Animal Science Papers and Reports vol. 37 (2019) no. 2, 149-168
Institute of Genetics and Animal Breeding, Jastrzębiec, Poland

Assessment of genomic prediction accuracy of discrete traits with imputation of missing genotypes

Yousef Naderi[1]* and Saadat Sadeghi[2]

[1] Islamic Azad University, Astara Branch, Department of Animal Science, Astara, Iran. Young Researchers Club, Islamic Azad University, Astara Branch, Astara, Iran

[2] Tabriz Branch, Islamic Azad University, Tabriz, Iran

(Accepted March 15, 2019)

Genomic selection as a promising tool for discovering genetic variants influencing complex traits, along with genotype imputation has an important role in increasing economic efficiency as well as genetic gain by accelerating animal breeding programs and potentially improving the accuracy of breeding values. The objectives of present research were: (i) to quantify the accuracy of genotype imputation and to evaluate factors affecting it, and (ii) to assess the effects of genotype imputation and genomic architecture on the performance of the Random Forest (RF), GBLUP and threshold Bayes A (TBA) methods for genomic predictions of binary traits. According to disease incidence and genomic architecture (heritability (h^2) = 0.25 or 0.05, QTL=81 or 810 and linkage disequilibrium (LD) = low or high), reference and validation sets were organised in different simulated scenarios for the 54K SNP panel. To evaluate imputation accuracy, we randomly masked (90 and 50 percent of markers) and subsequently imputed certain genotypes using the FImpute programme. The disease incidence slightly affected prediction accuracies. A negative effect of increased missing genotypes on accuracies of genomic prediction was observed when applying TBA and GBLUP rather than RF. The TBA method performed better than the RF and GBLUP methods for genomic prediction. Nonetheless, for a scenario affected by a high number of QTLs and a high level of heritability, RF was more precise with an extension of computational time. The results suggested that genotype imputation from sparse panels (5.4 K SNPs) with high LD to 50K panels could be a cost-effective approach for genomic selection.

KEY WORDS: complex trait / genomic selection / heritability / missing genotypes

The development of genotyping technologies has facilitated genetic progress in breeding programmes by implementing genomic selection (GS). In fact, the

*Corresponding author: y.naderi@iau-astara.ac.ir, yousefnaderi@gmail.com

accuracy of genomic evaluations has been enhanced via GS and thus has quickly spread in livestock breeding. These methodologies are of great value especially for the improvement of complex traits such as disease resistance in livestock [Yáñez *et al.* 2014, Garrick 2017]. However, the economic aspect of genotyping may be a limiting factor for the practical implementation of GS [VanRaden *et al.* 2011].

Genotype imputation method is an alternative in genomic applications to detect missing SNPs during the genotyping process through reference population information. The merit of imputing genotypes essentially depends on its imputation accuracy [Yoshida *et al.* 2018], which is affected by several factors, such as the proportion of missing genotypes [Zhang and Druet 2010, Hickey *et al.* 2012], the size of the training set [Druet *et al.* 2010], relatedness between the training and testing sets [Cleveland and Hickey 2013, Carvalheiro *et al.* 2014], the level of LD [Hickey *et al.* 2012], chromosomal position [Hozé *et al.* 2013] and minor allele frequency [Badke *et al.* 2014]. There are many investigations reporting that software and methods of imputation give reasonable levels of imputation accuracy [Calus *et al.* 2011, Mulder *et al.* 2012, Sun *et al.* 2012]. Hence, genotype imputation may supply an appropriate alternative to reduce genotyping costs; as a results it has been offered for commercial usage [Weigel *et al.* 2010]. Previous studies [Hayes *et al.* 2012, Mulder *et al.* 2012, Wang *et al.* 2013b] showed that genomic prediction accuracy of imputed genotypes is comparable to that of original genotypes when imputation accuracy was above 0.95. In addition to the factors affecting the accuracy of imputation, genomic prediction accuracy depends on factors related to the genetic architecture of population, such as marker density, trait heritability [Guo *et al.* 2014], LD [Yin *et al.* 2014], the number of QTLs and the discrete phenotype ratio allocated to the training set [Naderi *et al.* 2016], type of trait (categorical vs. continuous traits), statistical method [Wang *et al.* 2017a] and imputation accuracy [Toghiani *et al.* 2016].

As a matter of fact, GS focusing on continuous traits has been shown to improve the accuracy of GEBVs. Since many prominent traits in livestock, such as resistance to disease and degree of calving difficulty, present a binary distribution of phenotypes (and are often termed threshold traits), it seems important to consider these traits in animal breeding [Wang *et al.* 2013a]. Hence, GS methods must be adapted to cope with challenges of discrete traits. Therefore, threshold versions of Bayesian regressions, genomic best linear unbiased prediction (GBLUP) and machine learning methods are applied for genomic analyses of these traits [González-Recio and Forni 2011, Naderi *et al.* 2016]. Felipe *et al.* [2014] evaluated the impact of genotype imputation on the predictive ability of complex traits by several models of semi and non-parametric models. Their results indicated factors that may affect the applicability of genomic prediction accuracy, i.e. the breeding programme design, genetic architecture of the trait, the structure of the population, statistical model and accuracy of imputation. Chen *et al.* [2014] evaluated the effect of imputed genotypes on the predictive ability of Bayesian regression and GBLUP methods and indicated that imputation errors influenced performance of both Bayesian and GBLUP methods.

In genome studies, simulation allows researchers to discover the effects of the genetic architecture of the target trait for evaluating some sources of variability, which cannot be distinguished by most real data [Daetwyler *et al.* 2010]. Therefore, the main objective of this study was to explore whether non-parametric methods, such as threshold Bayes A, threshold GBLUP and Random Forest for different disease incidence rates and population structure, could track genetic signals to provide acceptable accuracy of genomic predictions from low-density panels with the need of imputing to higher density panels. In this way, the computational aspects are also vital factors which have to be considered, because in turn they may affect the general performance of each method.

Material and methods

Population structure

A population of five thousand animals genotyped for 54,000 markers was simulated using the QMSim software [Sargolzaei and Schenkel 2009]. In the first phase, over a time span of 1,000 generations, a historical population was derived from 4800 females and 200 males. In the second phase, in order to produce a realistic level of LD, a bottleneck was used. For this purpose the population size decreased over 100 generations to 400 individuals. In the third phase, the population size increased over 100 generations and returned to the first phase (4,800 females and 200 males). All 5000 individuals of the last historical generation served as founders and using a random mating design expanded the recent population by simulating an additional 10 generations. During these generations, the replacement ratio was set at 0.2 and 0.50 for females and males, respectively, while selection of candidate individuals were based on EBV and age. Each mating produced only one offspring with the same probability of being either male or female. Individuals of generations 6 to 9 were used as a training set, while the whole generation 10 was considered as the validation set (5000 individuals). With regard to heritability (0.05 and 0.025), the number of QTLs (either 81 or 810 QTLs) and LD (low and high), 4 different scenarios including I (54K SNPs, $h^2 = 0.25$, LD = low and 810 QTLs), II (54K SNPs, $h^2 = 0.25$, LD = low and 81 QTLs), III (54K SNPs, $h^2 = 0.05$, LD = low and 81 QTLs) and IV (54K SNPs, $h^2 = 0.05$, LD = high and 81 QTLs) were simulated to reflect variations. All QTLs were randomly located along 27 chromosomes 100 cM long from a gamma distribution with a shape parameter of 0.4. The mutation rate was fixed at 2.5×10^{-5} for both SNPs and QTLs per locus and per generation, as used in previous simulations [Naderi *et al.* 2016]. For all scenarios, 10 replicates were simulated to evaluate the models. Table 1 shows more explanation of parameters as used for the simulations.

To create a discrete trait, the phenotype of training individuals was coded as 1= sick or 0= healthy, and the percentage of sick animals within the training set was considered 20% (group 1) and 50% (group 2), respectively, whereas phenotypes in the testing set were assumed to be unknown. Markers were excluded if they showed extreme departure

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 (5,000)	1,000 (5,000)
no. of generations (population size) in phase 2	1,100 (5,000)	1,100 (400)
no. of generation (population size) in phase 3	1,200 (5,000)	1,200 (5,000)
Recent population		
no. of founder sires (dams)	200 (4,800)	
no. of generations	10	
no. of offspring per dam	1	
mating system	random	
replacement ratio for males (females)	0.5 (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	54000 or 5400	
heritability of the trait	0.05 or 0.25	

from the Hardy–Weinberg equilibrium ($P<10^{-5}$) or minor allele frequency (MAF) was under 0.03. To calculate the squared correlation coefficient (r^2) between marker pairs in the last generation, the PLINK programme [Purcell *et al.* 2007] was used.

Imputation

To evaluate imputation accuracy, 90 and 50 percent of markers randomly masked in 54K SNP platform; afterwards, masked markers were imputed by considering a family and population-based algorithm with the FImpute programme [Sargolzaei *et al.* 2011a]. 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 minimise dependence on allele frequency.

Genomic prediction

GEBVs and accuracy of genomic prediction were estimated using the threshold Bayes A, Random Forest and GBLUP methods for simulated discrete data. For the GBLUP method we used the AI-REML algorithm and implemented the DMU software package [Madsen and Jensen 2010], which provides specification of a generalised linear mixed model with a logit link function for discrete data. The random individual effect was involved by considering genomic relationships between individuals based on marker data. According to the method proposed by VanRaden [2008], the genomic

relationship matrix (G matrix) was constructed and the GEBV software [Sargolzaei *et al.* 2011b] was applied.

For the RF method, we used the java RanFoG package [González-Recio and Forni 2011], which may alleviate the problems of analysing genome-wide data using feature selection and bootstrapping [Efron and Tibshirani 1993]. The RF prediction for an observation, $\hat{f}_{rf}^{p}(x)$, is computed by averaging predictions over p trees, $[T(x.\Psi p)]$, for which a given observation was not used to build the tree and 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:

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

An optimum combination was found when RF parameters such as *ntree* (the number of trees to grow), *mtry* (the number of SNPs randomly selected at each tree node) and *nodesize* (the minimum size of terminal nodes of trees) were pre-determined and tuned. Animals not included in the bootstrapped sample were defined as "out of bag", being the validation set for each tree. Therefore, the out of bag error is a basic factor in RF and determines the best output RF by an optimum combination of RF parameters. In this study, RF was 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 (*mtry*). At each node, data were split in 2 branches based on the genotype at SNPj by minimising a loss function for classification (*nodesize*). In the current study, 5,000 trees (*ntree*) were constructed for 54K SNP chips in original and imputed data. 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 marker genotypes. The average of the predicted value for each tree was the probability of being susceptible to the disease.

For threshold Bayes A (TBA), the BGLR package [De Los Campos *et al.* 2009] of R software was used. TBA assumes that each marker has a different variance, but each has an effect. The TBA can be described as follows:

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

where: the underlying liability variable vector for **y** is λ, μ is the population mean, column vector (n×1) of ones is 1; b indicate [bj] the vector for the regression coefficient estimates of the p markers assumed to be normally and independently distributed a priori as $N(0, \sigma_j^2)$, which σ_j^2 is 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 $s^2_j = 0.002$ is assumed for the prior distribution of σ_j^2. Elements of the incidence matrix **X**, of order $n \times p$, was set for the additive model. The residuals (**e**) are assumed to be distributed as $N(\mu=0, \sigma_e^2=1)$, as stated above [González-Recio and Forni 2011].

Generally, the accuracy of genomic prediction was calculated by the *phi*-correlation coefficient between TBV and EBV for each model and simulated a scenario in the validation set. The following model was used to the phi-correlation coefficient [González-Recio and Forni 2011]:

$$r_{\phi} = \frac{\rho(\hat{\mathbf{y}} = 1 \,|\mathbf{y} = 1) - \rho(\hat{\mathbf{y}} = 1)\rho(\mathbf{y} = 1)}{\sqrt{\rho(\hat{\mathbf{y}} = 1)\rho(\mathbf{y} = 1)\rho(\hat{\mathbf{y}} = 0)\rho(\mathbf{y} = 0)}}$$

All operation steps of the current study are shown in Figure 1.

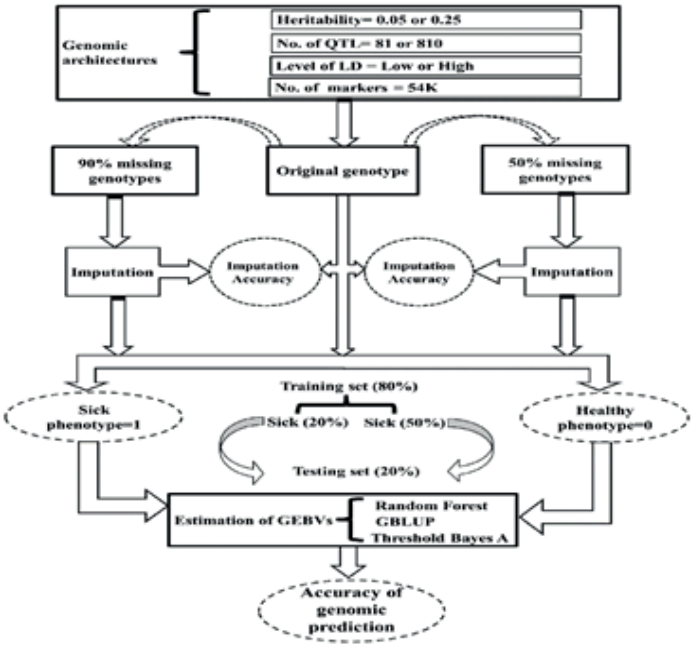

Fig.1. Overview of individual steps of research process

Computational time

We ran all the models using a computer with 32 gigabytes RAM and Core i7 – CPU 3.6 GHz. For each of the scenarios, computational time and the amount of required memory were recorded to compare the methods.

Results and discussion

Results indicated high imputation accuracies under different scenarios from different low-density panels (5.4 K and 27 K) to the 54 k SNP panel (Tab. 2). The average of imputation accuracy for the scenario with 50% of the genotypes masked (27K) was higher than for the scenario with 90% of the genotypes masked (5.4 K SNPs). We observed that the accuracy of imputation improved with increasing LD. The beneficial effect of LD on increasing imputation accuracy was greater when the sparse panels (5.4 K SNPs) were used.

Table 2. Accuracy of imputation and standard deviation for 90 and 50 percentage of masked genotypes for all scenarios 8,5

Scenarios	90% of masked genotypes		50% of masked genotypes	
	imputation accuracy	SD	imputation accuracy	SD
I	0.958	0.018	0.972	0.013
II	0.959	0.019	0.974	0.014
III	0.958	0.018	0.974	0.012
IV	0.97	0.013	0.986	0.01

I (54K SNP, $h^2 = 0.25$, LD = low and 810 QTL), II (54K SNP, $h^2 = 0.25$, LD = low and 81 QTL), III (54K SNP, $h^2 = 0.05$, LD = low and 81 QTL) and IV (54K SNP, $h^2 = 0.05$, LD = high and 81 QTL).

In animal breeding programmes the rate of genetic gain has been improved by genomic selection. In this regard, imputation using low density panels to high density panels and also genotyping of more individuals facilitated a reduction of the genotyping costs in the breeding practice. In recent years, some investigations have reported the predictive ability of models using subsets of SNPs, with and without imputation [Daetwyler *et al.* 2011, Mulder *et al.* 2012, Felipe *et al.* 2014, Chen *et al.* 2014]. In most cases, predictive ability improved with imputed genotypes, thus many researchers recommend the imputation strategy to reduce costs in genomic selection programmes. However, some studies reported that this strategy was not always useful for genomic selection programmes [Felipe *et al.* 2014], which in turn depends on different factors and will be considered in our study.

The results obtained showed that LD between SNPs in the original and low density panels helped to increase the accuracy of imputation. Also, similarity of LD patterns between the training panel and the set to be imputed serves as a basis for imputing the missing genotypes. In this study, the factors affecting imputation accuracy (the number of markers available than the original panel and LD among markers) are comparable to the reports published on maize [Hickey *et al.* 2012], Australian Holstein-Friesian cattle [Khatkar *et al.* 2012], Dutch Holstein cattle [Mulder *et al.* 2012], Fleckvieh and Holstein cattle [Pausch *et al.* 2013, Pausch *et al.* 2017], Yorkshire boars [Badke *et al.* 2014], a simulated population of Brazilian Nellore cattle [Boison *et al.* 2014] and Japanese Black cattle [Ogawa *et al.* 2016]. Generally, because of low linkage and LD among markers in low density panels and increasing imputation errors, the accuracy of imputation was reduced with a decrease of marker density (or an increase in the ratio of missing genotypes).

Accuracy of genomic predictions

Impact of disease incidence in training sets on accuracy of genomic prediction.

Accuracies of estimated GEBVs for different scenarios and marker densities are shown in Tables 3 and 4, with 20 and 50 percent of animals in the training set being sick, respectively. The distribution of sick individuals into the training sets

slightly affected accuracy of genomic predictions in all models considered. When the number of sick individuals decreased from 50% to 20%, prediction accuracy increased correspondingly. For both groups, the average of prediction accuracies from TBA always outperformed those from corresponding GBLUP and RF applications. Nonetheless, standard errors for RF across the 10 replicates were more homogeneous in comparison with TBA and GBLUP applications. The highest accuracy (0.608) was observed in scenario II when 20 percent of animals in the training set were sick. With regard to scenario I, RF performed better than TBA and GBLUP for both groups. For III and IV and in comparison to RF, TBA and GBLUP showed higher accuracies of genomic prediction and differentiated more accurately between both groups for original and imputed genotypes.

González-Recio and Forni [2011] simulated a genomic population considering a discrete trait, by including 2500 animals to establish a training set to investigate accuracy of genomic prediction via machine-learning and Bayesian regressions methods. Their results showed accuracy of 0.36 and 0.26 using RF and TBA, respectively. Ogutu *et al.* [2011] applied RR-BLUP and RF methodologies to a calibration set of 2,326 individuals genotyped by 10K SNPs. They reported a better predictive ability in RR-BLUP (0.6) rather than RF (0.48). Naderi *et al.* [2016] simulated different genomic architectures and selected different proportions of animals to establish a training set

Table 3. The accuracies of estimated GEBVs using the original and imputed SNP genotypes from genomic BLUP (GBLUP) random forest (RF), threshold Bayes A (TBA) applications when 20 percentage of animals in the training set are sick (the values in parentheses show the SD from 10 replicates)

Model	Imputation ratio	Scenarios				Average
		I	II	III	IV	
GBLUP	90%	0.508 (0.05)	0.475 (0.06)	0.34 (0.06)	0.448 (0.05)	0.443 (0.05)
	50%	0.52 (0.05)	0.498 (0.05)	0.366 (0.05)	0.454 (0.05)	0.460 (0.04)
	Original	0.544 (0.04)	0.524 (0.05)	0.397 (0.04)	0.466 (0.04)	0.483 (0.03)
RF	90%	0.567 (0.03)	0.45 (0.04)	0.319 (0.03)	0.429 (0.04)	0.441 (0.03)
	50%	0.578 (0.03)	0.469 (0.04)	0.341 (0.03)	0.436 (0.03)	0.456 (0.03)
	Original	0.592 (0.02)	0.493 (0.03)	0.369 (0.02)	0.446 (0.02)	0.475 (0.02)
TBA	90%	0.517 (0.06)	0.565 (0.05)	0.391 (0.06)	0.465 (0.05)	0.485 (0.05)
	50%	0.544 (0.05)	0.579 (0.04)	0.406 (0.05)	0.474 (0.04)	0.501 (0.04)
	Original	0.580 (0.04)	0.608 (0.04)	0.453 (0.04)	0.491 (0.03)	0.533 (0.03)

I (54K SNP, $h^2 = 0.25$, LD = low and 810 QTL), II (54K SNP, $h^2 = 0.25$, LD = low and 81 QTL), III (54K SNP, $h^2 = 0.05$, LD = low and 81 QTL) and IV (54K SNP, $h^2 = 0.05$, LD = high and 81 QTL).

Table 4. The accuracies of estimated GEBVs using the original and imputed SNP genotypes from genomic BLUP (GBLUP) random forest (RF), threshold Bayes A (TBA) applications when 50 percentage of animals in the training set are sick (the values in parentheses show the SD from 10 replicates)

Model	Imputation ratio	Scenarios				Average
		I	II	III	IV	
GBLUP	90%	0.506 (0.06)	0.472 (0.05)	0.321 (0.06)	0.423 (0.05)	0.431 (0.05)
	50%	0.517 (0.05)	0.486 (0.05)	0.345 (0.05)	0.436 (0.05)	0.446 (0.05)
	Original	0.541 (0.04)	0.516 (0.05)	0.376 (0.04)	0.447 (0.05)	0.47 (0.04)
RF	90%	0.531 (0.03)	0.41 (0.04)	0.275 (0.03)	0.396 (0.03)	0.403 (0.03)
	50%	0.54 (0.03)	0.424 (0.03)	0.3 (0.03)	0.404 (0.03)	0.417 (0.03)
	Original	0.556 (0.02)	0.453 (0.03)	0.329 (0.02)	0.416 (0.02)	0.4385 (0.02)
TBA	90%	0.491 (0.06)	0.556 (0.06)	0.372 (0.05)	0.455 (0.05)	0.469 (0.05)
	50%	0.51 (0.05)	0.572 (0.05)	0.389 (0.04)	0.462 (0.05)	0.483 (0.04)
	Original	0.549 (0.05)	0.608 (0.04)	0.429 (0.04)	0.485 (0.04)	0.518 (0.04)

I (54K SNP, $h^2 = 0.25$, LD = low and 810 QTL), II (54K SNP, $h^2 = 0.25$, LD = low and 81 QTL), III (54K SNP, $h^2 = 0.05$, LD = low and 81 QTL) and IV (54K SNP, $h^2 = 0.05$, LD = high and 81 QTL).

to evaluate efficiency of GBLUP and RF algorithms for a binary trait. In contrast with our results, they reported that genomic prediction accuracy of the GBLUP model increased from 0.23 to 0.36 for a low LD scenario due to an increase in the percentage of sick individuals in the training set from 2.5 to 25 percent, while for RF it increased from 0.13 to 0.30 until the 20% allocation scheme, and then, in agreement with our results, it decreased to a negligiblr 0.27 from the 25% allocation scheme. When 20 percent of animals in the training set were sick, prediction accuracies reported by Naderi *et al.* [2016] were consistently lower for both the RF (0.30-0.53) and GBLUP (0.32-0.50) methods than the present research results, i.e. 0.369- 0.592 for RF, 0.397-0.544 for GBLUP and 0.453-0.608 for TBA.

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 the results recorded by Naderi *et al.* [2018] for disease traits in Holstein Friesian cows. In brief, they specified that a correlation between pre-corrected phenotypes and genomic breeding values (rGBV) increased by the decrease in the percentage of sick cows in the training set from 37 to 20 percent for claw disorders, from 32 to 25 percent for clinical mastitis and from 29 to 19 percent for female infertility. One possible explanation for different reactions of various traits to the decreased percentage of sick individuals in training sets addresses the

different distributions of response variables. Generally, for binary traits as response variables the optimal individuals in the training sets had disease incidences that were close to the rates in the whole population. In this study, the phenotype of the animals was coded as 0 or 1 depending on whether their simulated phenotype was above or below the average population (the 50% allocation scheme) or above or below 20 percent of population phenotypes (the 20% allocation scheme), respectively. Because of the normal distribution for the simulated phenotypes, the phenotype mode tends to fluctuate around the average axis, so more individuals are in the vicinity of the average population for the 50% allocation scheme than the 0.2 axis for the 20% allocation scheme. This distribution leads to a situation when more individuals are coded without considering their merit and only using their phenotypes, which in turn generates more classification errors for binary phenotypes of the 50% allocation scheme than the 20% allocation scheme. In conclusion, the prediction accuracy is unintentionally decreased.

Impact of missing genotypes on accuracy of genomic predictions.

The effect of missing genotypes on the genomic prediction accuracy depends on methods and the missing rate (Fig. 2). Furthermore, performances of RF, GBLUP and TBA methods were affected by missing genotypes. Comparing to original genotypes, accuracy of genomic prediction dropped rapidly when the 90% missing rate panel was used. The genomic prediction accuracies for imputed genotypes were comparable to the original genotypes within each method and scenario. Generally, the imputation error leads to a reduction in imputation accuracy and subsequently decreasing imputation accuracy results in a reduced accuracy of genomic prediction. Results from the present study showed the negative effect of increased missing genotypes

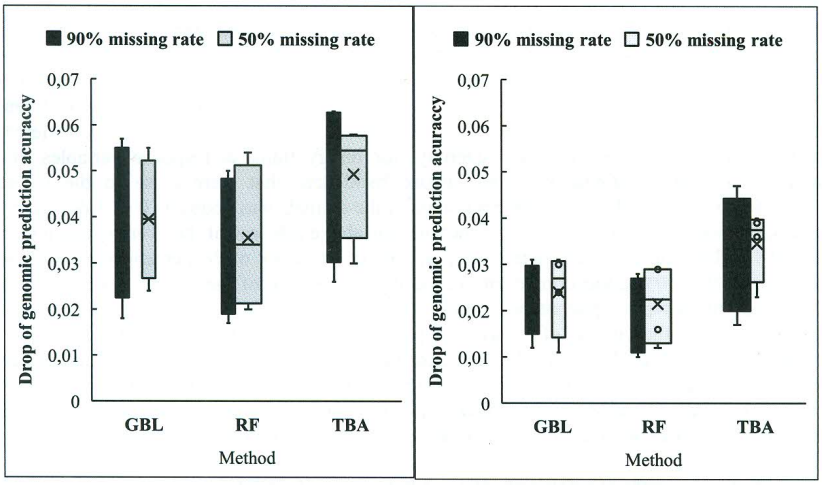

Fig. 2. Effect of missing genotype proportion on reduction of genomic prediction accuracy compared to the original panel for threshold Bayes A (TBA), GBLUP (GBL) and Random Forest (RF) methods when 20 (left) and 50 (right) percent of animals in the training set were sick.

on accuracies of genomic prediction when applying TBA and GBLUP rather than RF. Due to the negligibility of the imputation error, the noise added by imputation was lower for scenario IV (high imputation accuracy) than the other scenarios. As a result, with increasing imputation accuracy, differences between original and imputed genotypes (90%) in accuracies of genomic prediction reached the minimum.

One of the purposes of this study was to investigate whether imputation from a low-density to a higher density panels could be helpful for predictive ability of threshold methods. Overall, slight differences were seen among accuracies of genomic prediction via GBLUP, TBA and RF when original and imputed genotypes were compared. The results obtained from current research showed that the average accuracy of imputation was acceptable (0.961) when the sparse panels were used; therefore, it seems that the application of the 5.4 K SNP panel could be a good choice for genomic selection programs. This indicates that imputation yields extra information to the model and hence no increase in marker density is necessary to improve accuracy of genomic prediction. It is obvious that (i) imputation accuracy has a large influence on the accuracy of GEBVs [Wang *et al.* 2016], (ii) Mulder *et al.* [2012] after using a deterministic equation concluded that accuracy of GEBV increased linearly with an increase in imputation accuracy. Toghiani *et al.* [2016] and Weigel *et al.* [2010] investigated the accuracy of direct genomic values from the imputed SNP panels to original panels via Bayesian methods. They concluded that in scenarios where imputation accuracy was high, imputation could improve accuracy of genomic prediction. In the same context, researches [Mulder *et al.* 2012, Felipe *et al.* 2014] showed that where imputation error was high, as a result damage caused by imputation may be greater than its benefit in genomic selection programmes. Generally, studies showed that accuracy of genomic prediction is acceptable when accuracy of imputation is higher than 0.88 [Badke *et al.* 2014]. Felipe *et al.* [2014] investigated the accuracy of genomic prediction for true and imputed genotypes (with 90, 75 and 50% missing rates) in a mouse population using linear and semi and non-parametric models. They reported that genotype imputation had the same effect on the performance of the Bayesian LASSO model, while Bayesian Regularized Artificial Neural Network accuracies were more sensitive to the imputation error. The accuracy of genomic prediction obtained from imputed genotypes was significantly decreased compared to that from true data; nonetheless, it seemed that imputation is beneficial when relatedness between training and validation sets was poorer [Felipe *et al.* 2014].

Impact of genomic architecture on accuracy of genomic predictions

Impact of number of QTLs. For the original and imputed genotypes, the accuracy of genomic prediction from GBLUP, RF and TBA applications for scenarios I (810 QTLs) and II (81 QTLs), both considering a trait with a y 0.25heritabilit, are shown in Tables 3 and 4. For both groups, in contrast with TBA, with an increase in the number of QTLs accuracies of GEBV were partially higher for RF in original and imputed genotypes (Fig. 3). A reduction of genomic prediction accuracy due to a change of QTL number was generally greater when a higher number of genotypes was missed.

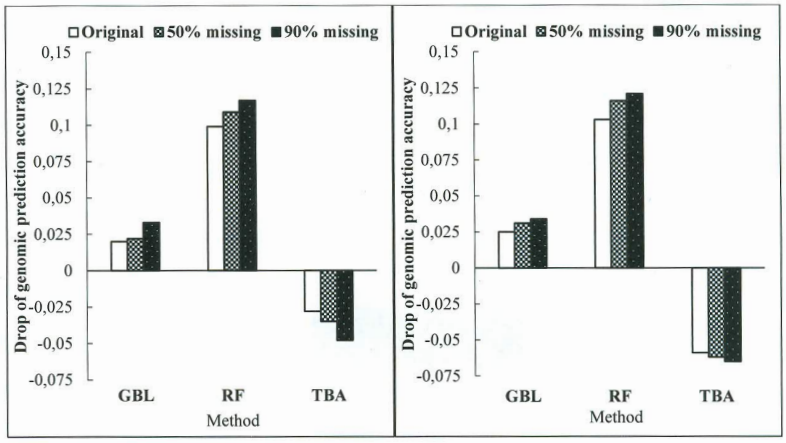

Fig. 3. The effect of decreasing the number of QTLs on a reduction of genomic prediction accuracy using threshold Bayes A (TBA), GBLUP (GBL) and Random Forest (RF) methods when 20 (left) and 50 (right) percent of animals in the training set were sick.

In the current study the highest accuracy was obtained for RF. An increase in the number of QTLs generally led to a major improvement in RF accuracies, while negligible positive and negative effects were found for GBLUP and TBA, respectively. In the present study, effects of QTL alleles were simulated with a gamma distribution. Gamma distribution of QTL effects would cause a minor proportion of them showing large effects, which may not be a desirable hypothesis for GBLUP compared to the BTA method. GBLUP was less sensitive to an increase in the number of QTLs (relatively stable results for both QTL scenarios in Fig. 4). For TBA, when the number of QTLs decreased, the total genetic variance was divided among fewer QTLs; therefore, the performance of methods increased to estimate such large QTL effects. In addition, RF was more sensitive to an increase in the number of QTLs compared with GBLUP, which in turn may be explained as follows: GBLUP assumed the same variance for each independent chromosome segment regardless of the effect of the segment, while RF was based on a sampling technique for predictors (SNPs). Hence, by using 50K panels combined with a large number of QTLs, SNPs in close distance to QTLs were sufficiently frequently sampled. Nevertheless, most of the important breeding traits are affected by many genes with small effects, and this supports the assumptions made for GBLUP applications [Hayes *et al.* 2009]. In scenario I, the highest accuracy of genomic prediction was recorded for TBA; it seems that several large QTLs are responsible for this phenomenon [Hayes *et al.* 2009]. At a constant heritability (h^2=0.3) and high-density SNP platforms, GBLUP was insensitive to genetic architecture (i.e. the number of QTLs), while the genomic prediction accuracy of the RF method improved as the number of QTLs increased [Naderi *et al.* 2016]. High accuracies of genomic prediction could be obtained when the number of QTLs decreased via Bayesian regressions methods, while accuracy of

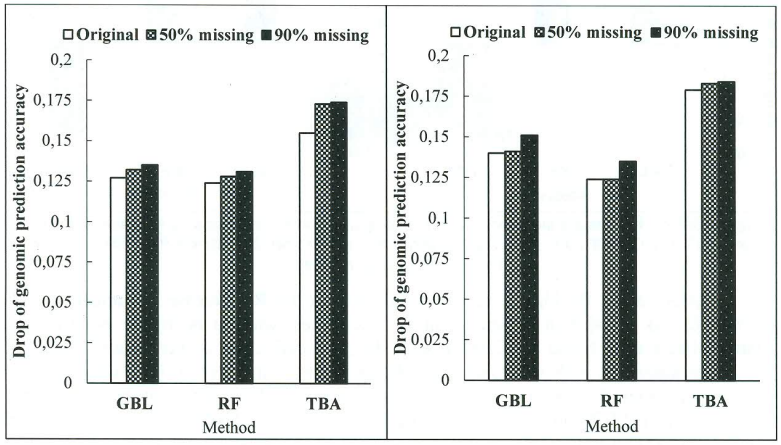

Fig. 4. The effect of decreasing heritability on the reduction of genomic prediction accuracy using threshold Bayes A (TBA), GBLUP (GBL) and Random Forest (RF) methods when 20 (left) and 50 (right) percent of animals in the training set were sick.

partial least square regression was unaffected [Coster *et al.* 2010]. Furthermore, in a later study the RF algorithm reached a higher accuracy (0.36) for smaller number of QTLs, while TBA methods showed a better predictive ability when a high number of QTLs was used [González-Recio and Forni 2011]. Generally, a different number of simulated chromosomes [Daetwyler *et al.* 2010], effective population sizes [Andonov *et al.* 2017], trait architecture [Ghafouri-Kesbi *et al.* 2017] and the additive nature of the simulated scenarios were in contrast to real data, with a more complex interaction between genes and biological pathways being a poential reason for inconsistency of earlier findings with our results.

Impact of heritability. The effect of heritability (h^2=0.25 and h^2=0.05) on accuracies of genomic prediction is depicted in Tables 3 and 4 (comparison of scenarios II and III). With an increase of heritability, the accuracy of genomic prediction considerably improved for all the methods applied for original and imputed genotypes in both groups. Because of decreasing heritability, the reduction of genomic prediction accuracy was rapid with an increase in the proportion of missing genotypes, which in turn was more pronounced for TBA rather than for GBLUP and RF (Fig. 4).

Resende *et al.* [2012] applied different Bayesian methods and RR–BLUP to a Pinus Taeda (loblolly pine) training population of 951 individuals genotyped with 5K SNPs. For all the methods the ability to predict phenotypes was linearly correlated with trait heritability. Our results are in accordance with the theory proposed by Bo *et al.* [2017] concerning the direct relationship between heritability and accuracy of genomic prediction. Furthermore, Neves *et al.* [2012] compared different methods (Kernel regression, LASSO, Random Forest, Ridge regression) in the evaluation of a mouse population with a wide range of heritability on the accuracy of genomic

prediction and found no significant differences between these methods. Wang *et al.* [2017a] investigated the effect of different heritabilities on the accuracies of genomic prediction using Bayesian methods in threshold traits. They reported that the accuracy of genomic prediction using threshold Bayes Cπ increased when heritability of the target trait increased. In many previous studies [Atefi *et al.* 2016], profitable effects of increasing heritability on the accuracy of genomic prediction via the Bayesian model were confirmed. These positive effects may be a result of higher genetic variations and contribute to accurate predictions of marker effects. Generally, high heritability means a strong role of the genes with additive expression to create variation in a trait, which, in turn, leads to a more accurate estimation of SNP effects. Regardless of the high heritability of the target trait, in that case the phenotype of the individuals is close to their genotype values; as a result, the effects of SNPs and thus alos genomic breeding values of individuals will be more precisely predicted [Goddard and Hayes 2009, Villumsen *et al.* 2009].

Impact of LD structure. We presented the pattern of different LD structures [i.e. scenario III (with average LD=0.233 at distances of 0.05 cM) vs. IV (with average LD=0.431 at distances of 0.05 cM)] on the accuracy of genomic prediction according to RF, TBA and GBLUP in imputed and original genotypes (Tab. 4 and Fig. 5). Generally, there was a decrease in prediction accuracies for all the methods with decreasing LD levels. Nonetheless, the accuracy of genomic prediction was rapidly reduced when the RF method was used.

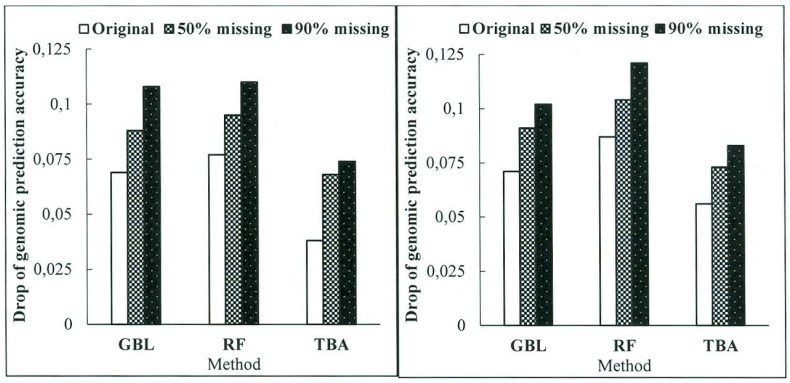

Fig. 5. The effect of a decrease in LD level on the reduction of genomic prediction accuracy using threshold Bayes A (TBA), GBLUP (GBL) and Random Forest (RF) methods when 20 (left) and 50 (right) percent of animals in the training set were sick.

The increase of LD can positively affect the accuracy of genomic prediction in two ways, including the effect on imputation accuracy and model estimation. In the case of only a weak genetic relationship between the reference and validation sets, LD is an important factor. Not only does a high LD mean a lower marker density requirement to cover the genome, but also a higher collinearity among linked markers [Liu *et al.*

2015]. In this study, LD had a different effect on the performance of RF, GBLUP and TBA accuracies, with imputed genotypes (especially 90% missing genotypes) being more sensitive than original data. Generally, RF was more sensitive than GBLUP and TBA to LD variations. Studies concerning the human genome showed that the existence of strong LD between neighbouring SNPs has a basic effect on the detection of disease-causing variants [Ke *et al.* 2004].

Theoretically, the extent of LD is related to the effective population size (*Ne*) [Wang *et al.* 2017b]. It is generally accepted that LD between markers and QTLs is a main source of information, which contributes to the accuracy of genomic prediction [Sun *et al.* 2016]. Jónás *et al.* [2017] reported that using LD information alongside with the genome to build haplotypes specifically for genomic prediction is a preferable step to increase the genomic prediction accuracy. However, Wientjes *et al.* [2013] indicated that LD has a small effect on predictive ability. Accuracies of the estimated genomic breeding value showed an increase alongside with the enlargement of LD size, especially for RF, which is in agreement with the simulated study by [Naderi *et al.* 2016]. Accuracy of the Bayes A was improved with an increase in LD of a historical population in the half-sib families [Sun *et al.* 2016]. A higher level of LD between QTLs and markers showed that more markers capture a higher proportion of the genetic variance [Goddard 2009], and are a prerequisite for an efficient performance of RF [Naderi *et al.* 2016].

Computational time

From the computational perspective, a wide variation was observed for the applied methods. Regarding computational time, GBLUP ranked first with 9 h per replicate, followed by TBA and RF with 12 and 24 h per replicate, respectively. Also concerning memory requirement, TBA, RF and GBLUP ranked from the highest to lowest with 8.3, 6.5 and 5.7 gigabytes, respectively.

In addition to the importance of models in achieving high accuracy, the computational aspects in this situation constitute new challenges from breeding and statistical viewpoints. For instance, Heslot *et al.* [2012] pointed out that BayesCpi should not be recommended for application in genomic selection despite the high accuracy in some traits, because it has a much greater computational cost compared to RR_GBLUP. They reported that the optimal method for genomic selection should be reliable as well as computationally efficient, while obviously being the most possibly accurate. In recent years, computing requirements have become more important than ever due to an increase in dimensionality of genomic selection programmes concerning the density of SNP chips and the number of genotyped individuals [Ober *et al.* 2012]. However, at continuous updating of computational systems its importance seems to be diminishing. In all the researches we studied [Neves *et al.* 2012; Ghafouri-Kesbi *et al.* 2017; Naderi *et al.* 2016], GBLUP was one of the most efficient methods regarding computational time, thus confirming our results. For example, Ghafouri-Kesbi *et al.* [2017] reported that computational time was reduced for GBLUP (10 min) compared

with RF (75 min). In the current study, computational time was the most considerable problematic issue of the RF. It is because when using machine learning models all the base learners in the ensemble have to be evaluated to obtain genomic predictions. This in turn is time-consuming, especially when the ensemble is noticeably large [Natekin and Knoll *et al.* 2013]. Therefore, a higher prediction accuracy of RF in scenario I was associated with the cost related with a longer computational time, which in turn may be a serious limitation when using RF methodology.

Conclusions

Genotype imputation can be reasonably applied to estimate the predictive ability of threshold methods, especially when sparse panels with high LD were used. The distribution of sick individuals into training sets slightly affected the predictive ability of RF, GBLUP and TBA. However, the accuracy of genomic prediction was greater when 20 percent of animals in the training set were sick. The structure of genomic architecture and accuracy of imputation were the most important factors when analysing discrete traits affecting accuracy of genomic prediction. For a scenario affected by a high number of QTLs and a high level of heritability, RF was more precise than the GBLUP and TBA methods. However, the cost of a longer computational time was a serious limitation when using RF methodology. Generally, genomic prediction accuracy of theTBA method was higher than those of the RF and GBLUP methods under different densities of the SNP panel. However, it seems that the use of imputed genotypes should be carefully evaluated, since the negative effect of increased imputation errors on the accuracies of genomic prediction in TBA was high.

REFERENCES

1. ANDONOV S., LOURENCO D., FRAGOMENI B., 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, 395-401.
2. 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. ***Journal of Animal Science*** 38, 447-53.
3. 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, 623-31.
4. BO Z., ZHANG J. J., HONG N., LONG G., PENG G., XU L.-Y., YAN C., ZHANG L.-P., GAO H.-J., XUE G., 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, 911-20.
5. BOISON S., NEVES H.H.D.R., O'BRIEN A.P., UTSUNOMIYA Y.T., CARVALHEIRO R., DA SILVA M., SÖLKNER J., GARCIA J.F., 2014 - Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. ***Livestock Science*** 166, 176-89.
6. CALUS M., VEERKAMP R., MULDER H., 2011 - Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework 1. ***Journal of Animal Science*** 89, 2042-9.

7. CARVALHEIRO R., BOISON S.A., NEVES H.H., SARGOLZAEI M., SCHENKEL F.S., UTSUNOMIYA Y.T., O'BRIEN A.M.P., SÖLKNER J., MCEWAN J.C., VAN TASSELL C.P., 2014 - Accuracy of genotype imputation in Nelore cattle. ***Genetics Selection Evolution*** 46, 69.
8. 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.
9. CLEVELAND M., HICKEY J., 2013 - Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. ***Journal of Animal Science*** 91, 3583-92.
10. COSTER A., BASTIAANSEN J.W., CALUS M.P., VAN ARENDONK J.A., 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, 9.
11. DAETWYLER H., HICKEY J., HENSHALL J., DOMINIK S., GREDLER B., VAN DER WERF J., HAYES B., 2010 - Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. ***Animal Production Science*** 50, 1004-10.
12. DAETWYLER H.D., WIGGANS G.R., HAYES B.J., WOOLLIAMS J.A., GODDARD M.E., 2011 - Imputation of missing genotypes from sparse to high density using long-range phasing. ***Genetics*** 189, 317-27.
13. DE LOS CAMPOS G., NAYA H., GIANOLA D., CROSSA J., LEGARRA A., MANFREDI E., WEIGEL K., COTES J.M., 2009 - Predicting quantitative traits with regression models for dense molecular markers and pedigree. ***Genetics*** 182, 375-85.
14. DRUET T., SCHROOTEN C., DE ROOS A., 2010 - Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. ***Journal of Dairy Science*** 93, 5443-54.
15. EFRON B., TIBSHIRANI R.J., 1993 - An introduction to the bootstrap. Number 57 in Monographs on statistics and applied probability. Chapman & Hall., New York.
16. FELIPE V.P., OKUT H., GIANOLA D., SILVA M.A., ROSA G.J., 2014 - Effect of genotype imputation on genome-enabled prediction of complex traits: an empirical study with mice data. ***BMC Genetics*** 15, 149.
17. GARRICK D., 2017 - The role of genomics in pig improvement. ***Animal Production Science*** 57, 2360-5.
18. 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, 229-36.
19. GODDARD M., 2009 - Genomic selection: prediction of accuracy and maximisation of long term response. ***Genetica*** 136, 245-57.
20. GODDARD M.E., HAYES B.J., 2009 - Mapping genes for complex traits in domestic animals and their use in breeding programmes. ***Nature Reviews Genetics*** 10, 381-91.
21. GONZÁLEZ-RECIO O., FORNI S., 2011 - Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. ***Genetics Selection Evolution*** 43, 7.
22. GUO Z., TUCKER D.M., BASTEN C.J., GANDHI H., ERSOZ E., GUO B., XU Z., WANG D., GAY G., 2014 - The impact of population structure on genomic prediction in stratified populations. ***Theoretical and Applied Genetics*** 127, 749-62.
23. HAYES B., BOWMAN P., DAETWYLER H., KIJAS J., VAN DER WERF J., 2012 - Accuracy of genotype imputation in sheep breeds. ***Animal genetics*** 43, 72-80.
24. HAYES B.J., BOWMAN P.J., CHAMBERLAIN A., GODDARD M., 2009 - Invited review: Genomic selection in dairy cattle: Progress and challenges. ***Journal of Dairy Science*** 92, 433-43.
25. HESLOT, N., YANG, H.-P., SORRELLS, M.E., JANNINK, J.L., 2012 - Genomic selection in plant breeding: a comparison of models. ***Crop Science*** 52, 146-160.

26. 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, 654-63.
27. HOZÉ C., FOUILLOUX M.-N., VENOT E., GUILLAUME F., DASSONNEVILLE R., FRITZ S., DUCROCQ V., PHOCAS F., BOICHARD D., CROISEAU P., 2013 - High-density marker imputation accuracy in sixteen French cattle breeds. ***Genetics Selection Evolution*** 45, 33.
28. 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, 2905-8.
29. KE X., HUNT S., TAPPER W., LAWRENCE R., STAVRIDES G., GHORI J., WHITTAKER P., COLLINS A., MORRIS A.P., BENTLEY D., 2004 - The impact of SNP density on fine-scale patterns of linkage disequilibrium. ***Human Molecular Genetics*** 13, 577-88.
30. 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, 538.
31. 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.
32. MADSEN P., JENSEN J., 2010 - A User's Guide to DMU: A Package for Analysing Multivariate Mixed Models. Version 5.0, release 6. Aarhus University, Foulum, Denmark.
33. MULDER H., CALUS M., 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, 876-89.
34. 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, 7261-73.
35. NADERI, S., BOHLOULI, M., YIN, T., KÖNIG, S., 2018 - Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets. ***Animal Genetics*** 49, 178-192.
36. NATEKIN A., KNOLL A., 2013 - Gradient boosting machines, a tutorial. ***Frontiers in Neurorobotics*** 7, 21.
37. 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.
38. OBER U., AYROLES J.F., STONE E.A., RICHARDS S., ZHU D., GIBBS R.A., STRICKER C., GIANOLA D., SCHLATHER M., MACKAY T.F., 2012 - Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. ***PLoS Genetics*** 8, e1002685.
39. 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, 3-12.
40. OGUTU J.O., PIEPHO H.-P., SCHULZ-STREECK T., 2011 - A comparison of random forests, boosting and support vector machines for genomic selection. In: ***BMC Proceedings*** p. S11. BioMed Central.
41. 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, 3.
42. 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, 24.

43. PURCELL S., NEALE B., TODD-BROWN K., THOMAS L., FERREIRA M.A., BENDER D., MALLER J., SKLAR P., DE BAKKER P.I., DALY M.J., 2007 - PLINK: a tool set for whole-genome association and population-based linkage analyses. ***The American Journal of Human Genetics*** 81, 559-75.
44. RESENDE M.F., MUÑOZ P., RESENDE M.D., GARRICK D.J., FERNANDO R.L., DAVIS J.M., JOKELA E.J., MARTIN T.A., PETER G.F., KIRST M., 2012 - Accuracy of genomic selection methods in a standard dataset of loblolly pine (Pinus taeda L.). ***Genetics*** 111.137026.
45. SARGOLZAEI M., CHESNAIS J., SCHENKEL F., 2011 - FImpute - An efficient imputation algorithm for dairy cattle populations. **Journal of Dairy Science** 94, 421.
46. SARGOLZAEI M., SCHENKEL F.S., 2009 - QMSim: a large-scale genome simulator for livestock. **Bioinformatics** 25, 680-1.
47. SARGOLZAEI M., VANRADEN P., KISTEMAKER G., SCHENKEL F., 2011 - Gebv software. L'alliance boviteq, sainthyacinthe, quebec and centre for genetic improvement of livestock, University of Guelph, Ontario, 27.
48. SUN C., WU X.-L., WEIGEL K.A., ROSA G.J., BAUCK S., WOODWARD B.W., SCHNABEL R.D., TAYLOR J.F., GIANOLA D., 2012 - An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. ***Genetics Research*** 94, 133-50.
49. 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, 77.
50. TOGHIANI S., AGGREY S., REKAYA R., 2016 - Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. ***Animal*** 10, 1077-85.
51. VANRADEN P.M., 2008 - Efficient methods to compute genomic predictions. ***Journal of Dairy Science*** 91, 4414-23.
52. VANRADEN P.M., O'CONNELL J.R., WIGGANS G.R., WEIGEL K.A., 2011 - Genomic evaluations with many more genotypes. ***Genetics Selection Evolution*** 43, 10.
53. VILLUMSEN T., JANSS L., LUND M., 2009 - The importance of haplotype length and heritability using genomic selection in dairy cattle. ***Journal of Animal Breeding and Genetics*** 126, 3-13.
54. 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, 213-9.
55. WANG C., HABIER D., PEIRIS B., WOLC A., KRANIS A., WATSON K., AVENDANO S., GARRICK D., FERNANDO R., LAMONT S., 2013 - Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. ***Poultry Science*** 92, 1712-23.
56. 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.
57. 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 Genetics*** 18, 45.
58. WANG Y., LIN G., LI C. STOTHARD P., 2016 - Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. ***Springer Science Reviews*** 4, 79-98.
59. WEIGEL K., DE LOS CAMPOS G., VAZQUEZ A., ROSA G., GIANOLA D., VAN TASSELL C., 2010 - Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. ***Journal of Dairy Science*** 93, 5423-35.
60. WIENTJES Y.C., VEERKAMP R.F., CALUS M.P., 2013 - The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. ***Genetics*** 193, 621-31.

61. YÁÑEZ J.M., HOUSTON R.D., NEWMAN S., 2014 - Genetics and genomics of disease resistance in salmonid species. ***Frontiers in Genetics*** 5, 415.
62. YIN T., PIMENTEL E., BORSTEL U.K.V., 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, 2444-54.
63. YOSHIDA G.M., CARVALHEIRO R., LHORENTE J.P., CORREA K., FIGUEROA R., HOUSTON R.D., YÁÑEZ J.M., 2018 - Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. ***Aquaculture*** 8(2), 719–726.
64. ZHANG Z., DRUET T., 2010 - Marker imputation with low-density marker panels in Dutch Holstein cattle. ***Journal of Dairy Science*** 93, 5487-94.