L}_t = \sum_{i=1}^{m} L_t(i)D_t(i)$, in which $L_t(i)$ is the error between the actual observation value and the predicted value of the i-th predicted individual, and $D_t(i)$ is the weight distribution of $f_t(x)$. After $f_t(x)$ is trained, the weight distribution $D_t(i)$ will become $D_{t+1}(i)$,

$$D_{t+1}(i) = \frac{D_t(i)\beta_t^{(1-L_t(i))}}{Z_t}, \qquad (11)$$

in which Z_t is a normalization factor chosen such that $D_{t+1}(i)$ will be a distribution. In current study, SVR and KRR were respectively used as weak learners of Adaboost.R2.

For these four ML methods, the vectors of genotypes (coded as 0, 1, 2) were the input independent variables and y_c were used as response variables, and Sklearn package for Python (V0.22) was used for genomic prediction.

Meanwhile, the optimal hyper-parameters for SVR, KRR, RF and Adaboost.R2 according to the grid search were shown in Table S1.

Accuracy of genomic prediction

Five-fold cross validation was used to estimate the accuracies of genomic prediction,

in which 2566 individuals were randomly split into five groups with 513 individuals each. For each cross validation, four of the five groups were defined as reference population, and the left one was treated as the validation population. The genotyped reference and validation sets in each replicate of 5-fold CV were same for all methods, and it should be noted that non-genotyped individuals were added in the reference population in ssGBLUP. For all methods, the accuracy of genomic prediction was calculated as the Pearson correlation between the GEBVs and corrected phenotypes y_c in validation population. In addition, the prediction unbiasedness was also calculated as the regression of y_c on the GEBVs of validation population. The 5-fold CV scheme was repeated 20 times, and the overall prediction accuracy and unbiasedness was the average of 20 replicates. The Hotelling-Williams Test [37] was performed to compare the prediction accuracy of different methods.

Meanwhile, prediction ability metrics *e.g.* mean squared error (MSE) and mean absolute error (MAE) were also used to evaluate the performance of regression models in the present study. MSE can take both prediction accuracy and bias into account [38], and the smaller the value of MSE, the better the accuracy of the model to describe the experimental data is. MAE can better reflect the actual situation of the predicted value error. Their formulas can be written as follows.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(f_i - y_i)^2, \text{ and}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|f_i - y_i|, \qquad (12)$$

where m represents the number of animals in each cross-validation test fold of the 5-fold CV, f is the vector of predicted values (GEBVs) and y is the vector of observed

values (y_c). The final MSE and MAE were the average of 20 replicates.

Results

Genotype imputation accuracy

Figure 1 illustrates the accuracy of imputing GenoBaits Porcine SNP 50K to PorcineSNP50 BeadChip across minor allele frequency (MAF) intervals and chromosomes. DR2, CR and COR were not sensitive to MAF except that COR was lower when the MAF was less than 0.05 and in the range of 0.45 to 0.5 (Figure 1a). DR2, CR and COR on each chromosome were 0.978~0.988, 0.984~0.988 and 0.957~0.972, respectively, and no significant differences were observed in DR2, CR and COR between chromosomes (Figure 1b). In the same scenarios, the values of COR were smaller than those of DR2, CR. The averaged DR2, CR and COR across all variants were 0.984, 0.985 and 0.964, respectively, indicating the imputation is enough accurate to analysis two SNP panel together.

Accuracy of genomic prediction

Comparison of ML methods with (ss)GBLUP and BayesHE

Table 2 shows the prediction accuracies and unbiasedness of machine learning methods, (ss)GBLUP and BayesHE on traits of TNB and NBA. The accuracies of ML methods were significantly higher than those of (ss)GBLUP and BayesHE. The improvements of ML methods over GBLUP, ssGBLUP and BayesHE were 19.3%, 15.0% and 20.8% on average, ranging from 8.9% to 24.0%, 7.6% to 17.5% and 11.1% to 24.6%, respectively. For trait TNB, compared with GBLUP, the average accuracy of all ML methods in this study has been improved, support vector regression (SVR) gained

284 improvement of 19.0% as same as Kernel ridge regression (KRR), Adaboost.R2 based
285 on SVR and KRR obtained the improvement of 18.1% and 17.7%, respectively, while
286 random forest (RF) yielded the lowest improvement of 8.9% advantage over GBLUP.
287 The similar advantage of ML were also over ssGBLUP, the improvements of SVR,
288 KRR, RF, Adaboost.R2_SVR and Adaboost.R2_KRR were 17.5%, 17.5%, 7.6%, 16.7%
289 and 16.3%, respectively. ML methods gained the largest advantage over BayeHE, the
290 accuracy from SVR, KRR, RF, Adaboost.R2_SVR and Adaboost.R2_KRR were
291 respectively improved by 21.4%, 21.4%, 11.1%, 20.6% and 20.2% compared with
292 BayeHE. For trait NBA, although ML methods still performed better than GBLUP,
293 ssGBLUP and BayesHE, Adaboost.R2_KRR gained the largest improvement in all
294 comparisons, and KRR obtained the second largest improvement. SVR and
295 Adaboost.R2 based on SVR yielded same improvements on GBLUP, ssGBLUP and
296 BayesHE. RF still gained the lowest improvement compared with other ML methods.
297 Meanwhile, GBLUP, ssGBLUP and BayesHE had similar performance, and no
298 statistical differences of prediction accuracy were found among them. Nevertheless,
299 ssGBLUP produced average improvement of 3.7% compared with GBLUP (1.2% for
300 TNB; 6.3% for NBA), while less bias was observed by GBLUP in all scenarios.
301 BayesHE yielded similar accuracy with GBLUP (0.243 and 0.248 for TNB; 0.207 and
302 0.208 for NBA), but the unbiasedness of BayesHE was much closer to 1 (1.015 for
303 TNB; 1.009 for NBA).
304 On the other hand, mean squared error (MSE) and mean absolute error (MAE) were
305 also used to assess the performance of different methods. As shown in Table 3, ML

methods were generally superior to GBLUP, ssGBLUP and BayesHE in terms of MSE and MAE. For two reproduction traits TNB and NBA, all ML methods yielded lower MSE and MAE than GBLUP, ssGBLUP and BayesHE. The performance of GBLUP, ssGBLUP and BayesHE was very close, and ssGBLUP produced a bit lower MSE (5.26 for TNB; 3.95 for NBA) and MAE (1.748 for TNB; 1.532 for NBA) among these three methods, while they were still higher than those obtained from RF, which performed the worst among four ML methods, and generated 5.212 and 3.901 of MSE and 1.747 and 1.527 of MAE on TNB and NBA, respectively. Among ML models, the performance of SVR and KRR was the best, and they yielded the smallest MSE and MAE in all scenarios.

Comparison between ML methods

Table 2 and 3 indicates that ML methods performed better than GBLUP, ssGBLUP and BayesHE. They also show RF had the lowest accuracy even though no significant differences were observed among the ML methods in this study. The accuracies of SVR, KRR, Adaboost.R2_SVR and Adaboost.R2_KRR were improved by an average of 5.8%, 6.2%, 5.5% and 6.1% compared to RF, ranging from 8.1% to 9.3% for TNB and from 2.4% to 4.0% for NBA, respectively. For TNB, SVR and KRR showed the highest accuracies (0.295 for both), and Adaboost.R2_KRR yielded the highest accuracies on NBA (0.258). In the meantime, in the comparison of unbiasedness, SVR produced the lowest genomic prediction bias, and the regression coefficient was close to 1.0, while Adaboost.R2 method with both base learner SVR and KRR produced larger bias. As a trade-off between accuracy and unbiasedness, SVR and KRR had the most robust

prediction ability, which also confirmed by the results of MSE and MAE, in which SVR and KRR had the smallest MSE and MAE in all scenarios.

It should be noted that the better performance of ML methods was acquired by tuning hyper-parameters (Table S1). Compared with using the default hyper-parameters, the accuracy was improved by 14.3% on average from the ML methods with optimal hyper-parameters (Table S2), the accuracy of SVR, KRR, RF and Adaboost.R2 with optimal hyper-parameters gained improvements by 15.7%, 11.7%, 9.8% and 15.0% respectively on the genomic prediction accuracies for TNB, and for NBA, the improvements were 13.4%, 15.3%, 10.2% and 23.4%, respectively. As for unbiasedness, except for SVR on TNB, the unbiasedness of all ML methods using the default parameters was lower than the unbiasedness using the optimal parameters.

Computing time

The computing time of each method is demonstrated in Table 4. Among all methods, KRR was the fastest algorithm, it took an average of 1.16 minutes in each iteration of cross-validation to complete the analysis, requiring considerably less time than GBLUP (2.07 minutes) and ssGBLUP (3.23 minutes). The computing efficiency of SVR (5.28 minutes) and Adaboost.R2_KRR (5.16 minutes) were comparable with KRR, GBLUP and ssGBLUP. However, RF (53.45 min) and Adaboost.R2_SVR (85.34 min) ran slowly among ML methods. Adaboost.R2 based on KRR (Adaboost.R2_KRR) was much more time-saving than Adaboost.R2_SVR. Since the MCMC algorithm required more iteration time to reach convergence, BayesHE was the slowest as expected, and it took 226.12 minutes for each cross-validation.

Discussion

Our results elucidated that ssGBLUP performed better than GBLUP in accuracy in all scenarios investigated, which was consistent with previous studies [25, 39-41] . It could be explained by the fact that GBLUP utilized phenotypic information only from genotyped individuals, while ssGBLUP simultaneously used information of both genotyped and non-genotyped individuals to construct a genotype-pedigree relationship matrix (H matrix). Since non-genotyped individuals were related to individuals in the validation population on the pedigree, ssGBLUP took advantage of the phenotypic information of the whole population to obtain better prediction results. However, in our research, ssGBLUP only produced slightly higher accuracies for the two reproduction traits, and the improvements were much lower than those obtained by all ML methods. The lower improvement of ssGBLUP may be due to the following reasons. (I) Weak relationship between the non-genotyped reference population and genotyped candidates in the pedigree. In our study, only 143 of the 789 non-genotyped reference population used by ssGBLUP had pedigree information, and only 46 and 40 individuals' sires and dams were included in the 2566 genotyped individuals, indicating that the relationship between non-genotyped reference animals and genotyped candidates was pretty weak, making tiny contribution to the genomic prediction. Li et al.[40] showed that the improvement of ssGBLUP over GBLUP on accuracy was almost entirely contributed by non-genotyped close relatives of candidates. It can also be observed from Figure S1 that the greater the weight of the A matrix, the lower the accuracy, indicating that the information obtained from pedigree is limited, resulting in ssGBLUP not exerting its

advantages greatly. (II) The low heritabilities of TNB and NBA. In this study, the heritabilities for the two traits were both 0.12, which was generally consistent with other reports [25, 42, 43], therefore, it cannot get enough accuracy from the pedigree information. This also confirmed by other studies, that a certain improvement can be achieved by adding a smaller reference population for traits with medium or high heritability[2, 44].

In this study, we investigated the performance of ML methods in genomic prediction, and demonstrated their superiorities compared to classical methods GBLUP, ssGBLUP and Bayesian methods. Generally, the following characteristics of ML methods make it potentially attractive to genomic prediction. (I) Although ML methods generally require moderate fine-tuning of hyper-parameters, and the default hyper-parameters usually do not perform badly [33]. According to our results, the average improvement of ML methods after tuning parameters was 14.3% over using the default hyper-parameters, nonetheless, all ML results without tuning hyper-parameters performed better than GBLUP except for RF in TNB, with an improvement from 0.5% to 8.2% (Table S2). (II) ML methods could handle the number of parameters larger than the sample size, it is very efficient in the case with high-density genetic markers for GS [45]. (III) ML methods do not make distribution assumptions about the genetic determinism underlying the trait, enabling to capture the possible non-linear relationships between genotype and phenotype in a flexible way [45], and it is different from GBLUP and Bayesian methods, which assumes that all marker effects follow the same normal distribution, or have different classes of shrinkage for different SNP effects. In addition,

ML methods can take the correlation and interaction of markers into account as well, while linear models based on pedigree and genomic relationships may not provide a sufficient approximation of the genetic signals generated by complex genetic systems [14]. Consequently, when traits are affected by non-additive effects, especially epistasis, ML methods can achieve more accurate predictions [23].These make ML methods gain large advantage over GBLUP and BayesHE even they only use genotyped animals.

Our results showed that ML methods have improved the prediction accuracy of the reproduction traits in Chinese Yorkshire pig population. SVR, KRR, RF and Adaboost.R2 reflected the superiority of the ML methods, with an average improvement of 20.5%, 21.0%, 14.1% and 20.5% respectively over GBLUP. Liang et al. [46] pointed out that the average improvement of SVR on beef cattle reached a staggering 12.7% . An et al. [13] designed a Cosine kernel-based KRR (K_cRR) and reported that the accuracy of K_cRR was improved by 13.1% compared with GBLUP in three traits of Chinese Simmental beef cattle population. Alves et al.[38] reported SVR has the highest genomic prediction ability in the comparison with GBLUP, BLASSO, Bayesian regularized ANN and RF in the genomic prediction on the reproductive traits of Nellore cattle.

Currently, many ML methods are available, and their performance varied in different scenarios. It is difficult to pick the optimal ML method for genomic prediction. In this study, we implemented SVR, KRR, RF and Adaboost.R2 in the genomic prediction. On the whole, SVR and KRR performed best, and our findings were consistent with other studies showing SVR and KRR had been widely used in the genomic prediction [13, 18,

[23, 47]. In the present study, for SVR and KRR, we used a non-linear kernel function (RBF kernel) to map the original input data to a high-dimensional feature space and then constructed a linear model in the feature space to estimate GEBVs, and finally constructed a nonlinear model. In all scenarios of this study, the prediction accuracy of SVR and KRR were almost equivalent. One explanation is that the main difference between SVR and KRR is that KRR assumes that most features hardly affect the estimation of GEBVs, so the coefficients of a large number of features are as close to zero as possible, and only certain features have a greater impact on GEBV [46]. SVR and KRR were therefore respectively chosen as weak learners for Adaboost.R2.

However, Adaboost.R2 did not show the advantages of its integration capabilities compared with single learning algorithms (SVR and KRR). It mainly because the currently SVR and KRR are sufficient to exert prediction abilities, which may limit the benefit of using ensemble learning. Besides, owing to the slow tuning process of Adaboost.R2, we did not precisely tune the hyper-parameters in this research, resulting in slightly lower prediction accuracy than SVR and KRR. One alternative strategy for Adaboost.R2 is integrating more learners. Liang et al. [48] developed a stacking ensemble learning framework (SELF) that integrated SVR, KRR, and ENET to predict GEBVs and showed excellent performance. Among all ML methods in this study, RF demonstrated low prediction ability and computational efficiency. The prediction accuracy of RF is mainly affected by the number and maximum depth of decision trees [46], but in order to weigh the practical application feasibility of RF, it is impractical to precisely tune the number of trees, resulting in not training the most ideal RF model,

thus compromising its prediction accuracy.

Although ML significantly outperformed GBLUP and Bayesian methods, one problem should be noted is the hyper-parameter optimization. In this study, the average improvement after tuning parameters was 14.3% over without tuning. Since ML models have multiple hyper-parameters and they are generally sensitive to changes in hyper-parameters, it might be time-consuming to perform strict hyper-parameter adjustments in the process of training models to obtain high accuracies. And the optimal hyper-parameter depends on the character of traits, data sets *etc.*. Usually, the effect of the default hyper-parameters did not perform poorly as discussed above, and failure to find suitable hyper-parameters may greatly reduce the prediction effect of ML methods [46]. If hyper-parameter automation can be realized during ML operation, it will greatly reduce the time used for hyper-parameter adjustment and greatly increase the application of ML methods in genomic prediction.

Conclusions

In this study, we compared four ML methods with GBLUP, ssGBLUP and BayesHE to explore their efficiency of genomic prediction on reproduction traits in pigs. We compared the prediction accuracy, unbiasedness, MSE, MAE and computation time of different methods through 20 replicates of 5-fold CV. Our results showed that ML methods possess a significant potential to improve genomic prediction over GBLUP, ssGBLUP and BayesHE. ML methods outperformed in all scenarios, they yielded higher accuracy and smaller MSE and MAE. Among ML methods, SVR and KRR performed the best overall, which yielded higher accuracies, lower bias, and higher

460 computing efficiency. Our findings demonstrated that ML methods are more efficient
461 than traditional genomic selection methods, it could be new options for genomic
462 prediction.

463 **List of abbreviations**

GS	genomic selection
GEBV	genomic breeding values estimation
GBLUP	genomic BLUP
ssGBLUP	single-step GBLUP
ML	machine learning
RF	random forest
SVR	support vector regression
KRR	kernel ridge regression
RKHS	Reproducing Kernel Hilbert space
SVM	support vector machine
TNB	total number of piglets born
NBA	number of piglets born alive
A matrix	pedigree relationship matrix
EBV	estimated breeding values
y_c	corrected phenotypes
DR2	the dosage R-squared measure
COR	the genotype correlation
CR	the genotype concordance rate

MAF	minor allele frequency
BayesHE	Bayesian Horseshoe
5-fold CV	five-fold cross validation
MSE	mean squared error
MAE	mean absolute error
K_cRR	cosine kernel-based KRR
SELF	stacking ensemble learning framework

Declarations

Ethics approval and consent to participate

Animal samples used in this study were approved by the Animal Care and Use Committee of China Agricultural University. There was no use of human participants, data or tissues.

Consent for publication

Not applicable

Availability of data and material

The datasets used or analyzed during the present study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This work was supported by grants from the National Key Research and Development

Project (2019YFE0106800), Modern Agriculture Science and Technology Key Project of Hebei Province (19226376D), China Agriculture Research System of MOF and MARA.

Authors' contributions

XDD designed the experiments. XW performed statistical analysis and wrote the manuscript. SLS provided help on BayesHE. XDD revised the manuscript. All authors read and approved the final manuscript.

Acknowledgement

The authors gratefully acknowledge the constructive comments from reviewers.

References

1. de Roos AP, Schrooten C, Veerkamp RF, van Arendonk JA. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J Dairy Sci. 2011;94(3):1559-67.

2. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92(2):433-43.

3. Heffner EL, Jannink JL, Sorrells ME. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. The Plant Genome. 2011;4(1).

4. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123(4).

5. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci USA. 2016;113(28):E3995-E4004.

6. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414-23.

7. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 2009;92(9):4648-55.

8. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.

9. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 2001;157(4):1819-29.

10. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.

11. Varona L, Legarra A, Toro MA, Vitezica ZG. Non-additive Effects in Genomic Selection. Front Genet. 2018;9:78.

12. Gianola D, Campos G, Gonzalez-Recio O, Long N, Okut H, Rosa G, et al. Statistical Learning Methods For Genome-based Analysis Of Quantitative Traits2018.

13. An B, Liang M, Chang T, Duan X, Gao HJBiB. KCRR: a nonlinear machine learning with a modified genomic similarity matrix improved the genomic prediction efficiency. Briefings in Bioinformatics. 2021.

14. Gianola D, Okut H, Weigel KA, Rosa GJ. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 2011;12:87.

15. González-Recio O, Rosa GJM, Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science. 2014;166:217-31.

16. Montesinos-Lopez OA, Martin-Vallejo J, Crossa J, Gianola D, Hernandez-Suarez CM, Montesinos-Lopez A, et al. A Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding. G3 (Bethesda). 2019;9(2):601-18.

17. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics. 2008;9:319.

18. Gonzalez-Camacho JM, Ornella L, Perez-Rodriguez P, Gianola D, Dreisigacker S, Crossa J. Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. Plant Genome. 2018;11(2).

19. Ornella L, Perez P, Tapia E, Gonzalez-Camacho JM, Burgueno J, Zhang X, et al. Genomic-enabled prediction with classification algorithms. Heredity (Edinb). 2014;112(6):616-26.

20. Noe F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. Curr Opin Struct Biol. 2020;60:77-84.

543 21. Kojima K, Tadaka S, Katsuoka F, Tamiya G, Yamamoto M, Kinoshita K. A
544 genotype imputation method for de-identified haplotype reference information by using
545 recurrent neural network. PLoS Comput Biol. 2020;16(10):e1008207.
546 22. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multi-
547 task deep neural networks. PLoS One. 2018;13(6):e0198216.
548 23. Long N, Gianola D, Rosa GJ, Weigel KA. Application of support vector regression
549 to genome-assisted prediction of quantitative traits. Theor Appl Genet.
550 2011;123(7):1065-74.
551 24. Madsen P, Jensen J, Labouriau R, Christensen O, Sahana G. DMU - A Package for
552 Analyzing Multivariate Mixed Models in quantitative Genetics and Genomics.
553 CanadaAugust 17-22, 2014.
554 25. Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G. Improving genetic
555 evaluation of litter size and piglet mortality for both genotyped and nongenotyped
556 individuals using a single-step method. J Anim Sci. 2015;93(2):503-12.
557 26. Browning BL, Browning SR. A unified approach to genotype imputation and
558 haplotype-phase inference for large data sets of trios and unrelated individuals. Am J
559 Hum Genet. 2009;84(2):210-23.
560 27. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-
561 generation PLINK: rising to the challenge of larger and richer datasets. Gigascience.
562 2015;4:7.

563 28. Forni S, Aguilar I, Misztal I. Different genomic relationship matrices for single-
564 step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol.
565 2011;43:1.
566 29. Shi S, Li X, Fang L, Liu A, Su G, Zhang Y, et al. Genomic Prediction Using
567 Bayesian Regression Models With Global-Local Prior. Front Genet. 2021;12:628205.
568 30. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin
569 classifiers. Proceedings of the fifth annual workshop on Computational learning
570 theory - COLT '921992. p. 144-52.
571 31. Müller AC, Guido S. Introduction to machine learning with Python: A guide for
572 data scientists. Sebastopol: CA: O'Reilly Media, Inc; 2017.
573 32. Exterkate P, Groenen PJF, Heij C, van Dijk D. Nonlinear forecasting with many
574 predictors using kernel ridge regression. International Journal of Forecasting.
575 2016;32(3):736-53.
576 33. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
577 34. Zhou Z. Machine Learning. In: Press TU, editor. Beijing,China: Tsinghua
578 University
579 Press2016. p. 247-63.
580 35. Drucker H, editor Improving Regressors using Boosting Techniques. ICML; 1997.
581 36. Shrestha DL, Solomatine DP. Experiments with AdaBoost.RT, an improved
582 boosting scheme for regression. Neural Comput. 2006;18(7):1678-710.
583 37. Steiger JH. Tests for comparing elements of a correlation matrix. Psychological
584 Bulletin. 1980;87(2):245-51.

585 38. Alves AAC, Espigolan R, Bresolin T, Costa RM, Fernandes Junior GA, Ventura
586 RV, et al. Genome-enabled prediction of reproductive traits in Nellore cattle using
587 parametric models and machine learning methods. Anim Genet. 2021;52(1):32-46.
588 39. Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, Ding X. Using imputation-based whole-
589 genome sequencing data to improve the accuracy of genomic prediction for combined
590 populations in pigs. Genet Sel Evol. 2019;51(1):58.
591 40. Li X, Wang S, Huang J, Li L, Zhang Q, Ding X. Improving the accuracy of genomic
592 prediction in Chinese Holstein cattle by using one-step blending. Genet Sel Evol.
593 2014;46:66.
594 41. Su G, Madsen P, Nielsen US, Mantysaari EA, Aamand GP, Christensen OF, et al.
595 Genomic prediction for Nordic Red Cattle using one-step and selection index blending.
596 J Dairy Sci. 2012;95(2):909-17.
597 42. Song H, Zhang Q, Ding X. The superiority of multi-trait models with genotype-by-
598 environment interactions in a limited number of environments for genomic prediction
599 in pigs. J Anim Sci Biotechnol. 2020;11:88.
600 43. Song H, Zhang J, Jiang Y, Gao H, Tang S, Mi S, et al. Genomic prediction for
601 growth and reproduction traits in pig using an admixed reference population. Journal
602 of Animal Science. 2017;95(8).
603 44. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals
604 and their use in breeding programmes. Nat Rev Genet. 2009;10(6):381-91.

605 45. Piles M, Bergsma R, Gianola D, Gilbert H, Tusell L. Feature Selection Stability
606 and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in
607 Pigs Using Machine Learning. Front Genet. 2021;12:611506.
608 46. Liang M, Miao J, Wang X, Chang T, An B, Duan X, et al. Application of ensemble
609 learning to genomic selection in chinese simmental beef cattle. J Anim Breed Genet.
610 2021;138(3):291-9.
611 47. Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi AJAPS.
612 Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic
613 Best Linear Unbiased Prediction in different scenarios of genomic evaluation. Animal
614 Production Science. 2016;57(2):229-36.
615 48. Liang M, Chang T, An B, Duan X, Du L, Wang X, et al. A Stacking Ensemble
616 Learning Framework for Genomic Prediction. Front Genet. 2021;12:600040.

617

618 **Table 1** Summary of two reproduction traits of Yorkshire pigs

Trait[a]	Number of records	Birth year	Genotyped animals	Mean	SD	Minimum	Maximum	σ^2_a	σ^2_e	h^2(SE)
TNB	4274	2016-2020	2566	13	3.38	3	24	1.26	8.95	0.12(0.034)
NBA	4274	2016-2020	2566	12	3.13	3	24	0.98	7.13	0.12(0.032)

619 [a] TNB: total number of piglets born; NBA: number of piglets born alive
620 SE: standard error

621

622 **Table 2** Accuracies and unbiasedness of genomic prediction on TNB and NBA from 7 methods in

623 20 replicates of 5-fold Cross-validation

Method	TNB		NBA	
	Accuracy	Unbiasedness	Accuracy	Unbiasedness
GBLUP	0.248[a] ±0.026	0.958±0.132	0.208[a] ±0.025	0.931±0.142
ssGBLUP	0.251[a] ±0.026	0.901±0.121	0.221[ab] ±0.026	0.844±0.113
BayesHE	0.243[a] ±0.025	1.015±0.148	0.207[a] ±0.026	1.009±0.171
SVR	0.295[b]±0.025	1.23±0.119	0.254[b] ±0.023	1.106±0.11
KRR	0.295[b] ±0.025	1.266±0.125	0.256[b] ±0.023	1.151±0.113
RF	0.270[ab] ±0.029	1.229±0.152	0.248[ab] ±0.028	1.188±0.147
Adaboost.R2_SVR	0.293[b] ±0.025	1.363±0.138	0.254[b] ±0.024	1.256±0.131
Adaboost.R2_KRR	0.292[b] ±0.025	1.344±0.136	0.258[b] ±0.024	1.249±0.129

624 The different superscript of accuracy indicates the significant difference by the Hotelling-Williams

625 test.

626

627 **Table 3** MAE and MSE of 7 methods for TNB and NBA as assessed with 20 replicates of 5-fold

628 CV

Method	TNB		NBA	
	MSE[a]	MAE[b]	MSE[a]	MAE[b]
GBLUP	5.259	1.749	4.168	1.606
ssGBLUP	5.26	1.748	3.95	1.532
BayesHE	5.32	1.763	4.023	1.556
SVR	5.129	1.730	3.880	1.521

KRR	5.134	1.731	3.876	1.521
RF	5.212	1.747	3.901	1.527
Adaboost.R2_SVR	5.158	1.739	3.892	1.528
Adaboost.R2_KRR	5.153	1.737	3.883	1.526

[a] MSE: mean squared error

[b] MAE: mean absolute error

Table 4 Average computing time in one each iteration of the 5-fold Cross validation for different genomic prediction methods

Method	TNB	NBA
GBLUP	2min 06s	2min 02s
ssGBLUP	3min 12s	3min 16s
BayesHE	3h 57min 1s	3h 35min 13s
SVR	5min 27s	5min 07s
KRR	1min 04s	1min 16s
RF	50min 38s	56min 16s
Adaboost.R2_(SVR)	1h 35min 13s	1h 15min 28s
Adaboost.R2_(KRR)	5min 03s	5min 16s

Figure captions

Figure 1 Imputation accuracy

Imputation accuracy of GenoBaits Porcine SNP 50K to PorcineSNP50 BeadChip at different minor allele frequency (MAF) intervals (a) and chromosomes (b).

DR2, the estimated squared correlation between the estimated allele dose and the true allele dose; Genotype concordance rate (CR), the ratio of the correctly imputed genotypes; Genotype correlation (COR), the correlation coefficient between the imputed variants and the true variants.

Figures

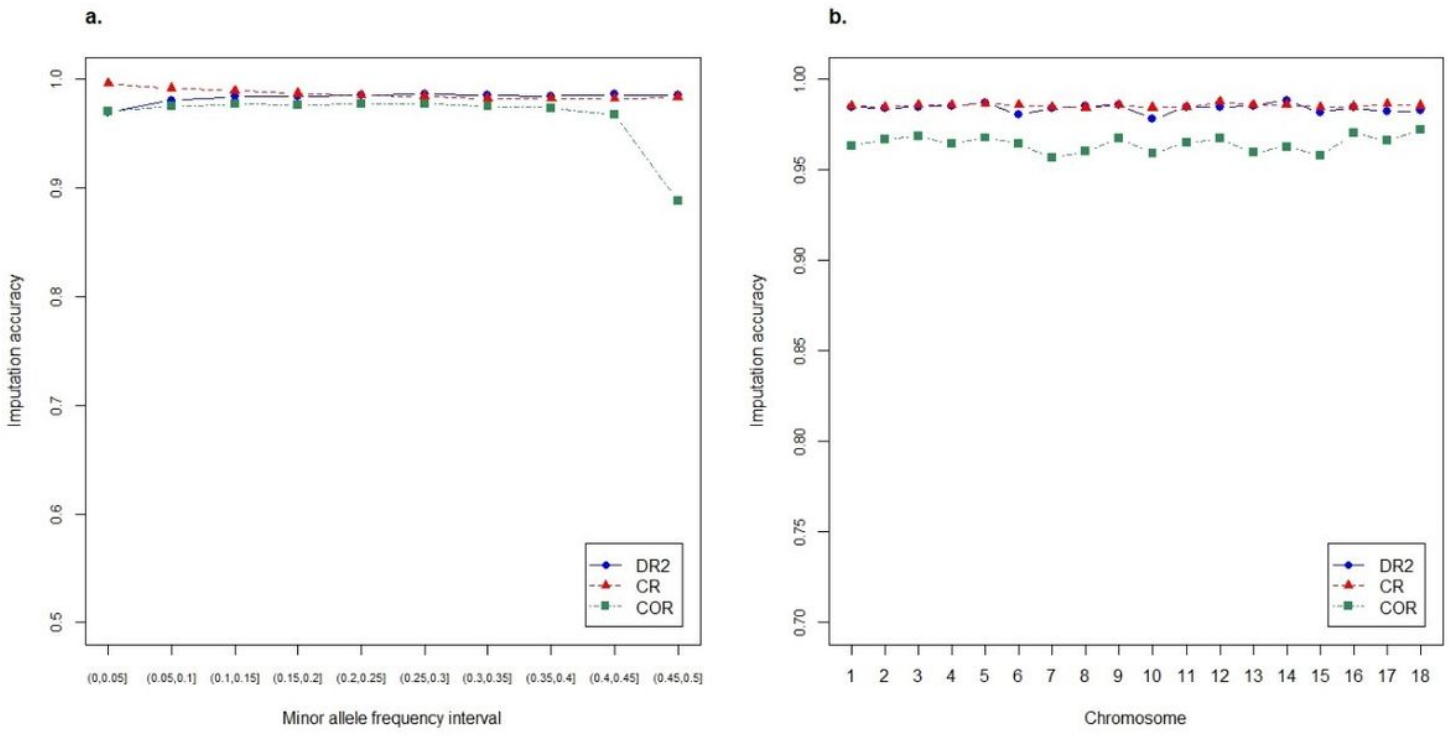

Figure 1

Imputation accuracy Imputation accuracy of GenoBaits Porcine SNP 50K to PorcineSNP50 BeadChip at different minor allele frequency (MAF) intervals (a) and chromosomes (b). DR2, the estimated squared correlation between the estimated allele dose and the true allele dose; Genotype concordance rate (CR), the ratio of the correctly imputed genotypes; Genotype correlation (COR), the correlation coefficient between the imputed variants and the true variants.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplymentary.docx