Acta Scientiarum. Animal Sciences
ISSN: 1807-8672
Editora da Universidade Estadual de Maringá - EDUEM

Sadeghi, Saadat; Rafat, Seyed Abbas; Alijani, Sadegh

Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods

Acta Scientiarum. Animal Sciences, vol. 40, e39007, 2018
Editora da Universidade Estadual de Maringá - EDUEM

DOI: https://doi.org/10.4025/actascianimsci.v40i1.39007

Available in: https://www.redalyc.org/articulo.oa?id=303158407042

Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8672
Doi: 10.4025/actascianimsci.v40i1.39007

ANIMAL BREEDING AND REPRODUCTION

Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods

Saadat Sadeghi*, Seyed Abbas Rafat and Sadegh Alijani

*Department of Animal Science, University of Tabriz, Zip code: 51666-14766, Tabriz, East Azarbaijan, Iran. *Author for correspondence. E-mail:saadat.sadeghi@tabrizu.ac.ir*

ABSTRACT. The objectives of this study were (1) to quantify imputation accuracy and to assess the factors affecting it; and (2) to evaluate the accuracy of threshold BayesA (TBA), Bayesian threshold LASSO (BTL) and random forest (RF) algorithms to analyze discrete traits. Genomic data were simulated to reflect variations in heritability (h^2 = 0.30 and 0.10), number of QTL (QTL = 81 and 810), number of SNP (10 K and 50 K) and linkage disequilibrium (LD=low and high) for 27 chromosomes. For real condition simulating, we randomly masked markers with 90% missing rate for each scenario; afterwards, hidden markers were imputed using FImpute software. In imputed genotypes, a wide range of accuracy was observed for RF (0.164-0.512) compared to TBA (0.283-0.469) and BTL (0.272-0.504). Comparing to original genotypes, using imputed genotypes decreased the average accuracy of genomic prediction about 0.0273 (range of 0.024 to 0.036). Comparing to Bayesian threshold, using RF was improved rapidly accuracy of genomic prediction with increase in the marker density. Despite the higher accuracy of BTL and TBA at different levels of LD and heritability, the increase in accuracy was greater for RF. Furthermore, the best method for prediction of genomic accuracy depends on genomic architecture of population.

Keyword: accuracy; genomic architecture; linkage disequilibrium; machine learning; masked genotypes.

Avaliação de dados genômicos imputados em características distintas usando os métodos de Random Forest e de limiares Bayesianos

RESUMO. Os objetivos deste estudo foram (1) quantificar a precisão de imputação e acessar os fatores que as afetam; e (2) avaliar a precisão do princípio de BayesA (TBA), do modelo Bayesiano LASSO (BTL), e o algoritmo Random Forest para analisar as características distintas. Dados genômicos foram simulados para indicar variações na herdabilidade (h^2 = 0.30 e 0.10), número de QTL (QTL = 81 e 810), número de SNP (10 k e 50 k) e desequilíbrio de ligação (LD = baixo e alto) para 27 cromossomos. Para uma simulação mais realista, nós cobrimos os marcadores aleatoriamente com 90% da taxa ausente para cada cenário, depois, os marcadores foram imputados usando o *software* FImpute. Nos genótipos imputados uma grande oscilação de precisão foi observada pelo modelo RF (0.164-0.512) comparado com TBA (0.283 - 0.469) e BTL (0.272 - 0.504). Comparando com os genótipos originais, os genótipos imputados decaíram a precisão média da predição genômica em cerca de 0.0273 (oscilação de 0.024 para 0.036). Comparando-se ao princípio Bayesiano, o uso de RF melhorou a precisão de predição com o aumento da densidade do marcador. Além disso, o melhor método para predição de precisão genômica depende da arquitetura genômica da sua população.

Palavras-chave: precisão; arquitetura genômica; desequilíbrio de ligação aprendizado maquinal; genótipos mascarados.

Introduction

Genomic selection (GS) plays an important role to estimate genomic breeding values (GEBVs) of continuous traits that follow approximately a Gaussian phenotypic distribution in livestock (Meuwissen, Hayes, & Goddard, 2001). However, some traits for instance, litter size, degree of calving difficulty and resistance to disease are the most prominent traits in animal breeding that often termed discrete traits and present a categorical distribution of phenotypes, where current livestock breeding programs are aiming at including discrete traits that reflect animal health, behavior, and product quality (König, Brügemann, & Pimentel, 2013). Discrete traits are influenced by multiple genes and deviate from Mendelian inheritance (Blazer & Hernandez, 2006). Obviously, the focused GS methods on continuous traits cannot be adequately useful for these traits (Wang et al., 2013). Hence, GS methods must be adapted to cope with challenges of discrete traits. Therefore, threshold versions of Bayesian regressions and machine

learning methods are applied for genomic prediction such kind of traits analyses (González-Recio & Forni, 2011). Machine-learning methods are improving predictive ability in repeated observation. In discrete traits, methods such as random forest algorithm (Breiman, 2001) could help to achieve high genomic accuracy for human (Sun et al., 2008) and livestock (Chen, Li, Sargolzaei, & Schenkel, 2014; Nguyen, Huang, Wu, Nguyen, & Li, 2015). Therefore, using these methods to include discrete traits in animal breeding schemes could increase accuracy of prediction and consequently, it results in higher genetic gain.

Marker density is one of the most important factors in order to achieve appropriate accuracy of genomic prediction. However, the economic aspect of genotyping should not be ignored. Nowadays, animal breeding researchers are trying to genotype more individuals with low-density chips and obtain the remaining genotypes through imputation from a higher density panel (Toghiani, Aggrey, & Rekaya, 2016). However, re-sequencing all individuals by the high density chip is not cost-effective. The technique known imputation allows researchers to have more accurate estimate of association evidence at genetic single nucleotide polymorphisms (SNPs) that are not directly genotyped (Li, Willer, Sanna, & Abecasis, 2009). For more detections of genes associated with discrete traits, genotypes imputation is more affordable compared to whole-genome sequencing at current prices (Yang et al., 2015). FImpute (Sargolzaei, Chesnais, & Schenkel, 2011) is way to impute missing genotypes based on pedigree information and linkage information, which was developed for animal applications (Toghiani et al., 2016).

In addition to marker density from low to high SNP chip, other factors such as reference population size, genetic relationships among genotyped individuals and the animals to be imputed and level of linkage disequilibrium (LD) have impact on accuracy of genotype imputation (Hickey, Crossa, Babu, & de los Campos, 2012). Many studies (Badke, Bates, Ernst, Fix, & Steibel, 2014; Sargolzaei, Chesnais, & Schenkel, 2014) were carried out to evaluate the efficiency of SNP genotype imputation under different architecture emphasizing on the accuracy of imputed genotypes. According to literature, little attention has been paid to the imputation accuracy and its impact on the quality of genomic accuracy.

Furthermore, the accuracy of genotype prediction is also depended on other factors related to population structure and genetic architecture, such as size of the reference data set (VanRaden & Sullivan, 2010), trait heritability (Guo et al., 2014), markers density (Meuwissen, 2009), the number of loci affecting the trait (Daetwyler, Villanueva, & Woolliams, 2008) and LD (Yin, Pimentel, Borstel, & König, 2014).

In GS, simulation allows researchers to discover the influences of the genetic architecture of the trait, the number of markers used for analysis, and the data also allows for evaluating some sources of variability, such as drift, which cannot be assessed with the most of real data (Daetwyler et al., 2010). In this respect, the simulation study can be carried out to investigate the advantage of the threshold methods in terms of accuracy with the GEBVs of discrete traits with considering different aspects of genomic structures. Therefore, the objective of this study was to compare the accuracy of genomic predictions using threshold Bayes A, Bayesian threshold LASSO and RF for simulated binary traits by altering heritability, number of QTL, marker density, and the LD structure of the genotyped population when original (before masking a proportion of SNPs) and imputed genotypes were used.

Material and methods

Simulation of population

The simulation was implemented using the QMSim software (Sargolzaei & Schenkel, 2009) to generate phenotypes, genotypes and true breeding values using the following parameters: at first, during 1000 generations, a historical population was provided from 10000 females and 200 males in order to produce a realistic level of LD for the platform. Bottleneck was used to create a population with a higher level of LD. However, we initiated the same simulation process, but after 1000 generations, the population size decreased over 100 generations to 400 individuals. Afterward, the population size was increased over 100 generations. Then, 10,000 females and 400 males from the last historical population were selected. In the second step, all individuals from the last generation of the historical population served as founders in the recent population. Using a random mating design, the recent population was expanded by simulating an additional 10 generations. Per mating produced only one offspring with a same probability of being each sex. Replacement rates were 80 and 20 percent for males and females, respectively.

Selection for both sexes was based on estimated breeding values. Biallelic SNP markers were evenly placed along 27 chromosomes of sheep, each 100

cM long. Simulations of 370 and 1,850 biallelic markers per chromosome depicted applications with 9,990 SNP (10 K chip) and 49,950 SNP (50 K chip), respectively. For each marker density, two different numbers of QTL (either 3 or 30 QTL on each chromosome) affected the trait. A gamma distribution was sampled for QTL effects with a shape parameter of 0.4. For each locus and generation, the mutation rate was fixed on 2.5×10^{-5} for all of SNPs and QTLs. Moreover, the total amount of additive-genetic variance was ascribed to the QTL. We considered two levels of heritability (low = 0.1 and moderate = 0.3). More explanation for parameters is summarized in Table 1. There were eight scenarios (I to VIII; Table 2) to reflect variations regarding too number of QTL, heritability, level of LD and number of markers. To create a binary phenotype, we defined code 1 as diseased and code 0 as healthy depending on whether simulated phenotype was lower or higher of the population phenotype mean, respectively. We performed 10 replicates for each scenario to evaluate the models.

Table 1. Parameters of the simulation process.

Parameter	Low linkage disequilibrium	High linkage disequilibrium
Historical population		
No. of generations (population size) in phase 1	1,000 (10,400)	1,000 (10,400)
No. of generations (population size) in phase 2	-	100 (400)
No. of generation (population size) in phase 3	-	200 (10,400)
Recent population		
No. of founder sires (dams)	400 (10,000)	
No. of generations	10	
No. of offspring per dam	1	
Mating system	Random	
Replacement ratio for males (females)	0.8 (0.2)	
Criteria for selection/culling	EBV/age	
Sex probability for offspring	0.5	
Genome		
No. of chromosomes	27	
Total length of chromosomes (cM)	2,700	
Marker distribution	Evenly spaced	
No. of QTL alleles	Random (2, 3, or 4)	
Effects of QTL alleles	Gamma (0.4)	
Marker and QTL mutation rate	$2.5\ 10^{-5}$	
Position of marker and QTL	Random	
No. of QTL	81 or 810	
No. of markers	9990 or 49950	
Heritability of the trait	0.1 or 0.3	

Table 2. The simulated scenarios (I to VIII) with respect to the number of markers and QTL, the heritability of the trait and the level of linkage disequilibrium.

	Scenarios							
Variable	I	II	III	IV	V	VI	VII	VIII
h^2	0.3	0.3	0.1	0.1	0.3	0.3	0.1	0.1
No. of QTL	810	81	81	81	810	81	81	81
No. of SNP	10 k	10 k	10 k	10 k	50 k	50 k	50 k	50 k
Level of linkage disequilibrium	low	low	low	high	low	low	low	high

Calculation of linkage disequilibrium

The level of LD in the simulated scenarios was assessed by calculating the squared correlation coefficient (r^2) between all possible pairs of markers according to Hill and Robertson (1968):

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

where, D = f(AB)-f(A)f(B), and f(AB), f(A), f(a), f(B), f(b) are observed frequencies of haplotypes AB and of alleles A, a, B, b, respectively. The *PLINK* software (Purcell et al., 2007) was used to estimate LD between marker pairs in the last generation.

Imputation

To simulate a real condition, we randomly masked a major proportion of markers (90%) in low-density SNP platform (10 K) and medium-density SNP platform (50 K); afterwards, masked markers were imputed by considering a family and population-based algorithm with FImpute program (Sargolzaei et al., 2011). The FImpute software uses a deterministic approach that combines family and population imputation methods. The population imputation method is based on the assumption that all individuals have some degree of relationship and share haplotypes that may differ in frequency and length depending on the relationships. Imputation by FImpute is a two-step procedure, *i.e.* first it searches for long haplotypes by applying a family imputation method, and second, it identifies short segments (two SNPs) by applying a population imputation method that analyzes overlapping sliding windows. FImpute uses deterministic methods to infer missing or un-typed marker genotypes. FImpute offers the option to impute genotypes based on Mendelian inheritance and segregation rules without using population information.

Accuracy of imputation (per SNP in all chromosomes) was assessed by correlation between imputed and original genotypes for all replications as an appropriate approach to minimize the dependency on allele frequency. It estimates the ability of a linear model to depict the relationship between two variables. These imputed genotype probabilities, one for each genotype class (e.g. AA, AB, or BB), are transformed to dosage values by multiplying by 0, 1 or 2 for each genotypic class.

Genome-enabled evaluation models

To estimate genomic breeding values, we applied three different evaluation models (two linear regressions using a Bayesian framework (Threshold

Bayes A and LASSO), and one machine-learning ensemble algorithms (Random forest).

Model 1: Threshold Bayes A (TBA)

Meuwissen et al. (2001) had proposed Bayesian regressions on the genomic markers. We utilized TBA as proposed González-Recio and Forni (2011). Wright (1934) postulates an underlying random variable, called liability (λ) that followes a continuous distribution, and that the observed dichotomy is the result of the position of the liability with respect to a fixed threshold (t):

$$\text{Phenotype} = \begin{cases} 0 \text{ if } t > \lambda \\ 1 \text{ if } t \leq \lambda \end{cases}$$

where, λ is taken as the response variable. The suggested change consists of the linear regression of the single nucleotide polymorphism (SNP) coefficients on a liability variable with Gaussian distribution. The TBA can be described as follows:

$$\lambda = \mu 1 + \mathbf{X}b + e$$

where, the underlying liability variable vector for y is λ, μ is the population mean, column vector (n×1) of ones is 1; b indicates (bj) the vector for the regression coefficient estimates of the *p* markers assumed normally and independently distributed a priori as N (0, σ_j^2), which σ_j^2 is assumes to an unknown variance related with SNP *j*. The scaled inverse chi square $\sigma_j^2 \sim \upsilon_j s^2_j \chi^{-1}_{\upsilon_j}$ with $\upsilon_j = 4$ and $s^2_j =$ 0 002 assume for prior distribution of σ_j^2. Elements of the incidence matrix **X**, of order $n \times p$, may be set up as for different additive, dominant or epistatic models. In the more practical scenario, it takes values -1, 0 or 1 for marker genotypes *aa, Aa* and *AA*, respectively. The residuals (e) are assumed to be distributed as N ($\mu = 0$, $\sigma_e^2 = 1$), as stated above. As in a regular threshold model, threshold and the residual variance have to be set fixed (0 and 1, respectively) since these parameters are not identifiable in a liability model.

This method can be solved through the Gibbs sampler described in Meuwissen et al. (2001), with the simple incorporation of the data augmentation algorithm to sample the individual liabilities from their corresponding truncated normal distribution as described in Tanner and Wong (1987). The joint posterior distribution of the n liabilities is:

$$\text{Prob}(\lambda|\mu, b, t) = \prod_{i=1}^{n} \{ \frac{\Phi[t - (\mu + x_i b)]}{\sigma_e} \}^{1-y_i} \{1 - \frac{\Phi[t - (\mu + x_i b)]}{\sigma_e} \}^{y_i}$$

Model 2: Threshold Bayesian LASSO (BTL)

BTL was described by Park and Casella (2008), afterwards, De Los Campos et al. (2009) has been applied BTL genomic version for continuous traits, and furthermore, González-Recio, Maturana, Vega, Engelman, and Broman (2009) extended for binary traits. This methodology considers a Laplace (double exponential) prior distribution on the markers effects. BTL depends on shrinkage parameters over the distribution of the effects of marker. As stated in the previous model, the response variable is a liability response (λ) that follows a continuous distribution. BTL can be solved as:

$$\lambda = \mu 1 + \mathbf{X}\hat{\beta} + e$$

where λ represents the vector of liabilities for all individuals, μ is the average of population, 1 shows a column vector (n × 1) of ones; $\hat{\beta}$ are the LASSO estimates with their respective incidence matrix **X** as described for model TBA. The residuals (e) were considered the vector of independently and identically distributed residuals, as N (0, σ_e^2). As described for model TBA, the threshold and the residual variance fixed 0 and 1 respectively; alternate choices result in the same model.

In a fully Bayesian context, the LASSO estimates ($\hat{\beta}$) can be interpreted as posterior modes estimates when the regression parameters have independent and identical double-exponential priors (Tibshirani, 1996). Park and Casella (2008) have proposed a conditional Laplace prior specification for the LASSO estimates of the form:

$$P(\beta|\sigma_e^2) = \prod_{i=1}^{n} \frac{\gamma}{2\sqrt{\sigma_e^2}} e^{-\gamma|\beta_j|/\sqrt{\sigma_e^2}}$$

where σ_e^2 s the residual variance, and γ is a parameter controlling the shrinkage of the distribution. Inferences about γ may be done in different ways (Park & Casella, 2008). To follow the Bayesian specifications, a gamma prior is proposed here for γ^2, with known rate (r) and shape (δ) hyper-parameters, as described by De Los Campos et al. (2009). Samples from posterior distributions of those estimates are drawn from the Gibbs sampling algorithm as described by De Los Campos et al. (2009), with the corresponding data augmentation algorithm for liabilities, as described for TBA.

Model 3: Random Forest (RF)

One of the machine learning ensemble algorithms is RF which was first proposed by

Breiman (2001). González-Recio and Forni (2011) used the java package RanFoG for RF analyses in GS of discrete traits. RF was also explored for genome-wide association studies by Li et al. (2014) and Nguyen et al. (2015). This algorithm is strongly non-parametric, powerful to over fitting, able to capture complex interaction structures in the data, which may alleviate the problems of analyzing genome-wide data. In validation data many classification trees were constructed by bootstrapping (Efron & Tibshirani, 1994) in the RF analysis. RF uses bagging strategy and reduces error prediction.

The RF prediction for an observation, $\hat{f}_{rf}^{P}(x)$, is computed by averaging the predictions over P trees, $(T(x, \Psi p))_1^P$, for which the given observation was not used to build the tree and Ψp characterises the p_{th} RF tree in terms of split variables, cut points at each node, and terminal node values. The RF framework was used in the following model:

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

RF used on mean almost two-thirds of the data and a random subset p of the m SNP ($p \sim 2/3 \times m$) for the construction of each tree. Animals not included in the bootstrapped sample were defined as "out of bag", being the validation set for each tree. At each node, data were split in 2 branches based on the genotype at SNPj by minimizing a loss function for classification. Repetition of this procedure implied a large number of trees i.e., RF, until the convergence criterion was achieved. The convergence criterion used classification errors of out of bag samples. In current study, 2,000 and 5,000 trees were constructed for 10K and 50 K SNP chips, respectively. Random sampling of the data contributed to the formation of de-correlated trees. Each tree reflected the most frequent outcome for a given combination of SNP genotypes. The average of the predicted value of each tree was the probability of being susceptible to the disease.

Prediction accuracy

Predicted accuracy was calculated by phi-correlation coefficient between the true BVs and the genomic predicted BVs ($r_{p,t}$) or genomic imputed BVs ($r_{i,t}$) for all scenarios per the testing set. Analysis of variance was performed to investigate the different effects of method, heritability, LD, QTL and marker density for the accuracy using the R software. Figure 1 shows all operation steps that are applied at current research.

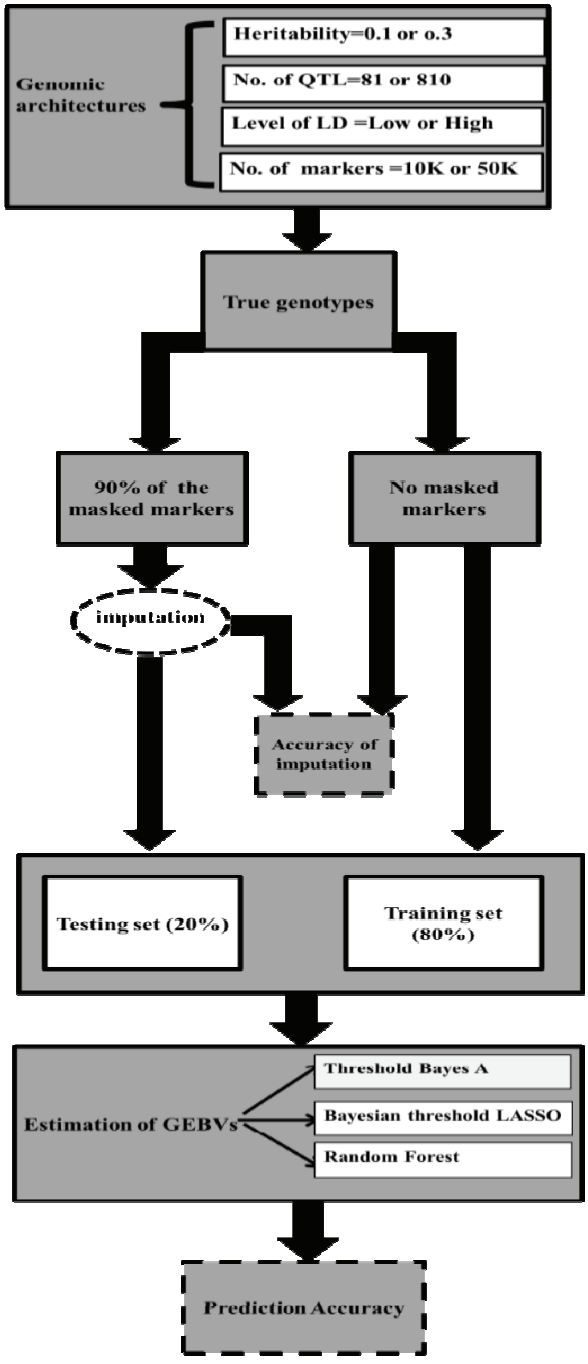

Figure 1. Schematic of the whole process from simulated scenarios to the prediction accuracy.

Results and discussion

The first purpose of this study was to investigate the accuracy of imputation in simulated data with 1K and 5 K SNPs up to 10 K and 50 K SNPs, respectively, based on different patterns of genomic architecture.

Effect of genomic architecture on imputation accuracy

For different patterns of genomic architectures, the box-plots of correlation between imputed and

original genotypes are shown in Figure 2. Imputation accuracies show numerically for different genomic architectures in Table 3. The accuracy of imputation ranged from 0.929 to 0.979. The average of imputation accuracy increased by 2.34% from 10 to 50 K scenarios. The accuracies were 0.931 and 0.956 for low (III) and high LD (IV) scenarios, respectively. The highest accuracy of imputation was belonged to VIII scenario. Since the average accuracy of imputation was lower when the sparse panels (1K SNPs) were used, it seems that use the 5 K chip could be a good choice to improve imputation accuracy. The within-breed accuracy of imputation had ranged from 0.578 to 0.854 when markers were imputed from 5 to 50 K SNPs for Romney sheep Ventura et al. (2016). Previous studies showed that a 7 K marker panel can give better accuracy than a 3 K SNP panel to reach a 50 K marker panel (Boichard et al., 2012; Dassonneville, Fritz, Ducrocq, & Boichard, 2012). While, based on Toghiani et al. (2016) results, imputation of genotypes from 3 K panel to HD panels leads to acceptable results.

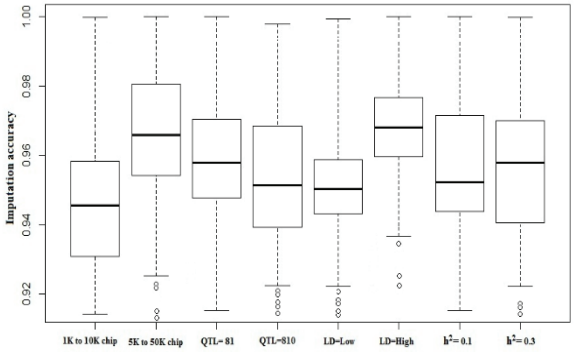

Figure 2. The box-plots of correlation between imputed and original genotype for the main effects.

Table 1. Mean and standard deviation (in bracket) of correlation between imputed and observed genotypes by scenarios.

Scenarios	Correlation between imputed and observed genotypes
I (10K SNP, h2 = 0.30, 810 QTL and LD=Low)	0.929 (0.012)
II (10K SNP, h2 = 0.30, 81 QTL and LD=Low)	0.932 (0.011)
III(10K SNP, h2 = 0.10, 81 QTL and LD=Low)	0.931 (0.011)
IV (10K SNP, h2 = 0.10, 81 QTL and LD=High)	0.956 (0.010)
V (50K SNP, h2 = 0.30, 810 QTL and LD=Low)	0.949 (0.012)
VI (50K SNP, h2 = 0.30, 81 QTL and LD=Low)	0.953 (0.011)
VII (50K SNP, h2 = 0.10, 81 QTL and LD=Low)	0.955 (0.012)
VIII(50K SNP, h2 = 0.10, 81 QTL and LD=High)	0.979 (0.010)

To infer masked genotypes, imputation methods depend partially on density and LD among markers. Nevertheless, factors affecting the accuracy of genotype imputation obtained in this study are comparable to reports published by Hickey et al. (2012) in maize and Khatkar, Moser, Hayes, and Raadsma (2012) in Australian Holstein-Friesian cattle, Mulder, Calus, Druet, and Schrooten (2012) in Dutch Holstein cattle, Pausch et al. (2013) in the Fleckvieh cattle, Badke et al. (2014) in Yorkshire boar, Boison et al. (2014) in simulated population of Brazilian Nellore cattle, Ogawa et al. (2016) in Japanese Black cattle and Pausch et al. (2017) in Fleckvieh and Holstein cattle. However, differences in level of LD showed limited effects on imputation accuracy in French cattle breeds (Hozé et al., 2013). According to Ogawa et al. (2016) studies, an increase on accuracy of imputation was observed with increasing the density of markers. Carvalheiro et al. (2014) was evaluated genomic-imputation accuracy for different low density chips in Nellore cattle. To predict 99.1% missing rate, they obtained high imputation accuracy (0.925) using 7 K SNPs chips. As previously reported in the literature (Van Raden et al., 2013), some regions of the genome have less than 0.60 imputation accuracy. A more careful analysis revealed that these regions contain very low levels of LD between markers, which emphasize the role of LD on imputation accuracy.

Accuracy of genomic prediction

The second aim of the study was to investigate the effect of different genomic architectures, accuracy of imputation and also to compare RF, TBA and BTL models on the accuracy of genomic prediction in imputed and original genotypes.

Effect of genotype imputation on accuracy of genomic prediction

Table 4 presents the accuracies of estimated GEBVs using original and imputed genotypes (with the 90 % missing rate) via RF, TBA and BTL models.

In all scenarios, little mean difference on the accuracies was evident, when original genotypes and imputed genotypes were compared. Due to low imputation accuracy, the decay of genomic prediction accuracies was higher in low density scenarios in comparison with medium density scenarios; this decline was 8.56 and 6.95% for 10 and 50 K SNPs chip, respectively. These results show that improvement in accuracy of genomic prediction is mainly, due to the increase in markers density and imputation accuracy. Toghiani et al. (2016) reported the accuracies of estimated GEBVs for the true and imputed SNP genotypes (with the 92.86% missing

rate) via BayesA method and concluded that accuracy was higher for the true SNP genotypes. Comparing to original genotype, the imputed 50 K SNP genotypes reduced accuracies of genomic perdition by 0.6 using Bayesian methods. Also, including genotypes imputed from the 6 K panel achieved almost the same accuracy of genomic prediction as that of the 50 K panel (Chen et al., 2014). It is obvious that (i) imputation accuracy has a large influence on the accuracy of GEBVs (Wang, Lin, Li, & Stothard, 2016), (ii) Mulder et al. (2012) after using a deterministic equation, concluded that accuracy of GEBV increased linearly with increase in imputation accuracy. Pimentel, Edel, Emmerling, and Götz (2015) showed that performance of genomic accuracy was influenced by imputation errors. Decrease of genomic accuracy based on imputed genotypes in previous published results (Cleveland & Hickey, 2013) acknowledges our results. Comparing genomic accuracy through genotype imputation, Badke et al. (2014) reported no difference among accuracy of genomic prediction when markers had imputed with high accuracy (R^2 = 0.95) instead of true genotypes. However, accuracy of genomic evaluation significantly decreased when genotypes were imputed with lower accuracy (R^2 = 0.88).

Effect of marker density

Accuracy of genomic prediction for original genotypes ($r_{p,t}$) and imputed genotypes ($r_{i,t}$) is shown in Table 4. For the low-density 10K SNP panels, the total average of genomic prediction accuracy for imputed genotypes were 0.246, 0.362 and 0.366 using RF, TBA and BTL, respectively (Figure 3). In addition, application of original genotypes increased accuracy 11.7, 6.89% and 7.03 % for RF, TBA and BTL, respectively. According to the results, a small absolute improvement (0.023 to 0.033 and on average 0.0265 across all scenarios) in prediction accuracy has been seen when prediction was based on original genotypes. There was more increase in prediction accuracy (0.027 to 0.033 and on average 0.029) for RF. Generally, prediction accuracies from RF always underperformed those from corresponding TBA and BTL methods; and standard deviations from RF were more homogeneous compared with Bayesian threshold methods for both genotype sets. González-Recio and Forni (2011) with simulation of 10 K SNP chips for a binary trait observed that accuracies ranging from 0.30 to 0.36 for RF, 0.26 to 0.32 for TBA and 0.33 to 0.35 for BTL. Similarly, Naderi, Yin, and König (2016) simulated different scenarios to investigate the performance of RF and GBLUP.

In the case of the medium-density panel, a wide range of accuracy was observed for RF comparing to TBA and BTL for both imputed and original genotypes (Figure 3). The later findings showed that accuracy improvement was more obvious for RF ($r_{i,t}$=52.8% and $r_{p,t}$= 47.6%) than TBA ($r_{i,t}$=12.0% $r_{p,t}$= 12.01%) and BTL ($r_{i,t}$=17.6 % and $r_{p,t}$=16.3%) by the increase of marker density. When Bayes A was used, imputed genotypes (3K SNP to 42K SNP panel) had decreased the accuracy of genomic prediction by 12.8% ($r_{i,t}$=0.528 and $r_{p,t}$= 0.596) in comparison with true genotypes (Toghiani et al., 2016). Naderi et al. (2016) reported the range of 0.30 to 0.53 using RF for scenarios with similar genomic architecture in simulated genotypes. In contrast to with our current study, Spindel et al. (2015) reported that with increase in marker density, RF has more accuracy than Bayesian regression methods in rice. According to other study Wang, Li et al. (2017), increase in marker densities generally resulted in raised accuracy predicted by Bayes A and Bayesian LASSO. In human for height trait, accuracy of genomic prediction improved rapidly with increase of marker density (approximately 150,000 markers), while plateaued at between 200,000 and 400,000 markers (Desta & Ortiz, 2014). As in GS, all genetic variance is described by the markers which are distributed in the whole genome; the predictive ability of GEBVs is deeply dependent with on marker density (Bo et al., 2017; Wang, Yu et al., 2017). Generally, because of increasing marker density, the level of LD among QTL and SNPs increased and then the accuracy of genomic prediction improved.

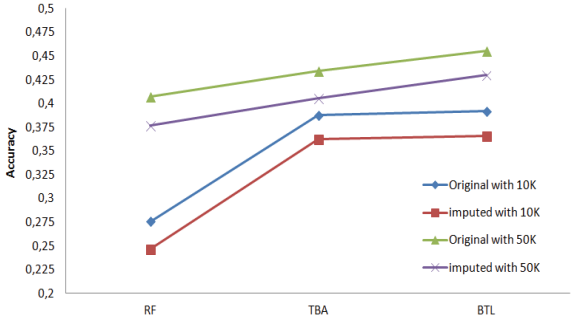

Figure 3. Effect of different marker density on accuracies of GEBVs estimated by threshold BayesA (TBA), Bayesian threshold LASSO (BTL) and random forest (RF) for original and imputed genotypes.

Table 4. The accuracies of estimated GEBVs using the original and imputed SNP genotypes from RF, TBA, BTL models (values in parentheses show the SD from 10 replicates).

Scenarios	$r_{i,t}$			$r_{p,t}$		
	RF	TBA	BTL	RF	TBA	BTL
I (10K SNP, h2 = 0.30, 810 QTL and LD=Low)	0.284(0.01)	0.408(0.04)	0.443(0.03)	0.311(0.01)	0.433(0.03)	0.467(0.02)
II (10K SNP, h2 = 0.30, 81 QTL and LD=Low)	0.316(0.02)	0.445(0.03)	0.458(0.03)	0.349(0.02)	0.478(0.03)	0.484(0.04)
III(10K SNP, h2 = 0.10, 81 QTL and LD=Low)	0.164(0.02)	0.283(0.02)	0.272(0.03)	0.191(0.02)	0.306(0.03)	0.297(0.03)
IV (10K SNP, h2 = 0.10, 81 QTL and LD=High)	0.223(0.01)	0.314(0.04)	0.291(0.03)	0.252(0.02)	0.343(0.04)	0.319(0.04)
V (50K SNP, h2 = 0.30, 810 QTL and LD=Low)	0.512(0.02)	0.469(0.08)	0.504(0.07)	0.548(0.02)	0.494(0.06)	0.531(0.07)
VI (50K SNP, h2 = 0.30, 81 QTL and LD=Low)	0.397(0.03)	0.457(0.06)	0.482(0.06)	0.430(0.04)	0.488(0.06)	0.508(0.06)
VII (50K SNP, h2 = 0.10, 81 QTL and LD=Low)	0.244(0.03)	0.325(0.03)	0.342(0.04)	0.271(0.02)	0.354(0.03)	0.366(0.03)
VIII(50K SNP, h2 = 0.10, 81 QTL and LD=High)	0.355(0.04)	0.373(0.07)	0.394(0.08)	0.380(0.04)	0.401(0.06)	0.418(0.07)

RF=Random forest; TBA=Threshold BayesA; BTL=Bayesian threshold LASS.

Effect of the number of QTL

For the low-density panel, the accuracy of genomic prediction was evaluated for scenarios with identical architecture, except for two different QTL numbers [i.e., scenario I (81 QTL) vs. II (810 QTL)] from RF, TBA and BTL in imputed and original genotypes (Table 4 and Figure 4). Under scenarios of 81 or 810 QTL, TBA and BTL methods showed better accuracy than RF. Significant difference between accuracies of Bayesian regressions and RF were found in both imputed and original genotypes. By the decrease of QTL numbers in imputed and original genotypes, accuracy of TBA improved more than RF; it seems a few large QTL that affects scenarios are reason for higher accuracy of BayesA methods (Hayes, Bowman, Chamberlain, & Goddard, 2009). Ghafouri-Kesbi, Rahimi-Mianji, Honarvar, and Nejati-Javaremi (2017) results, increase in QTL number have an inverse minor effect on accuracy of genomic prediction.

In the case of the medium-density panel, with considering 0.30 heritability in imputed and original genotypes, accuracies of GEBVs have been assessed for TBA, BTL and RF in scenarios V (810 QTL) and VI (81 QTL) (Figure 4). In the VI scenario, BTL had better performance, whereas higher accuracy was belonged to V scenario in RF. In contrast with low-density, with increase in number of QTLs, accuracies were partially higher for TBA. In current study, increasing the number of QTLs had negligible effect on genomic prediction accuracy for TBA and BTL methods, while for RF, a significant effect was found. Using Bayesian regressions and LASSO methods, Coster, Bastiaansen, Calus, Van Arendonk, and Bovenhuis (2010) showed that high accuracies could be achieved when the number of QTLs decreased, while accuracy of partial least square regression was unaffected. Abdollahi-Arpanahi, Peñagaricano, Aliloo, Ghiasi, and Urioste (2013) simulated a trait with different QTL levels and observed that with increasing the number of QTLs, accuracy was decreased. When number of QTL increased, the total genetic variance was divided among more QTL, therefore, the efficiency of methods decreased for estimating such small QTL effects. The same result was reported by Wientjes et al. (2015). With higher number of QTL, greater accuracies were reported with Bayesian regression comparing to machine learning methods (González-Recio & Forni, 2011). At constant heritability (h^2=0.3) and high-density SNP platforms, GBLUP was insensitive to genetic architecture (i.e., the number of QTL), while the accuracy of RF method improved as the number of QTL increased (Naderi et al., 2016). Different number of simulated chromosomes (Daetwyler et al., 2010), effective population sizes (Andonov et al., 2017) and architectures (Ghafouri-Kesbi et al., 2017) might be reasons for inconsistency of earlier finding with our results. With increase in both QTL and marker numbers, accuracy could be impact more by application of RF than other methods. Generally, the higher sensitivity of RF on QTL alterations than Bayesian threshold methods can be explained by RF based on a sampling technique for predictors (SNP). Therefore, by applying 50 K chip combined with 810 QTLs, SNPs in close distance to a QTL were sufficiently sampled.

Effect of heritability

For the low-density panel, the effect of different heritability levels on accuracy of genomic prediction in imputed and original genotypes is represented in Table 4 and Figure 5 (scenarios II and III). With increase in heritability, we recognized an evident increase on accuracy; as this increase was more pronounced for Bayesian threshold methods than for RF in both genotypes. Our results are in accordance to Bo et al. (2017) theory concerning direct relationship between heritability and accuracy of genomic prediction. Furthermore, Neves, Carvalheiro, and Queiroz (2012) compared different methods for evaluation of mice population with a wide range of heritability (0.16 – 0.89) on accuracy of genomic prediction and did not find any significant differences among these methods.

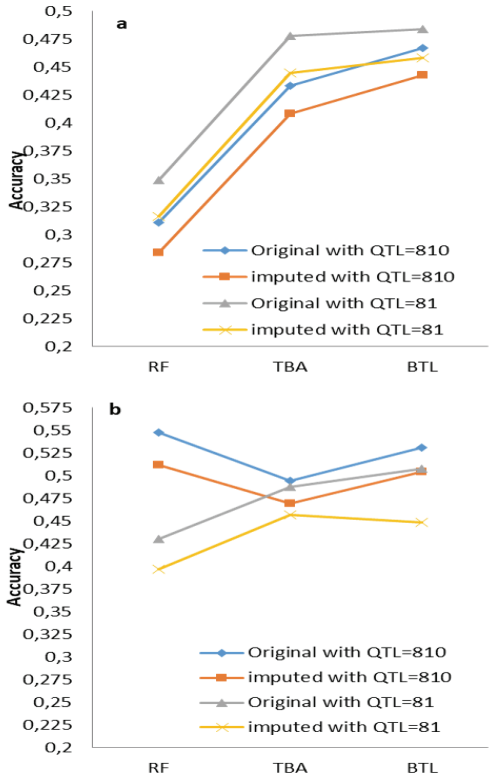

Figure 4. Effect of different number of QTL on accuracies of GEBVs estimated by threshold BayesA (TBA), Bayesian threshold LASSO (BTL) and random forest (RF) in original and imputed genotypes for 10 K (a) and 50 K (b) SNP panels.

In the case of the medium-density panel, the accuracy of genomic prediction in imputed and original genotypes was evaluated for different heritability levels [i.e., scenarios VI (h^2 = 0.3) vs. VII (h^2 = 0.1)] by RF, TBA and BTL methods (Tale 4 and Figure 5). As was expected, accuracy improvement for both genotypes was accompanied by the growth in heritability. Whereas, the increase of heritability had stronger effect on accuracy of RF; nonetheless, BTL in imputed ($r_{i,t}$=0.482) and original ($r_{p,t}$=0.508) genotypes had better performance than other methods. In several previous studies (Atefi, Shadparvar, & Hossein-Zadeh, 2016; Wang, Li et al., 2017), the profitable effects of increasing heritability on accuracy of genomic prediction has been proved by Bayesian model. These positive effects may be result of higher genetic variations and contributing to accurate predictions of marker effects.

Effect of LD structure

For the low-density panel, we presented the pattern of LD different structures [i.e., scenario III (LD = low) vs. IV (LD = high)] on accuracy of genomic prediction according to RF, TBA and BTL in imputed and original genotypes (Tale 4 and Figure 6).

The average LD (r^2) for III scenario and IV scenario were 0.175 and 0.323, respectively, at distances of 0.05 cM. Increasing level of LD had obvious effects on improvement of accuracy for RF in imputed (35.9%) and original (31.9%) genotypes. Nonetheless, TBA model had higher accuracy than RF. It is considerable that this difference was slightly higher than BTL model within each scenario. Jónás, Ducrocq, and Croiseau (2017) reported that using LD information along the genome to build haplotypes specifically for genomic prediction is a favorable step to improve the accuracy of genomic prediction. Wientjes, Veerkamp, and Calus (2013) results indicated that LD has a small effect on the reliability of genomic prediction.

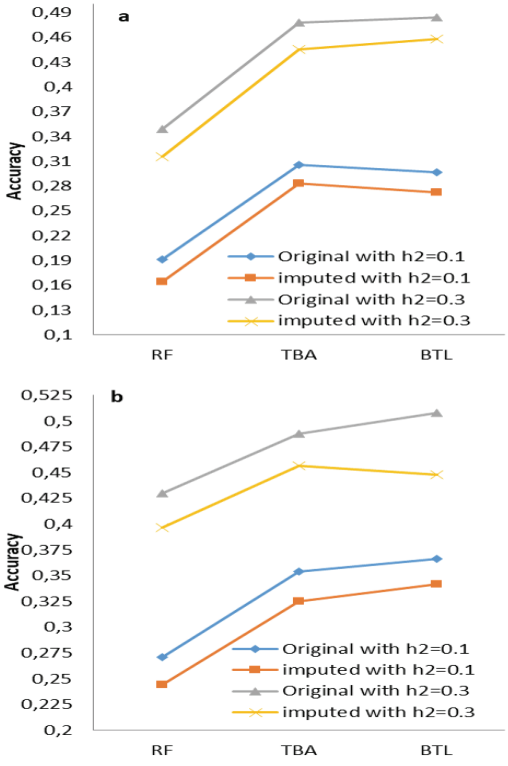

Figure 5. Effect of different level of heritability on accuracies of GEBVs estimated by threshold BayesA (TBA), Bayesian threshold LASSO (BTL) and random forest (RF) in original and imputed genotypes for 10 K (a) and 50 K (b) SNP panels.

In the case of the medium-density panel and with considering the similar levels of h^2=0.1 and QTL = 81 in imputed and original genotypes, the genomic accuracy was investigated for different levels of LD (e.g., VII scenario (LD = low) and VIII scenario (LD = high)) from RF, TBA and BTL (Tale 4 and Figure 6). At distances of 0.05 cM, the average observed LD (r^2) for VII and VIII scenarios were 0.241 and 0.438, respectively. Compared to the low LD scenario, the accuracy of regression

threshold models increased obviously for RF in the high LD scenario; nonetheless the increase level of LD was more effective on RF. The detection of disease-causing variants by association with neighboring SNPs depends on the existence of strong LD between them in the human genome (Ke et al., 2004). Theoretically, the extent of LD in a population is related to the effective population size (Ne) (Wang, Yu et al., 2017; Bohlouli, Alijani, Javaremi, König, & Yin, 2017). It is generally accepted that LD between markers and QTL is a main source of information, which is contributed to the accuracy of genomic prediction (Sun, Fernando, & Dekkers, 2016). Accuracies of estimated genomic breeding value showed an increase alongside with the enlargement of LD size, especially for RF, which is in agreement with simulated study by Naderi et al. (2016). Accuracy of the BayesA was improved with increase in LD of historical population in the half-sib families (Sun et al., 2016). A higher level of LD between QTL and marker showed that more markers are capturing higher proportion of the genetic variance (Goddard, 2009), and are prerequisite for an efficient performance of RF (Naderi et al., 2016).

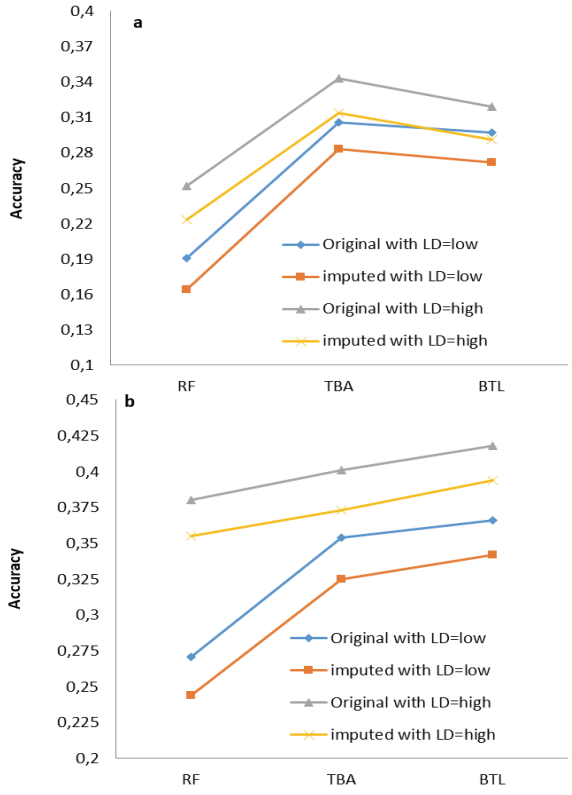

Figure 6. Effect of different level of LD on accuracies of GEBVs estimated by threshold BayesA (TBA), Bayesian threshold LASSO (BTL) and random forest (RF) in original and imputed genotypes for 10 K (a) and 50 K (b) SNP panels.

Conclusion

Imputation can be used to prediction of missing genotypes for the 10K and 50K SNP panels with imputation accuracy higher than 0.929 (on average 0.948) in simulated scenarios with 90% missing rate. In addition to quantifying imputation accuracy, results of current study shed light on the effects of level of LD and marker density on imputation accuracy. More importantly, application of these imputed genotypes will have little effect on the accuracy of estimated GEBVs. Anyway, a medium-density marker panel could be imputed from an available lower density marker panel, which will also have a lower cost.

The structure of genomic architecture and accuracy of imputation were the most important factors to analyze discrete traits affecting prediction accuracy in RF, TBA and BTL. The effect of structures including number of QTL, level of LD, marker density and heritability were more pronounced on the accuracy of GEBVs for RF than TBA and BTL. Generally, prediction accuracies were higher when using the Bayesian regressions (especially BTL). Only in the scenario combining the highest heritability, the dense marker panel, and the largest number of QTL, RF (despite the high computational time) was more precise.

References

Abdollahi-Arpanahi, R., Peñagaricano, F., Aliloo, H., Ghiasi, H., & Urioste, J. I. (2013). Comparison of Poisson, probit and linear models for genetic analysis of number of inseminations to conception and success at first insemination in Iranian Holstein cows. *Livestock Science, 153*(1), 20-26.

Andonov, S., Lourenco, D. A. L., Fragomeni, B. O., Masuda, Y., Pocrnic, I., Tsuruta, S., & Misztal, I. (2017). Accuracy of breeding values in small genotyped populations using different sources of external information — A simulation study. *Journal of Dairy Science, 100*(1), 395-401.

Atefi, A., Shadparvar, A. A., & Hossein-Zadeh, N. G. (2016). Comparison of whole genome prediction accuracy across generations using parametric and semi parametric methods. *Acta Scientiarum. Animal Sciences, 38*(4), 447-453.

Badke, Y. M., Bates, R. O., Ernst, C. W., Fix, J., & Steibel, J. P. (2014). Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. *G3: Genes, Genomes, Genetics, 4*(4), 623-631.

Blazer, D. G., & Hernandez, L. M. (2006). *Genes, behavior, and the social environment: Moving beyond the nature/nurture debate*. Washington, DC: National Academies Press.

Bo, Z. H. U., Zhang, J.-j., Hong, N. I. U., Long, G. U. A. N., Peng, G. U. O., XU, L.-y., ... Xue, G. A. O.

(2017). Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. *Journal of Integrative Agriculture, 16*(4), 911-920.

Bohlouli, M., Alijani, S., Javaremi, A.N., König, S., & Yin, T. (2017). Genomic prediction by considering genotype× environment interaction using different genomic architectures. *Annals of Animal Science,* 17, 683-701.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., ... Sonstegard, T. S. (2012). Design of a bovine low-density SNP array optimized for imputation. *PloS One,* 7(3), e34130.

Boison, S. A., Neves, H. H. R., O'Brien, A. M. P., Utsunomiya, Y. T., Carvalheiro, R., Silva, M. V. G. B., ... Garcia, J. F. (2014). Imputation of non-genotyped individuals using genotyped progeny in Nellore, a *Bos indicus* cattle breed. *Livestock Science, 166*, 176-189.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Carvalheiro, R., Boison, S. A., Neves, H. H. R., Sargolzaei, M., Schenkel, F. S., Utsunomiya, Y. T., ... Van Tassell, C. P. (2014). Accuracy of genotype imputation in Nelore cattle. *Genetics Selection Evolution, 46*(1), 69.

Chen, L., Li, C., Sargolzaei, M., & Schenkel, F. (2014). Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. *PLoS One, 9*(7), e101544.

Cleveland, M. A., & Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. *Journal of Animal Science, 91*(8), 3583-3592.

Coster, A., Bastiaansen, J. W. M., Calus, M. P. L., Van Arendonk, J. A. M., & Bovenhuis, H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution, 42*(1), 9.

Daetwyler, H. D., Hickey, J. M., Henshall, J. M., Dominik, S., Gredler, B., Van Der Werf, J. H. J., & Hayes, B. J. (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Animal Production Science, 50*(12), 1004-1010.

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One, 3*(10), e3395.

Dassonneville, R., Fritz, S., Ducrocq, V., & Boichard, D. (2012). Imputation performances of 3 low-density marker panels in beef and dairy cattle. *Journal of Dairy Science, 95*(7), 4136-4140.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., ... Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics, 182*(1), 375-385.

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science, 19*(9), 592-601.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

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. *Animal Production Science, 57*(2), 229-236.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica, 136*(2), 245-257.

González-Recio, O., & Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution, 43*(1), 1-7.

González-Recio, O., de Maturana, E. L., Vega, A. T., Engelman, C. D., & Broman, K. W. (2009). Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. *BMC Proceedings*, 3(Suppl 7), S63.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., ... Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics, 127*(3), 749-762.

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

Hickey, J. M., Crossa, J., Babu, R., & de los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science, 52*(2), 654-663.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics, 38*(6), 226-231.

Hozé, C., Fouilloux, M.-N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., ... Croiseau, P. (2013). High-density marker imputation accuracy in sixteen French cattle breeds. *Genetics Selection Evolution, 45*(1), 33.

Jónás, D., Ducrocq, V., & Croiseau, P. (2017). The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. *Journal of Dairy Science, 100*(4), 2905-2908.

Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., ... Bentley, D. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics, 13*(6), 577-588.

Khatkar, M. S., Moser, G., Hayes, B. J., & Raadsma, H. W. (2012). Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics, 13*(1), 538.

König, S., Brügemann, K., & Pimentel, E. C. G. (2013). Züchterische strategien für tier-und klimaschutz: Was ist möglich und was brauchen wir? *Zuchtungskunde, 85*, 22-33.

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics, 10*, 387-406.

Meuwissen, T. H. E. (2009). Accuracy of breeding values of'unrelated'individuals predicted by dense SNP genotyping. *Genetics Selection Evolution, 41*(1), 41-35.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157, 1819-29.

Mulder, H. A., Calus, M. P. L., Druet, T., & Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *Journal of Dairy Science, 95*(2), 876-889.

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. *Journal of Dairy Science, 99*(9), 7261-7273.

Neves, H. H., Carvalheiro, R., & Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC genetics*, 13: 100.

Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T., & Li, M. J. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, *16*(Suppl 2), S5.

Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., Takasuga, A., Sugimoto, Y., & Iwaisaki, H. (2016). Accuracy of imputation of single nucleotide polymorphism marker genotypes from low□density panels in Japanese Black cattle. *Animal Science Journal, 87*(1), 3-12.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association, 103*(482), 681-686.

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K.-U., & Fries, R. (2013). Imputation of high-density genotypes in the Fleckvieh cattle population. *Genetics Selection Evolution, 45*(1), 3.

Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D., & Goddard, M. E. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution, 49*(1), 24.

Pimentel, E. C. G., Edel, C., Emmerling, R., & Götz, K. U. (2015). How imputation errors bias genomic predictions. *Journal of Dairy Science, 98*(6), 4131-4138.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

Sargolzaei, M., Chesnais, J., & Schenkel, F. (2011). FImpute-An efficient imputation algorithm for dairy cattle populations. *Journal of Dairy Science, 94*(1), 421.

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC genomics*, *15*(1), 478.

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

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., ... McCouch, S. R. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics, 11*(2), e1004982.

Sun, X., Fernando, R., & Dekkers, J. (2016). Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genetics Selection Evolution, 48*(1), 77.

Sun, Y. V., Bielak, L. F., Peyser, P. A., Turner, S. T., Sheedy, P. F., Boerwinkle, E., & Kardia, S. L. R. (2008). Application of machine learning algorithms to predict coronary artery calcification with a sibship based design. *Genetic Epidemiology, 32*(4), 350-360.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528-540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 58*(1), 267-288.

Toghiani, S., Aggrey, S. E., & Rekaya, R. (2016). Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. *Animal, 10*(7), 1077-1085.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., ... Van Kaam, J. (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science, 96*(1), 668-678.

VanRaden, P. M., & Sullivan, P. G. (2010). International genomic evaluation methods for dairy cattle. *Genetics Selection Evolution, 42*(1), 1-9.

Ventura, R. V., Miller, S. P., Dodds, K. G., Auvray, B., Lee, M., Bixley, M., ... McEwan, J. C. (2016). Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution, 48*(1), 1-20.

Wang, C., Li, X., Qian, R., Su, G., Zhang, Q., & Ding, X. (2017a). Bayesian methods for jointly estimating genomic breeding values of one continuous and one threshold trait. *PloS One, 12*(4), e0175448.

Wang, C. L., Ding, X. D., Wang, J. Y., Liu, J. F., Fu, W. X., Zhang, Z., ... Zhang, Q. (2013). Bayesian methods for

estimating GEBVs of threshold traits. *Heredity, 110*(3), 213-219.

Wang, Q., Yu, Y., Yuan, J., Zhang, X., Huang, H., Li, F., & Xiang, J. (2017b). Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. *BMC Genetics, 18*(1), 1-17.

Wang, Y., Lin, G., Li, C., & Stothard, P. (2016). Genotype imputation methods and their effects on genomic predictions in cattle. *Springer Science Reviews, 4*(2), 79-98.

Wientjes, Y. C. J., Veerkamp, R. F., Bijma, P., Bovenhuis, H., Schrooten, C., & Calus, M. P. L. (2015). Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution, 47*(1), 1-14.

Wientjes, Y. C. J., Veerkamp, R. F., & Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics, 193*(2), 621-631.

Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics, 19*(6), 506-536.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., ... Van Vliet-Ostaptchouk, J. V. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics, 47*(10), 1114-1123.

Yin, T., Pimentel, E. C. G., Borstel, U. K., & König, S. (2014). Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature× humidity-dependent covariate. *Journal of Dairy Science, 97*(4), 2444-2454.

Received on August 08, 2017.
Accepted on December 07, 2017.