Spanish Journal of Agricultural Research
18 (3), e0405, 10 pages (2020)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2020183-15228
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)

RESEARCH ARTICLE

OPEN ACCESS

The importance of disease incidence rate on performance of GBLUP, threshold BayesA and machine learning methods in original and imputed data set

Yousef Naderi (Naderi, Y)[1] and Saadat Sadeghi (Sadeghi, S)[2]

[1] *Department of Animal Science, Young Researchers and Elite Club, Astara Branch, Islamic Azad University, Astara, Iran* [2] *Animal breeding and genetics group, Ghiam Dairy Complex, Isfahan, Iran.*

Abstract

Aim of study: To predict genomic accuracy of binary traits considering different rates of disease incidence.

Area of study: Simulation.

Material and methods: Two machine learning algorithms including Boosting and Random Forest (RF) as well as threshold BayesA (TBA) and genomic BLUP (GBLUP) were employed. The predictive ability methods were evaluated for different genomic architectures using imputed (*i.e.* 2.5K, 12.5K and 25K panels) and their original 50K genotypes. We evaluated the three strategies with different rates of disease incidence (including 16%, 50% and 84% threshold points) and their effects on genomic prediction accuracy.

Main results: Genotype imputation performed poorly to estimate the predictive ability of GBLUP, RF, Boosting and TBA methods when using the low-density single nucleotide polymorphisms (SNPs) chip in low linkage disequilibrium (LD) scenarios. The highest predictive ability, when the rate of disease incidence into the training set was 16%, belonged to GBLUP, RF, Boosting and TBA methods. Across different genomic architectures, the Boosting method performed better than TBA, GBLUP and RF methods for all scenarios and proportions of the marker sets imputed. Regarding the changes, the RF resulted in a further reduction compared to Boosting, TBA and GBLUP, especially when the applied data set contained 2.5K panels of the imputed genotypes.

Research highlights: Generally, considering high sensitivity of methods to imputation errors, the application of imputed genotypes using RF method should be carefully evaluated.

Additional key words: accuracy; boosting; disease susceptibility; imputation; random forest

Abbreviations used: GBLUP (genomic best linear unbiased prediction); GS (genomic selection); LD (linkage disequilibrium); OOB (out of bag); QTLs (quantitative trait loci); RF (random forest); SNP (single nucleotide polymorphism); TBA (threshold BayesA)

Authors' contributions: Analyzed the data: SS. Wrote the paper: YN. Both authors read and approved the final manuscript.

Citation: Naderi, Y; Sadeghi, S (2020). The importance of disease incidence rate on performance of GBLUP, threshold BayesA and machine learning methods in original and imputed data set. Spanish Journal of Agricultural Research, Volume 18, Issue 3, e0405. https://doi.org/10.5424/sjar/2020183-15228

Received: 29 May 2019. **Accepted:** 11 Jul 2020

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Correspondence should be addressed to Yousef Naderi: y.naderi@iau-astara.ac.ir; yousefnaderi@gmail.com

Introduction

For several decades, most phenotypic variation in dairy cattle populations has focused on continuous traits, especially milk yield (Egger-Danner *et al.*, 2015). However, economic benefits require better understanding of novel functional traits and their direct inclusion in dairy cattle breeding program (Naderi *et al.*, 2016). Functional traits (*e.g.* resistance to disease and direct information on animal health) are vital, due to the importance of animal welfare and the human tendency for healthy and high-quality products. These traits are generally categorical, influenced by multiple genes, and deviate from Mendelian inheritance and normal distribution, all of which pose statistical challenges for genomic estimated breeding values (GEBV) estimation (Wang *et al.*, 2013; Naderi *et al.*, 2016).

Meuwissen *et al.* (2001) introduced the statistical pattern of genomic selection (GS) which has shown a comprehensive gain in the types of statistical models applied in genomic evaluation for approximately two decades (De Los Campos *et al.*, 2009). Generally, these methods (*e.g.* genomic best linear unbiased prediction –GBLUP– and Bayesian methods) are based on linear regression. In recent years, machine learning methodology (Breiman, 2001) as

a non-parametric method along with GBLUP (VanRaden, 2008) and threshold versions of Bayesian methods have commonly extended to solve the challenges of genomic selection in threshold traits (González-Recio & Forni, 2011; Wang C *et al.*, 2017). Random forest (RF) and boosting are powerful machine learning methods to recognize gene-gene, protein-protein and gene-environment interactions. These methods are able to detect disease associated genes and to model the relationship among combinations of markers in order to select genes associated with the target trait. In addition, the regulatory elements in DNA or protein sequences are identified by these methods to classify various samples in gene expression of microarrays data (Yang *et al.*, 2010) and also genomic prediction accuracy has been improved using RF and Boosting (González-Recio & Forni, 2011; Ghafouri-Kesbi *et al.*, 2017).

Genomic accuracy of statistical algorithms depends mainly on the genetic architecture of target traits, including the number of QTL (Wientjes *et al.*, 2015), level of linkage disequilibrium (LD) (Yin *et al.*, 2014), marker density (Wang Q *et al.*, 2017) and heritability (Bohlouli *et al.*, 2017). Furthermore, the rate of disease incidence into training set is another important factor affecting the accuracy of genomic prediction in threshold traits. Recently, some studies have shown that genomic accuracy can be influenced by the different compositions of the training set (Mc Hugh *et al.*, 2011; Pimentel *et al.*, 2013). A study by Naderi *et al.* (2016) found that the genomic prediction accuracy increased with high rate of disease incidence into the training set, especially when applying the RF method.

Despite the important role of genomic selection in achieving high genomic accuracy and its long-term cost-effectiveness, the cost of genotyping and the economic aspects should not be disregarded in the short term. Animal breeding programmers require harmony among these factors in order to maximize the benefits to farmers. In this regard, genotype imputation could be applied to infer higher density genotypes with an acceptable estimation of genomic accuracy to reduce the cost of genotyping (Ventura *et al.*, 2016; Friedrich *et al.*, 2018). Furthermore, imputation could assist GS by allowing screening on a larger number of young individuals (Chen *et al.*, 2014; Lakhssassi & Recio, 2017).

Whereas real data offer the advantage of reflecting complexity, simulated data allow the researcher to explore important aspects, such as the genetic architecture of the trait, number of markers used for analysis and degree of relatedness between the training and prediction populations. It also offers the possibility of evaluating some sources of variability, such as drift, which cannot be assessed with most real data (Daetwyler *et al.*, 2013). The objective of this study has been to assess the performance of GBLUP, threshold BayesA and machine learning methods (RF and Boosting) for the evaluation of binary diseases traits, considering different genotyping strategies (from very low to mid density), different incidence rates of diseases in the populations and different genetic architectures for the disease trait, in order to define an optimal strategy to evaluate these type of traits.

Material and methods

Population structure

QMSim software was used (Sargolzaei & Schenkel, 2009) to generate phenotypes, genotypes, and true breeding values by applying stochastic simulations. Along the genome, 50010 bi-allelic single nucleotide polymorphism (SNP) markers (1667 per chromosome) were evenly spaced along 30 chromosomes, each 100 cM long. During the first phase of the historical population, the population started to achieve the intended level LD for a basic population with an effective population size (Ne) = 1472 (400 males and 4600 females), which in turn were randomly mated for 1000 generations. In the second phase of the historical population, the effective size of over 100 generations was decreased from 1472 to 500 individuals by a "bottleneck" to produce a higher level of LD. In the third phase of the historical population, the effective size was increased from 500 to 1472 for 100 generations, by considering 400 males and 4600 females. After that, 5000 animals from the last generation were used as founders of the recent population andexpanded via a random mating design for other 10 generations. In the meantime, one offspring for each mating was considered, with an equal proportion of both genders, and replacement proportions were 0.2 and 0.5 for females and males, respectively. In each generation, the criteria for selection/culling was estimated breeding value (EBV)/age. Four different scenarios were considered to reflect variations of genomic architecture, including the level of LD and heritability and number of QTL (Table 1).

Two different QTL sets (450 and 150 QTLs) were randomly located along each chromosome with effects sampled from a gamma (β=0.4) distribution (Meuwissen *et al.*, 2001). To simulate a wide range of polymorphic SNP loci, the mutation rate was considered to be 2.5×10^{-5} and 2.5×10^{-3} for QTL per locus and per generation and

Table 1. Different scenarios with respect to level of LD and heritability and number of QTL

Variable	Scenarios			
	I	II	III	IV
Heritability	0.25	0.25	0.1	0.1
No. of QTL[a]	450	150	150	150
Level of LD[b]	low	low	low	high

[a]QTL: quantitative trait loci. [b]LD: linkage disequilibrium.

marker, respectively. Phenotypes were simulated with low (0.1) and moderate (0.25) heritabilities.

Discrete phenotype

Individuals of the last generation (1210 generation) as the validation set and individuals of three generations before the validation set were considered as the training set (1207 to 1209 generation). Three strategies were used to create a binary phenotype. In the first strategy, the phenotype of the individuals was coded as 0 (12600 healthy individuals; approx. 84%) or 1 (2400 diseased individuals; a disease incidence rate of approx. 16%) depending on whether their simulated phenotype was respectively above or below $\overline{x}$- SD (standard deviation). In the second strategy, 5100 healthy individuals from the first strategy were randomly recoded as sick to create a disease incidence rate of approx. 50%. In the third strategy, 10200 healthy individuals from the first strategy were randomly recoded as sick to create a disease incidence rate of approx. 84%. Phenotypes of the validation set were assumed unknown. To ensure data quality control, minor allele frequency (MAF) < 3% was considered as the criteria to filter out markers of low frequency. Ten repetitions were considered at all stages of the process in each scenario.

Scenarios for masking genotypes

Different changes in simulated scenarios with 50K SNPs densities (original scenarios) were made to imitate the real condition of genotyping for uncalled genotypes with imputing genomic data. For this purpose, 95%, 75% and 50% of marker (to create 2.5K, 12.5K and 25K panels) were removed. Afterwards, missing genotypes were imputed using FImpute program (Sargolzaei *et al.*, 2011). FImpute uses the family imputation algorithm followed by population imputation steps based on a sliding window technique. The imputation accuracy was calculated per animal and per SNP by the correlation between the imputed and original genotypes for all replications as an appropriate approach to minimize dependence on the allele frequency. This criterion was calculated for every marker and individual with genotypes coded 0, 1 and 2 as described above.

Statistical methods

—Random Forest (RF)

RF uses different variables at each split for each tree. This algorithm uses an ensemble of unpruned decision trees. It also uses a random subset of predictors to detect the best split at each node grown on bootstrap samples of observations. The RF predictions for each observation ($\hat{f}_{rf}^{P}(x)$) were calculated through averaging the performances over P trees ($[T(x.\Psi p)]_1^P$) for those observations that are not applied to build the tree. Ψp determines the p_{th} tree of RF in terms of split variables, cut points at each node, and terminal node amounts. The java package RanFoG (González-Recio & Forni, 2011) was used for RF analysis in the framework of the following model:

$$\hat{f}_{rf}^{P}(x)=\frac{1}{P}\sum_{p=1}^{P}[T(x.\Psi p)]$$

Out of bag (OOB) error is a basic feature in RF. Each tree is grown through a bootstrapping sample of the data irrespective of 1/3 observations. Some individuals will emerge more than once and others will not emerge at all. The ones that do not emerge are called OOB observations and used as internal training set for trees. These OOB samples are the source of data used in RF for estimation of the OOB error by which the performance of RF can be assessed. To achieve the best performance of the model, parameters of RF should be tuned. These include *ntree* (the number of trees to grow), *mtry* (number of variables randomly sampled as candidates at each split) and *nodesize* (minimum size of terminal nodes). In this study, different combinations of tuning parameters were tested to ensure the optimum combination of these parameters using OOB error value. Eventually, the best combination of tuning parameters, including *ntree*=5000, *mtry*=10000 and *nodesize*=2, were used to analyze the RF method.

—Stochastic gradient Boosting

Boosting algorithm (Freund & Schapire, 1996) involves training multiple models in a sequence in which the error function that is used to train a particular model depends on the performance of the previous models. In Boosting model, the base classifiers are trained in sequence, and each base classifier is trained using a weighted form of the data set in which the weighting coefficient associated with each data point depends on the performance of the previous classifiers (Bishop, 2006). Boosting methods improve predictive ability as it concerns the interactions among predictive variables and enabling variable selection, unaffected by numerous correlated and irrelevant variables, outliers and missing data. The following formula was applied to the Boosting method (Ghafouri-Kesbi *et al.*, 2017):

$$f(x)=\sum_{m=1}^{m}\beta_m\, b(x;\gamma_m)$$

where β_m (m= 1,2,…,M) represents the basis expansion coefficients and $b(x;\gamma_m)$ denotes sample functions of the multivariate argument x, along with a collection of

parameters (γ=γ_1,γ_2, …, γ_M). Prediction takes place by weighting the ensemble outputs of all the regression trees. The package *gbm* (Wimmer *et al.*, 2015) in R software was applied to run Boosting. The number of tree (*ntree*), shrinkage rate or learning rate (*lr*) and tree depth or tree complexity (*tc*) are the most tuning parameters in Boosting. Different combinations of parameters were applied to achieve the best combinations of tuning parameters using the least cross-validation error. In this process, the 10-fold cross-validation was used to evaluate the efficiency of Boosting. Eventually, the best combination of tuning parameters, including *ntree*=5000, *lr*=0.02 and *tc*=8, was used to analyze data in all the scenarios using the Boosting method.

—Genomic best linear unbiased prediction (GBLUP)

In order to analyze the GBLUP method, AI-REML algorithms of the DMU software package were used (Madsen & Jensen, 2013). They allow the specification of a generalized linear mixed model with a logit link function for discrete data. For this purpose, the following model was applied:

$$logit\ (\pi_r) = \log\left[\frac{\pi_r}{1-\pi_r}\right] = \emptyset + \gamma_r$$

where π_r is the probability of disease incidence for animal r; $\emptyset$ is the total mean effect; γ_r is the random individual effect. It was included by considering the genomic relationships among individuals based on SNP marker data. Gmatrix software (Su & Madsen, 2013) was employed to calculate the genomic relationship matrix (G) according to the method presented by VanRaden (2008). To circumvent problems with matrix singularity, a value of 0.01 was added to the diagonal of genomic relationship matrix G.

—Threshold Bayes A (TBA)

To infer SNPs effects in genomic selection, Meuwissen *et al.* (2001) proposed BayesA methods. Later González-Recio & Forni (2011) and Wang *et al.* (2013) developed a version of BayesA to estimate genomic breeding values of animal discrete traits. This method assumed that all SNPs are effective, and each has a different variance. The R-package BGLR (De Los Campos *et al.*, 2009) was applied to analyze TBA using the following model.

$$\lambda = \mu 1 + Xb + e$$

where λ represents the underlying liability variable vector for a vector of phenotypes recorded (0 or 1); μ represents the population mean; 1 represents a column vector of ones ($n \times 1$); b= [bj] corresponds to the vector for the regression coefficient estimates of the p markers, SNP assumed independently and normally distributed a priori as N (0, σ_j^2); the prior distribution of σ_j^2 (an unknown variance associated with SNP j) assumed to be $\sigma_j^2 \sim \upsilon_j s^2_j \chi^{-1}_{\upsilon j}$ with $\upsilon_j = 4$ and $s^2_j = 0.002$.X=[x_i] represents a $n \times p$ matrix (n animals genotyped for p SNPs) including values 0, 1 or 2; e represents the residuals assuming N (μ_e=0, σ_e^2=1). The predictive accuracy of the methods for the training set was assessed using the correlation coefficient between the predicted GBV and true GBV. For this purpose, 10 replicates were considered.

Results and discussion

Imputation accuracy

Table 2 presents the imputation accuracy for different proportions of missing genotypes which were imputed to the 50K SNPs panels under different scenarios. In the simulated scenarios, the average imputation accuracy across all replicates was 0.955 (0.908 to 0.989). The correlation between imputed and true genotypes increased markedly as the percentage of missing genotypes decreased from the 2.5K to 25K panels. In comparison with low-LD scenarios, the correlation between imputed and original genotypes was higher in high-LD scenarios (0.938 to 0.989), showing an average genotypes correlation of 0.952 (0.908 to 0.978). However, there was a considerable increase in the accuracy of imputation when the sparse panels from original panels were imputed. The explanation is that similarity of LD patterns between the imputed panel and the reference population serves as an important source for imputing the missing genotypes.

In addition to quantifying imputation accuracy in different scenarios, the results of current study shed light on the effects that the two factors (proportion of missing genotypes and LD patterns) have on imputation accuracy. These outcomes are in agreement with several researchers who reported that the accuracy of conventional imputation

Table 2. Correlation between imputed and original 50k SNPs genotypes by scenario.

Scenarios[a]	Proportion of missing genotypes		
	95% (2.5K)	75% (12.5K)	50% (25K)
I	0.910 (0.016)	0.963 (0.016)	0.975 (0.012)
II	0.914 (0.021)	0.966 (0.019)	0.976 (0.015)
III	0.908 (0.020)	0.967 (0.019)	0.978 (0.014)
IV	0.938 (0.019)	0.980 (0.015)	0.989 (0.011)

[a]I (h^2 = 0.25, LD = low and 450 QTL), II (h^2 = 0.25, LD = low and 150 QTL), III (h^2 = 0.1, LD = low and 150 QTL) and IV (h^2 = 0.1, LD = high and 150 QTL).

methods are strongly dependent on proportion of missing genotypes in validation set and the similarity of LD patterns in reference population (Hickey *et al.*, 2012; Kabisch *et al.*, 2017; Pausch *et al.*, 2017; Friedrich *et al.*, 2018; Sadeghi *et al.*, 2018). The current results showed that imputation based on family and population-based algorithm, such as the one was implemented in FImpute program, could produce reasonable accuracy of imputation for different scenarios containing different proportions of missing genotypes. In other words, because of high LD among SNPs in dense panels and decreasing imputation errors, the accuracy of imputation improved with the increase of markers density. However, as the results in the low-LD sparse panels have shown, this is not always the case (Pimentel *et al.*, 2013).

Rate of disease incidence

Accuracy of genomic prediction in the validation set via the GBLUP and TBA, RF and Boosting methods when the rates of disease incidence allocated to the training set were 50, 84 and 16%, respectively are shown in Tables 3, 4 and 5. The rate of disease incidence allocated to the training set directly affected the predictive ability of all models. Highest prediction accuracies were observed for disease incidences in training sets that reflected the population disease incidence of 0.16%. Generally, the accuracy of genomic prediction improved as long as the rate of disease incidence decreased, showing the best accuracy for Boosting in all scenarios. In other methods, the accuracy of genomic prediction was also influenced by the type of considered scenarios. For example, Boosting and TBA for scenario II and RF and GBLUP for scenario I had the highest values when different rates of disease incidence were used.

Specifically, the total average of prediction accuracies increased from 0.449, 0.498, 0.411 and 0.479 in the third strategy to 0.496, 0.564, 0.478, 0.529 in the first strategy for GBLUP, Boosting, RF and TBA, respectively. There are very limited researches available about the effect of different rates of disease incidence on genomic prediction accuracy using different methods. Naderi *et al.* (2016) simulated different rates of disease incidence in order to compare the performance of RF and GBLUP for genomic predictions of threshold phenotypes based on cow calibration groups. Their results indicated that distribution of binary phenotypic in training set affected the predictive ability of RF and GBLUP so that their performance improved by the increase in proportions of disease incidence up to 20%, and then decreased insignificantly, yet it was more tangible for RF. González-Recio & Forni (2011) investigated genomic accuracy of binary traits (with the same rate of disease incidence in training set) including 2500 animals using machine learning and Bayesian regressions methods. They observed better performance of L_h-Boosting (0.41) and RF (0.36) than TBA (0.26). Also, Sadeghi *et al.* (2018) simulated different scenarios of binary traits (considering the rate of disease incidence equal to 0.5 into training set) to evaluate the accuracy of genomic prediction via RF, TBA, and Bayesian LASSO by altering genetic architectures. They reported that genomic prediction for each method depends on the genomic architecture of population. Shirali *et al.* (2012) investigated the accuracy of BayesC and GBLUP for different rates of

Table 3. Accuracies of genomic estimated breeding values (GEBVs) from genomic BLUP (GBLUP), Boosting, Random Forest (RF) and threshold BayesA (TBA) methods for 50% threshold point (standard deviation across 10 replicates in parentheses)

Model	SNP panels	Scenarios[a]			
		I	II	III	IV
GBLUP	2.5K	0.457 (0.06)	0.442 (0.06)	0.289 (0.07)	0.398 (0.05)
	12.5K	0.495 (0.05)	0.463 (0.06)	0.338 (0.06)	0.436 (0.06)
	25K	0.500 (0.05)	0.479 (0.06)	0.349 (0.07)	0.436 (0.06)
	50K	0.522 (0.06)	0.508 (0.06)	0.376 (0.05)	0.448 (0.06)
Boosting	2.5K	0.482 (0.04)	0.493 (0.05)	0.493 (0.05)	0.429 (0.05)
	12.5K	0.529 (0.04)	0.565 (0.06)	0.409 (0.05)	0.465 (0.04)
	25K	0.550 (0.04)	0.575 (0.03)	0.422 (0.04)	0.471 (0.03)
	50K	0.586 (0.04)	0.592 (0.04)	0.446 (0.05)	0.498 (0.03)
RF	2.5K	0.473 (0.03)	0.397 (0.04)	0.269 (0.04)	0.382 (0.04)
	12.5K	0.527 (0.04)	0.434 (0.05)	0.306 (0.04)	0.411 (0.04)
	25K	0.542 (0.04)	0.443 (0.04)	0.316 (0.05)	0.413 (0.03)
	50K	0.578 (0.03)	0.470 (0.03)	0.343 (0.04)	0.421 (0.03)
TBA	2.5K	0.436 (0.06)	0.489 (0.06)	0.348 (0.07)	0.409 (0.06)
	12.5K	0.482 (0.06)	0.557 (0.06)	0.382 (0.06)	0.447 (0.06)
	25K	0.495 (0.07)	0.568 (0.07)	0.381 (0.06)	0.457 (0.06)
	50K	0.514 (0.06)	0.581 (0.06)	0.432 (0.07)	0.472 (0.06)

[a]I (h^2 = 0.25, LD = low and 450 QTL), II (h^2 = 0.25, LD = low and 150 QTL), III (h^2 = 0.1, LD = low and 150 QTL) and IV (h^2 = 0.1, LD = high and 150 QTL).

Table 4. Accuracies of genomic estimated breeding values (GEBVs) from genomic BLUP (GBLUP), Random Forest (RF), Boosting and threshold BayesA (TBA) methods for 84% threshold point (standard deviation across 10 replicates in parentheses)

Model	SNP panels	Scenarios[a]			
		I	II	III	IV
GBLUP	2.5K	0.427 (0.06)	0.418 (0.07)	0.271 (0.07)	0.370 (0.06)
	12.5K	0.471 (0.05)	0.418 (0.07)	0.311 (0.05)	0.412 (0.06)
	25K	0.489 (0.06)	0.467 (0.06)	0.322 (0.07)	0.417 (0.06)
	50K	0.520 (0.06)	0.498 (0.05)	0.351 (0.07)	0.427 (0.06)
Boosting	2.5K	0.437 (0.04)	0.446 (0.04)	0.301 (0.05)	0.390 (0.06)
	12.5K	0.487 (0.05)	0.536 (0.06)	0.367 (0.06)	0.439 (0.04)
	25K	0.502 (0.04)	0.536 (0.05)	0.394 (0.04)	0.443 (0.05)
	50K	0.537 (0.04)	0.568 (0.03)	0.411 (0.04)	0.478 (0.03)
RF	2.5K	0.424 (0.05)	0.357 (0.03)	0.220 (0.04)	0.343 (0.03)
	12.5K	0.475 (0.03)	0.398 (0.05)	0.264 (0.03)	0.381 (0.04)
	25K	0.493 (0.04)	0.397 (0.03)	0.271 (0.04)	0.382 (0.03)
	50K	0.532 (0.04)	0.424 (0.04)	0.298 (0.04)	0.390 (0.04)
TBA	2.5K	0.420 (0.07)	0.465 (0.06)	0.316 (0.07)	0.396 (0.05)
	12.5K	0.469 (0.06)	0.519 (0.07)	0.356 (0.07)	0.435 (0.06)
	25K	0.479 (0.06)	0.534 (0.07)	0.363 (0.07)	0.441 (0.06)
	50K	0.493 (0.05)	0.562 (0.06)	0.405 (0.07)	0.456 (0.05)

[a]I (h^2 = 0.25, LD = low and 450 QTL), II (h^2 = 0.25, LD = low and 150 QTL), III (h^2 = 0.1, LD = low and 150 QTL) and IV (h^2 = 0.1, LD = high and 150 QTL).

Table 5. Accuracies of genomic estimated breeding values (GEBVs) from genomic BLUP (GBLUP), Random Forest (RF), Boosting and threshold BayesA (TBA) methods for 16 % threshold point (standard deviation across 10 replicates in parentheses)

Model	SNP panels	Scenarios[a]			
		I	II	III	IV
GBLUP	2.5K	0.459 (0.05)	0.448 (0.06)	0.316 (0.07)	0.442 (0.06)
	12.5K	0.509 (0.06)	0.494 (0.06)	0.372 (0.05)	0.481 (0.06)
	25K	0.520 (0.06)	0.502 (0.06)	0.388 (0.07)	0.486 (0.07)
	50K	0.542 (0.06)	0.524 (0.06)	0.415 (0.05)	0.502 (0.06)
Boosting	2.5K	0.529 (0.05)	0.515 (0.04)	0.415 (0.05)	0.449 (0.04)
	12.5K	0.573 (0.05)	0.589 (0.05)	0.469 (0.05)	0.504 (0.04)
	25K	0.587 (0.05)	0.601 (0.04)	0.473 (0.04)	0.519 (0.03)
	50K	0.605 (0.04)	0.619 (0.04)	0.495 (0.05)	0.539 (0.03)
RF	2.5K	0.496 (0.04)	0.427 (0.04)	0.273 (0.04)	0.364 (0.03)
	12.5K	0.552 (0.04)	0.471 (0.04)	0.324 (0.05)	0.418 (0.04)
	25K	0.562 (0.04)	0.483 (0.03)	0.338 (0.04)	0.429 (0.04)
	50K	0.587 (0.04)	0.504 (0.03)	0.368 (0.05)	0.455 (0.03)
TBA	2.5K	0.454 (0.06)	0.509 (0.06)	0.387 (0.07)	0.444 (0.06)
	12.5K	0.493 (0.06)	0.562 (0.06)	0.408 (0.06)	0.479 (0.06)
	25K	0.514 (0.07)	0.587 (0.07)	0.413 (0.06)	0.487 (0.06)
	50K	0.535 (0.06)	0.614 (0.06)	0.454 (0.07)	0.515 (0.06)

[a]I (h^2 = 0.25, LD = low and 450 QTL), II (h^2 = 0.25, LD = low and 150 QTL), III (h^2 = 0.1, LD = low and 150 QTL) and IV (h^2 = 0.1, LD = high and 150 QTL).

disease incidence in threshold traits and showed that an obvious decrease in the proportions of disease incidence resulted in approximately a loss of 30-40% in the genomic accuracy.

In the current study, a reduction in the number of sick individuals in the training set was associated with an increase in the accuracy of genomic prediction in all three models and was in agreement with results presented by Naderi *et al.* (2018) for disease traits in Holstein Friesian cow. In brief, these authors specified that correlation between pre-corrected phenotypes and genomic breeding values (rGBV) increased with the decrease in the percentage of sick cows in the training set from 37 to 20% for claw disorders, from 32 to 25% for clinical mastitis and from 29 to 19% for female infertility. One possible explanation for different reactions of different traits to the decreased percentage of sick individuals in the training sets is the different distributions of response variables. Generally, for binary traits as response variables, the optimal individuals training sets had disease incidences

close to the main population disease incidence (Naderi & Sadeghi, 2019).

Design and optimization of the training set, disease incidence rate and recording registration are among the most important factors affecting accuracies of genomic predictions in threshold traits. As a result, random assignment of a number of healthy individuals in the first strategy as sick in the second and third strategies leads to more individuals without considering their merit be encoded. Therefore, this theorem leads to more classification errors for binary phenotypes in these strategies. In conclusion, the prediction accuracy unintentionally decreased.

Original and imputed 50K SNPs panels

Tables 3, 4 and 5 show the genomic prediction results for different models using original and imputed 50K SNPs panels (with different proportions of missing genotypes including 95, 75 and 50%). For imputed scenarios, the highest accuracy was 0.619 using Boosting on imputed 50% genotypes for the first strategy. Also, the accuracy of genomic prediction for imputed 95% genotypes was the lowest (0.220) when RF was used for the third strategy. The higher sensitivity of machine learning methods on very sparse scenarios (2.5K panels) reduced with increase in imputation rate (12.5K and 25K panels).

Currently, nearly all Bayesian and GBLUP methods for genomic prediction are improved using imputed genotypes, such that many researchers recommend this strategy to decrease costs in animal breeding programs (Chen *et al.*, 2014; Wang *et al.*, 2016). In this study, the application of RF and Boosting comparing the above-mentioned methods was kept constant regarding the use of imputation. We observed that the accuracies of genomic prediction of machine learning methods were more sensitive to imputation errors. The results of this study are in line with the reports by Felipe *et al.* (2014), where different proportions of masking genotypes were used to evaluate accuracy of genomic prediction and showed that genotype imputation of low-density panels is not always helpful.

Number of QTL

To investigate the effect of the number of QTL on the genomic prediction of GBLUP, RF, Boosting and TBA methods under different rates of disease incidence, scenarios I (450 QTL) and II (150 QTL) were used (Tables 3 to 5). Boosting and RF method performed similarly, slightly better than GBLUP and TBA when the binary traits were affected by many QTL each with a small effect and considerably better than RF and GBLUP and similar to TBA when the binary traits were influenced by a few large QTL. The highest accuracy was identified for the scenario of 150 QTL when 16% of phenotypes were considered as sick.

The influence of the number of QTL on the genomic prediction accuracy depends on the statistical model (Sadeghi *et al.*, 2018). Despite the positive effect of decreasing the number of QTL on prediction of genomic accuracy via Boosting and especially TBA, and being in agreement with Ghafouri-Kesbi *et al.* (2017), the accuracy of genomic prediction via GBLUP and RF dropped with decrease in the number of QTL, as was previously shown by Naderi *et al.* (2016). Regarding prediction accuracies in scenarios I and II, Boosting outperformed other models although in some cases the differences were negligible. It seems that higher accuracy obtained from Boosting is due to the capability of this method to define interactions among markers by changing the tree depth parameter aimed at finding the value.

Heritability

To evaluate the effect of heritability on the genomic prediction of GBLUP, RF, Boosting and TBA methods under all the SNP types and different rates of disease incidence, scenarios II (h^2=0.25) and III (h^2=0.1) were used (Tables 3 to 5). Generally, both TBA and Boosting methods performed better than GBLUP and RF. The highest and least of prediction accuracies were observed for the first and third strategies, respectively. We recognized an increase in genomic accuracy with increase in heritability and a pronounced decrease in disease incidence rate, which was more obvious for RF. Furthermore, the accuracy of genomic predicted via RF, GBLUP, TBA and Boosting increased with the increase in heritability by a rate of 56.7, 51.5, 31.4 and 27.3%, respectively.

Hayes *et al.* (2009) showed that by increasing heritability from 0.1 to 0.9 the accuracy of genomic prediction increased from 0.3 to 0.7. Moreover, Daetwyler *et al.* (2013), through the accuracy formula $r=\sqrt{N_p h^2/N_p h^2 + M_e}$, showed that genomic prediction has a direct relationship with heritability. In this formula N_p is the number of individuals in the training, M_e the number of independent chromosome segments and h^2 represents the heritability of the target trait. Guo *et al.* (2014) indicated that the accuracy of genomic prediction improved by increasing genomic heritability (or less environmental noise) in training set, which is largely attributed to improved estimations of SNPs effects via presenting more genetic variation into sets.

LD structure

To evaluate the effect of LD structure on genomic prediction of different methods under all the SNP types

and different rates of disease incidence, scenarios III (r^2 = 0.229 at distances of 0.05 cM) and IV (r^2 = 0.417 at distances of 0.05 cM) were simulated (Tables 3 to 5). Results showed a gain in genomic accuracies with the increase in the level of LD in all of the models and increase in disease incidence rate.

In the present study, imputed scenarios were more sensitive than original scenarios to the LD variation. For example, for 84% threshold point to the training set, accuracy via RF, GBLUP, Boosting and TBA improved with increasing LD by a rate of 56, 36.2, 29.2 and 25.6% for 2.5K panel and 30.6, 21.6, 16.3 and 12.6% for original data, respectively. Moreover, LD affects the genomic prediction accuracy in imputed data in two ways: 1) direct effect on imputation accuracy, and 2) direct effect on the models' predictive ability. Not only high LD means that lower marker density cover the genome, but also higher collinearity among linked markers is required (Liu *et al.*, 2015). With regard to the high LD scenario, the highest accuracy was identified for 16% threshold point to the training set when applying the Boosting method. As the level of LD between SNP and QTL increases, the more markers capture a higher rate of the genetic variance (Goddard, 2009), a prerequisite for the efficient performance of machine learning methods.

Conclusions

In this study, the advantage of imputing genotypes was shown to be highly dependent on the number of SNPs available and the LD levels of the reference set. For GBLUP, RF, Boosting and TBA methods, the composition of disease incidence in training set was one of the major factors affecting the accuracy of genomic prediction. To achieve the highest prediction accuracy, optimal training set was characterized by 16% threshold point. Generally, for different genomic architectures, the Boosting outperformed the TBA, GBLUP and RF method under all SNPs panels and with different rates of disease incidence and the markers set being imputed. Looking at the change in genetic architecture of all scenarios, the RF resulted in a bigger reduction than Boosting, TBA, GBLUP, especially when the data set containing 2.5K panels was used. Therefore, due to their high sensitivity to imputation errors, the application of imputed genotypes using RF methods should be carefully evaluated.

References

Bishop CM, 2006. Pattern recognition and machine learning (information science and statistics). Springer-Verlag, NY.

Bohlouli M, Alijani S, Javaremi AN, König S, Yin T, 2017. Genomic prediction by considering genotype × environment interaction using different genomic architectures. Ann Anim Sci 17: 683-701. https://doi.org/10.1515/aoas-2016-0086

Breiman L, 2001. Random forests. Machine Learning 45: 5-32. https://doi.org/10.1023/A:1010933404324

Chen L, Li C, Sargolzaei M, Schenkel F, 2014. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS One 9: e101544. https://doi.org/10.1371/journal.pone.0101544

Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM, 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193: 347-365. https://doi.org/10.1534/genetics.112.147983

De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM, 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375-385. https://doi.org/10.1534/genetics.109.101501

Egger-Danner C, Cole J, Pryce J, Gengler N, Heringstad B, Bradley A, Stock KF, 2015. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal 9: 191-207. https://doi.org/10.1017/S1751731114002614

Felipe VP, Okut H, Gianola D, Silva MA, Rosa GJ, 2014. Effect of genotype imputation on genome-enabled prediction of complex traits: an empirical study with mice data. BMC Gent 15: 149. https://doi.org/10.1186/s12863-014-0149-9

Freund Y, Schapire RE, 1996. Experiments with a new boosting algorithm. Icml 96: 148-156. https://dl.acm.org/doi/10.5555/3091696.3091

Friedrich J, Antolín R, Edwards S, Sánchez-Molano E, Haskell M, Hickey J, Wiener P, 2018. Accuracy of genotype imputation in Labrador Retrievers. Anim Genet 49: 303-311. https://doi.org/10.1111/age.12677

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, 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.

Goddard M, 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245-257. https://doi.org/10.1071/AN15538

González-Recio O, Forni S, 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol 43: 7. https://doi.org/10.1186/1297-9686-43-7

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G, 2014. The impact of population structure on genomic prediction in stratified

populations. Theor Appl Genet 127: 749-762. https://doi.org/10.1007/s00122-013-2255-x

Hayes B, Daetwyler H, Bowman P, Moser G, Tier B, Crump R, Khatkar M, Raadsma H, Goddard M, 2009. Accuracy of genomic selection: comparing theory and results. Proc Assoc Advmt Anim Breed Genet, pp: 34-37.

Hickey JM, Crossa J, Babu R, de los Campos G, 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci 52: 654-663. https://doi.org/10.2135/cropsci2011.07.0358

Kabisch M, Hamann U, Bermejo JL, 2017. Imputation of missing genotypes within LD-blocks relying on the basic coalescent and beyond: consideration of population growth and structure. BMC genomics 18: 798. https://doi.org/10.1186/s12864-017-4208-2

Lakhssassi K, González-Recio O, 2017. A haplotype regression approach for genetic evaluation using sequences from the 1000 bull genomes Project. Span J Agric Res 15 (4): e0407. https://doi.org/10.5424/sjar/2017154-11736

Liu H, Zhou H, Wu Y, Li X, Zhao J, Zuo T, Zhang X, Zhang Y, Liu S, Shen Y, 2015. The impact of genetic relationship and linkage disequilibrium on genomic selection. PLoS One 10: e0132379. https://doi.org/10.1371/journal.pone.0132379

Madsen P, Jensen J, 2013. A users guide to DMU. A package for analysing multivariate mixed models, Version 6. Center for Quantitative Genetics and Genomics, University of Aarhus, Denmark. https://dmu.ghpc.au.dk/

Mc Hugh N, Meuwissen T, Cromie A, Sonesson A, 2011. Use of female information in dairy cattle genomic breeding programs. J Dairy Sci 94: 4109-4118. https://doi.org/10.3168/jds.2010-4016

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

Naderi S, Yin T, König S, 2016. Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. J Dairy Sci 99: 7261-7273. https://doi.org/10.3168/jds.2016-10887

Naderi S, Bohlouli M, Yin T, König S, 2018. Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets. Anim Genet 49: 178-192.

Naderi Y, Sadeghi S, 2019. Assessment of the genomic prediction accuracy of discrete traits with imputation of missing genotypes. Anim Sci Papers Rep 37: 149-168.

Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, Goddard ME, 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol 49: 24. https://doi.org/10.1186/s12711-017-0301-x

Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH, 2013. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet Sel Evol 45: 12. https://doi.org/10.1186/1297-9686-45-12

Sadeghi S, Rafat sA, Alijani S, 2018. Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods. Acta Sci Anim Sci 40: e39007. https://doi.org/10.4025/actascianimsci.v40i1.39007

Sargolzaei M, Schenkel FS, 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681. https://doi.org/10.1093/bioinformatics/btp045

Sargolzaei M, Chesnais J, Schenkel F, 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. J Dairy Sci 94: 421.

Su G, Madsen P, 2013. User's Guide for GMATRIX version 2, a program for computing genomic relationship matrix.

VanRaden PM, 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91: 4414-4423. https://doi.org/10.3168/jds.2007-0980

Ventura RV, Miller SP, Dodds KG, Auvray B, Lee M, Bixley M, Clarke SM, McEwan JC, 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet Sel Evol 48: 71. https://doi.org/10.1186/s12711-016-0244-7

Wang C, Ding X, Wang J, Liu J, Fu W, Zhang Z, Yin Z, Zhang Q, 2013. Bayesian methods for estimating GEBVs of threshold traits. Heredity 110: 213-219. https://doi.org/10.1038/hdy.2012.65

Wang Y, Lin G, Li C, Stothard P, 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Spr Sci Rev 4: 79-98. https://doi.org/10.1007/s40362-017-0041-x

Wang C, Li X, Qian R, Su G, Zhang Q, Ding X, 2017. Bayesian methods for jointly estimating genomic breeding values of one continuous and one threshold trait. PloS One 12: e0175448. https://doi.org/10.1371/journal.pone.0175448

Wang Q, Yu Y, Yuan J, Zhang X, Huang H, Li F, Xiang J, 2017. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. BMC Gent 18: 45. https://doi.org/10.1186/s12863-017-0507-5

Wientjes YC, Calus MP, Goddard ME, Hayes BJ, 2015. Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet Sel Evol 47: 42. https://doi.org/10.1186/s12711-015-0124-6

Wimmer V, Auinger HJ, Albrecht T, Schoen CC, 2015. Framework for the analysis of genomic prediction

data using R (synbreed). https://cran.rproject.org/web/packages/synbreed/index.html.

Yang P, Hwa Yang Y, B Zhou B, Y Zomaya A, 2010. A review of ensemble methods in bioinformatics. Curr Bioinform 5: 296-308. https://doi.org/10.2174/157489310794072508

Yin T, Pimentel E, Borstel UKv, König S, 2014. Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature× humidity-dependent covariate. J Dairy Sci 97: 2444-2454. https://doi.org/10.3168/jds.2013-7143