Felipe et al. BMC Genetics (2014) 15:149
DOI 10.1186/s12863-014-0149-9

BMC Genetics

RESEARCH ARTICLE Open Access

Effect of genotype imputation on genome-enabled prediction of complex traits: an empirical study with mice data

Vivian PS Felipe[1*], Hayrettin Okut[2], Daniel Gianola[1], Martinho A Silva[3] and Guilherme JM Rosa[1]

Abstract

Background: Genotype imputation is an important tool for whole-genome prediction as it allows cost reduction of individual genotyping. However, benefits of genotype imputation have been evaluated mostly for linear additive genetic models. In this study we investigated the impact of employing imputed genotypes when using more elaborated models of phenotype prediction. Our hypothesis was that such models would be able to track genetic signals using the observed genotypes only, with no additional information to be gained from imputed genotypes.

Results: For the present study, an outbred mice population containing 1,904 individuals and genotypes for 1,809 pre-selected markers was used. The effect of imputation was evaluated for a linear model (the Bayesian LASSO - BL) and for semi and non-parametric models (Reproducing Kernel Hilbert spaces regressions – RKHS, and Bayesian Regularized Artificial Neural Networks – BRANN, respectively). The RKHS method had the best predictive accuracy. Genotype imputation had a similar impact on the effectiveness of BL and RKHS. BRANN predictions were, apparently, more sensitive to imputation errors. In scenarios where the masking rates were 75% and 50%, the genotype imputation was not beneficial. However, genotype imputation incorporated information about important markers and improved predictive ability, especially for body mass index (BMI), when genotype information was sparse (90% masking), and for body weight (BW) when the reference sample for imputation was weakly related to the target population.

Conclusions: In conclusion, genotype imputation is not always helpful for phenotype prediction, and so it should be considered in a case-by-case basis. In summary, factors that can affect the usefulness of genotype imputation for prediction of yet-to-be observed traits are: the imputation accuracy itself, the structure of the population, the genetic architecture of the target trait and also the model used for phenotype prediction.

Keywords: Genotype imputation, Genome-enabled prediction, Complex traits, Non-linear models

Background

Genome-enabled prediction of quantitative traits is a topic of current interest in genetic improvement of agricultural animal and plant species, as well as in preventive and personalized medicine in humans. In agriculture, it has been applied to prediction of genetic merit for breeding purposes [1] and to management decisions based on predicted phenotypes [2,3]. In human medicine, it has been applied for example to prediction of risk to disease [4,5]. The original idea was proposed by Meuwissen et al. [6] and involves the use of prediction models including thousands of Single Nucleotide Polymorphisms (SNPs) fitted simultaneously as predictor variables, generally using shrinkage-based estimation techniques (e.g. [7]). The implementation of such models involves two steps. First, a group of individuals having both phenotypic and genotypic information (generally referred to as reference sample) is used to train the model. Cross-validation techniques can be used to compare different models. Secondly, the trained model is applied to a group of individuals with genotypic information only (the target sample), for prediction of their genetic merit or of their yet-to-be-observed phenotypes.

A commonly used technique in this field is genotype imputation. Genotype imputation can be employed to fill in missing data from the laboratory or allow merging data sets generated from different SNP chips. Genotype

* Correspondence: vfelipe@wisc.edu
[1]Department of Animal Sciences, University of Wisconsin, Madison 53706, USA
Full list of author information is available at the end of the article

imputation has been proposed also to impute from genotypes scored with low-density chips to higher densities, as a way to reduce genotyping costs [3,8,9]. Other authors have proposed to use cosegregation information from chips built with evenly spaced low-density SNPs or SNPs selected by their estimated effects to track signals of high density SNP alleles [10]. Weigel et al. [11] showed that a low-density panel containing selected SNPs can retain most of the prediction ability of high-density panels. Furthermore, in a later study, Weigel et al. [3] also showed that imputed genotypes can provide similar levels of predictive ability to those derived from high density genotypes in scenarios where a suitable reference population is available.

The benefit of imputing genotypes essentially depends on its imputation accuracy [3], which, in turn, depends on a number of factors including population structure [3,12], and genetic architecture of the target trait [13]. Many studies have shown that currently available imputation methods and software give a satisfactory level of accuracy of uncovering unknown genotypes [8,14-16]. Hence, imputation may provide a suitable alternative for reducing genotyping costs, and it has been suggested for commercial applications such as the pre-screening of young bulls and heifers in dairy cattle [3]. Moreover, VanRaden et al. [17] reported that the reliability of genomic predictions can be improved at a lower cost by combining information from chips containing varied marker densities, to increase both the number of markers and animals included in genome-based evaluation.

So far, all studies conducted to evaluate the effect of genotype imputation on whole-genome prediction have assumed a linear relationship between phenotype and genotype, aimed at capturing additive genetic effects only. However, complex traits are known to be affected by complex gene effects and interactions [18]. For this reason, interest in non- and semi-parametric methods for prediction of complex traits using genomic information has been increasing. Such methods include Reproducing Kernel Hilbert Spaces (RKHS) regressions on markers [19-21] radial basis functions [22,23], and artificial neural networks [24,25]. Gianola et al. [24] argued that these non-parametric regressions can capture complex interactions and nonlinearities, which is not possible with Bayesian linear regressions commonly used in genomic prediction.

Recently, Heslot et al. [26] evaluated the prediction accuracy of several models including Bayesian regression methods and machine learning techniques. Their results indicated a slight superiority of non-linear models for phenotype prediction in plants. As another example, Okut et al. [25] used Bayesian Regularized Neural Networks (BRANN) to predict body mass index (BMI) in mice using information on 798 SNPs, and obtained an overall correlation between observed and predicted data that varied between 0.25 and 0.3. Similar results were obtained by de los Campos et al. [27] using a Bayesian LASSO approach but using a panel that was 13 times larger, comprising 10,946 SNPs. Perez-Rodriguez et al. [28] compared linear and nonlinear models for genome-enabled prediction in wheat and showed that nonlinear models in general performed better. However, the author found that in this case the BRANN did not outperformed the BL. Lastly, Howard et al. [29] indicated a clear superiority of RKHS when predicting epistatic traits using simulation.

The objective of our study was to investigate the effect of genotype imputation in the context of whole-genome prediction of complex traits in mice using parametric, semi-parametric and non-parametric models applied to different sizes of subsets of SNPs. Our underlying hypothesis was that more elaborated prediction models, such as those capable to accommodate non-additive genetic effects, would not benefit significantly from genotype imputation for prediction of yet-to-be-observed phenotypes.

Results

Results indicated a good accuracy of imputation of unknown genotypes for all scenarios (Table 1). The lowest imputation accuracy (0.75) was for the scenario with approximately 90% of the genotypes masked and the reference panel was not related to the imputing set. Although Beagle software does not use pedigree information, a higher genetic relatedness among individuals in the reference panel and in the set containing missing genotypes can enhance imputation accuracy. The explanation is that similarity of linkage disequilibrium (LD) patterns between the set to be imputed and the reference panel serves as a basis for imputing the unknown genotypes. The most common error found was the switch between heterozygotes and homozygotes for the allele at higher frequency (about 65%).

Correlations between predicted and observed phenotypes in the testing set are shown in Tables 2 and 3 for body weight (BW) and body mass index (BMI), respectively. The distribution of individuals into training and testing sets affected the predictive ability of all models considered. A higher genetic relatedness between these two sets provided better prediction accuracy for BW. On the other hand, for BMI, the average correlation between predicted and observed phenotypes was higher for the across families layout. Therefore, information from closely related individuals for SNP effect estimation was beneficial for prediction of new phenotypes, at least for BW.

As expected, the predictive ability for BW was higher than for BMI, since the latter has a lower heritability. Differences on results for each trait are also probably due to differences between their underlying genetic

Table 1 Overall imputation accuracy and error distribution for 90, 75 and 50% of masked genotypes

	90%		75%		50%	
	Across families	Within families	Across families	Within families	Across families	Within families
Accuracy	0.75	0.79	0.91	0.94	0.97	0.98
0*<–>1* error[a]	0.16	0.17	0.22	0.25	0.26	0.20
1<–>2* error[b]	0.50	0.54	0.61	0.63	0.62	0.65
0<–>2 error[c]	0.09	0.08	0.08	0.06	0.09	0.13

[a]Error due to change from 0 to 1 genotype code or vice versa.
[b]Error due to change from 1 to 2 genotype code or vice versa.
[c]Error due to change from 0 to 2 genotype code or vice versa.
*Genotypes are coded as 0, 1 and 2 as the number of copies of the more frequent allele.

architectures. As discussed by Legarra et al. [30], in this data set there is some confounding between family and cage effects since most animals allocated to the same cage were full sibs, so it is possible that the additive genetic effect is understated. For the present study however, it is reasonable to assume that this issue would impact the predictive ability of the different models considered in a similar way.

In general, the method with the best prediction results was RKHS using kernel averaging, and the worst was BRANN, probably due to overfitting. BRANN showed high correlation (above 0.9) between predicted and measured phenotype for the training sets (results not shown). Table 2, which describes results for BW, shows that imputation seemed to be beneficial for phenotype prediction when relatedness between reference and target samples was poorer, especially for BL and RKHS. Table 3, in contrast, shows a markedly noticeable benefit of imputation when the number of markers available in the testing set was low (201 SNPs) for the within-family layout when predicting BMI. Regarding the methods, imputation seemed to have similar impact on efficiency of BL and RKHS, whereas for BRANN it resulted in less robust predictions due to imputation error. In scenarios with good imputation accuracy and masking rates of 75% and 50%, the genotype imputation did not bring great benefit, as seen in Tables 2 and 3. However, when genotype information was sparse (90% masking rate – 201 observed genotypes) imputation could bring

Table 2 Correlations between predicted and observed body weight for all masking rates and family layouts

90% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[a]	201	1809	1809i[a]	201
BL	0.347	0.259	0.169	0.500	0.330	0.407
RKHS	0.347	0.312	0.210	0.527	0.417	0.499
BRANN	0.330	0.217	0.144	0.490	0.274	0.392
75% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[b]	453	1809	1809i[b]	453
BL	0.343	0.291	0.262	0.499	0.447	0.430
RKHS	0.348	0.317	0.293	0.528	0.506	0.501
BRANN	0.320	0.241	0.255	0.492	0.414	0.428
50% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[c]	905	1809	1809i[c]	905
BL	0.342	0.324	0.271	0.499	0.496	0.477
RKHS	0.343	0.345	0.306	0.530	0.530	0.520
BRANN	0.320	0.281	0.252	0.492	0.478	0.461

[a]Imputed from 201 SNPs.
[b]Imputed from 453 SNPs.
[c]Imputed from 905 SNPs.
*BL: Bayesian LASSO; RKHS: Reproducing Kernel Hilbert Spaces (RKHS) and; BRANN: Bayesian Regularized Neural Networks.

Table 3 Correlations between predicted and observed body mass index for all genotype masking rates and family layouts

90% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[a]	201	1809	1809i[a]	201
BL	0.227	0.193	0.191	0.199	0.164	−0.047
RKHS	0.238	0.195	0.199	0.208	0.132	−0.054
BRANN	0.112	0.092	0.147	0.163	0.041	0.054
75% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[b]	453	1809	1809i[b]	453
BL	0.228	0.219	0.199	0.200	0.196	0.184
RKHS	0.238	0.226	0.211	0.208	0.204	0.200
BRANN	0.118	0.115	0.145	0.172	0.154	0.170
50% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[c]	905	1809	1809i[c]	905
BL	0.227	0.231	0.225	0.199	0.197	0.189
RKHS	0.238	0.238	0.236	0.207	0.206	0.202
BRANN	0.118	0.131	0.149	0.172	0.168	0.149

[a]Imputed from 201 SNPs.
[b]Imputed from 453 SNPs.
[c]Imputed from 905 SNPs.
*BL: Bayesian LASSO; RKHS: Reproducing Kernel Hilbert Spaces (RKHS) and; BRANN: Bayesian Regularized Neural Networks.

information about important markers to improve phenotypic prediction.

The results for predicted mean squared error (PMSE) are summarized in Tables 4 and 5 for BW and BMI, respectively. For BW, the lowest values of PMSE were found for predictions made within families with the full data set (1,809 SNPs). This agrees with the results obtained for predictive correlation described earlier. In general, higher masking rates resulted in a higher PMSE for BW and data containing imputed genotypes provided a better goodness of fit compared to the data with no genotype imputation when markers were masked. With BMI, however, the PMSE showed no changes according to genotype masking rates or genotype imputation for BL and RKHS models. Overall, BRANN had the highest PMSE values, in agreement with the results using correlation between observed and predicted phenotypes.

Discussion

Recently, some studies have investigated the predictive ability of models using subsets of SNPs, with and without imputation [8,31,32]. In general, predictive ability improved with imputed genotypes, such that many researchers recommend this strategy to decrease costs on genomic selection programs. However, most studies with genotype imputation in whole-genome predictions considered only linear models, such as ridge regression, Bayesian LASSO or GBLUP approaches [3,8,12] specifically suited to model additive genetic signals but not tailored to capture non-additive genetic effects such as dominance and epistasis. The goal of our study was to explore if more elaborated models, such as semi-parametric and non-parametric methods, could track genetic signals from low-density chips without the need of imputing to higher density chips.

The results obtained indicated that imputation of the missing genotypes was not always advantageous for phenotypic prediction. The benefit of imputing genotypes depended on the degree of relatedness between reference and target samples, genetic architecture of the trait, number of markers available in the original panel, and the method used to predict marker effects.

Weigel et al. [3] investigated the effect of imputation from a low-density chip to a 50K chip on the accuracy of direct genomic values in Jersey cattle using BL. They found that genotype imputation improved predictive ability in scenarios where imputation accuracy was high; otherwise, a reduced panel containing the original number of SNPs was preferred. In the same context, Mulder et al. [8] showed that due to the magnitude of imputation errors, the noise added by imputation can be greater than its benefit when predicting breeding values.

Table 4 Prediction mean squared errors for body weight analysis by family layouts and genotype masking rates

90% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[a]	201	1809	1809i[a]	201
BL	5.03	5.32	5.67	4.18	4.99	4.71
RKHS	4.92	5.20	5.36	4.15	4.75	4.66
BRANN	5.36	5.52	5.54	5.26	5.40	5.52
75% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[b]	453	1809	1809i[b]	453
BL	5.05	5.25	5.44	4.18	4.45	4.52
RKHS	4.92	5.04	5.11	4.13	4.23	4.21
BRANN	5.38	5.44	5.44	5.26	5.32	5.33
50% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[c]	905	1809	1809i[c]	905
BL	5.06	5.12	5.49	4.18	4.19	4.32
RKHS	4.94	4.94	5.01	4.06	4.08	4.12
BRANN	5.20	5.24	5.44	5.26	5.27	5.28

[a]Imputed from 201 SNPs.
[b]Imputed from 453 SNPs.
[c]Imputed from 905 SNPs.
*BL: Bayesian LASSO; RKHS: Reproducing Kernel Hilbert Spaces (RKHS) and; BRANN: Bayesian Regularized Neural Networks.

Table 5 Prediction mean squared errors for body mass index analysis by family layouts and genotype masking rates

90% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[a]	201	1809	1809i[a]	201
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
BRANN	0.013	0.014	0.010	0.042	0.036	0.024
75% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[b]	453	1809	1809i[b]	453
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
BRANN	0.021	0.023	0.015	0.044	0.045	0.041
50% genotype masking rate						
Model*	**Across families**			**Within families**		
	1809	1809i[c]	905	1809	1809i[c]	905
BL	0.002	0.002	0.002	0.002	0.002	0.002
RKHS	0.002	0.002	0.002	0.002	0.002	0.002
BRANN	0.021	0.023	0.016	0.040	0.040	0.047

[a]Imputed from 201 SNPs.
[b]Imputed from 453 SNPs.
[c]Imputed from 905 SNPs.
*BL: Bayesian LASSO; RKHS: Reproducing Kernel Hilbert Spaces (RKHS) and; BRANN: Bayesian Regularized Neural Networks.

Hence, only those SNPs with high imputation accuracy would have a positive effect on the reliability of direct genomic value predictions. In the present study, results also suggested that if imputation accuracy was low, the model containing only observed marker genotypes gave a better prediction than the imputed set. The correlation between predicted and measured BW within families using either a full data set containing 1,809 genotyped SNPs, or the full data set containing 90% imputed genotypes, or a reduced panel of marker genotypes (201 SNPs) was respectively 0.52, 0.42 and 0.50 using RKHS. This indicates that imputation brought no additional information to the model.

For scenarios with different masking rates the imputed testing set gave, on average, a 4% higher correlation. For BMI, the reduced testing sets (201, 453 or 905 SNPs) provided 89% of the predictive ability of their respective complete imputed testing sets and 78% of the predictive ability of the complete testing sets, averaged across all scenarios tested. So, in general, the results indicated that imputation can be useful for phenotypic prediction.

When comparing correlations for across and within families cross-validation strategies, genotype imputation seemed to be more effective in improving prediction accuracy in cases where there was a weaker genetic relationship among individuals in the reference and testing data sets. Other studies regarding the role of within and across-family information [30] also indicate the need of genotyping and phenotyping closely related individuals, in order to improve predictive ability. As such, this information is an important issue for designing genome-assisted breeding programs.

Regarding the models considered, it was expected that the non-parametric methods would give smaller differences between the complete set with imputed markers and the reduced panel. However, our results indicated that the effect of imputation was similar for BL and RKHS predictions. An exception was the case of BRANN, which was not able to cope with imputation errors and tended to give worse predictions for the complete testing set containing imputed markers. Therefore, it seems that imputation accuracy is a fundamental factor to be considered when using BRANN for predicting phenotypes. The imputation from 905 markers to the full panel (1,809 SNPs) tended to slightly improve prediction using BRANN perhaps due to the low imputation errors rates for these panels.

Another discussion, beyond the scope of this paper, is on differences between chips containing either equally spaced SNPs or SNPs pre-selected based on their estimated effects for genome-enabled prediction (e.g., [33]). The main advantage of the former is that it avoids the need of trait-specific low-density SNP panels and, in general, it has given reliability of genomic breeding values similar to the latter [13]. Comparing the results obtained with the available literature on genomic selection applied to this same data set, it was found that no important differences in predictive ability were observed when using the entire set of SNPs. For example, de los Campos [27] used 10,946 SNPs with a BL model and observed a rank correlation of 0.306 between phenotypic observations and genomic predictions for BMI. Here, we obtained almost 95% of this correlation using the same method but with only 1,809 evenly spaced SNPs. In addition, Okut et al. [25] reported a correlation between predictions and observations in the testing set of 0.18 for BMI using BRANN and 798 pre-selected markers. We obtained a correlation of 0.15 with the same model and 905 evenly spaced markers, which suggests that BRANN can work better using selected markers with larger effects.

Similar results were observed in terms of PMSE. Apparently, higher imputation errors caused higher values of PMSE, making the results from models using the reduced SNP panel better than those containing imputed marker genotypes.

The results of the present study can be generalized for different scenarios, regardless the number of SNPs and/or sample size of a particular study, based on the impact of imputation accuracy on the predictive quality of genomic models. Clearly, the predictive ability of a model not only depends on how well genotypes are imputed but also on the genetic architecture of the target trait and the breeding program design. Therefore, the general reasoning provided by the results of the present study is that the use of genotype imputation should the evaluated in a case-by-case basis. For example, the use of imputed genotypes when employing the non-parametric method (BRANN, in this case) is not recommended given that this model tends to approximate the noise inserted by imputation errors.

Conclusions

Genotype imputation did not always improve the predictive ability of parametric and semi-parametric models. For BW, genotype imputation improved predictive ability when there was a relatively low genetic relatedness between the reference panel and the target population set. For BMI, the use of genotype imputation was more beneficial when the genotype set was very sparse (201 SNPs), especially for BL and RKHS. In other scenarios, imputation just slightly improved or even deteriorated predictive ability; the latter happened in cases in which the genotype imputation had low accuracy. Lastly, BRANN seemed more sensitive to imputation errors; therefore the use of imputed genotypes with this model should be carefully evaluated when using neural networks.

Methods

Data

A publicly available dataset on mice (http://mus.well.ox.ac.uk/mouse/HS/) was used. This is a sample from an outbred mice population that descended from eight inbred strains created for fine-mapping QTL and high-resolution whole-genome association analysis of quantitative traits [34]. The data set contains genotypic information from 1,904 fully pedigreed mice on 13,459 SNPs coded as 0, 1 and 2 as the number of copies of the more frequent allele. Traits such as weight, immunology, obesity and behavior, to name a few, are also available for a proportion of these animals. A full description of this mice population is in [35] and [36]. This data have also been utilized in genomic-enabled prediction studies using Bayesian regression methods [2,27,30,37] and neural networks [25].

In our analysis, only animals with both phenotypic and genotypic information were considered. Loci with a minor allele frequency lower than 0.05, a call rate lower than 95% or not in Hardy-Weinberg equilibrium ($p<0.01$) were discarded from the original dataset. The two traits, BW at ten weeks of age, and body mass index BMI were pre-corrected by fitting the following linear mixed model:

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{Wc} + \boldsymbol{Zu} + \boldsymbol{e},$$

where y is the vector of observations on one of the measured phenotypes (BW or BMI); θ is an unknown vector of fixed effects of age, gender, month and cage density; c is a random vector of unknown cage effects; u is a random vector of unknown additive genetic effects; $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{Z}$ are the incidence matrices of fixed, random cage and additive genetic effects, respectively, and e is a vector of residual effects assumed to follow a multivariate normal distribution $\boldsymbol{e} \sim N\left(\boldsymbol{0}, \boldsymbol{I\sigma_e^2}\right)$, where σ_e^2 is the residual variance. The random additive genetic and cage effects were assumed independent from each other and with distributions $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \boldsymbol{A\sigma_u^2}\right)$ and $\boldsymbol{c} \sim N\left(\boldsymbol{0}, \boldsymbol{I\sigma_c^2}\right)$, respectively, where A is the additive genetic relationship matrix, I is an identity matrix of appropriate order, and σ_u^2 and σ_c^2 are additive genetic and cage components of variance, respectively. The target response variable after correction was $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} - \mathbf{W}\hat{\mathbf{c}}$, which presumably includes all types of genetic effects (additive, dominance and epistasis) as well as additional environmental effects not accounted for by the mixed model employed. From now on the pre-corrected phenotype $\mathbf{y}^*$ will be simply referred to as $\mathbf{y}$.

After data cleaning, 10,348 SNPs remained from which 1,809 equally spaced SNPs were selected and regarded as full genotyped data due to computational limitations on number of markers that can be fitted when using Bayesian Regularized Neural Networks. Then, subsets containing 905, 453 and 201 (50, 75 and 90% masking rates, respectively) equally spaced SNPs were taken from the full genotype set. In total, 1,881 and 1,823 individuals were included in the analysis of BW and BMI, respectively. For a cross-validation (CV) model comparison, in each case, approximately 2/3 of the individuals were designated as training set (reference sample) and 1/3 as testing set (target sample) (See Table 6). Two CV scenarios were considered, denoted as "across" and "within" families as also applied by [30]. In the across families approach, whole families were randomly assigned to training and testing sets, whereas in the within families approach, individuals from each family were randomized to training and testing sets. Subsequently, phenotypic predictions were performed using the three methods (BL, RKHS and BRANN) for both traits and for data sets containing either the full genotype set or subsets (201, 453 or 905 SNPs), with or without genotype imputation. Details on the imputation approach and models considered are provided below.

Imputation

Testing sets containing 201, 453 and 905 SNPs were imputed to 1,809 SNPs using the Beagle software [38]. This software is based on Hidden Markov Models that cluster haplotypes at each locus. The clustering adapts to the amount of information available so that the number of clusters increases globally with sample size and locally with increasing linkage disequilibrium levels [14]. The training set, which contained 1,809 markers, was used as a reference sample for imputation of SNPs in the testing set. Imputation was carried out for both prediction scenarios ("across" and "within") using only population structure and ignoring pedigree information. To check the global imputation accuracy, the imputed sets were compared with the full data set to calculate the percentage of correctly imputed genotypes.

Table 6 Number and distribution of individuals by trait and cross validation strategy employed

Trait*	Across families		Within families		Total no. of individuals
	Training set	Testing set	Training set	Testing set	
BW	1,200	681	1,200	681	1,881
BMI	1,165	658	1,161	662	1,823

*BMI: body mass index; BW: body weight at ten weeks of age.

Bayesian LASSO

Tibshirani [39] proposed a regression method called Least Angle Shrinkage Selection Operator (LASSO) that combines feature subset selection and shrinkage estimation. In this model, a penalty term proportional to the norm of regression coefficients is added to the optimization problem formula, allowing for variable selection and shrinkage of coefficients simultaneously. The optimization problem can be expressed as:

$$\min_{\beta}\left\{\sum_{i}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2+\lambda\sum_{j}\left\|\beta_j\right\|\right\},$$

where $\sum_{i}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2$ is the residual sum of squares and $\lambda\sum_{j}\left\|\beta_j\right\|$ is the penalization factor, with $\boldsymbol{x_i}$ and $\boldsymbol{\beta}$ representing the incidence and parameter vectors, respectively, and λ is a regularization parameter. A larger λ means stronger shrinkage and some β's are even zeroed out.

A Bayesian version of the LASSO was proposed by [40], who described a Gibbs sampling implementation. In this Bayesian interpretation, the LASSO solution can be viewed as a conditioned posterior mode in a Bayesian model with Gaussian likelihood, $p\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma_e^2\right)=\prod_{i=1}^{n}N\left(y_i|\boldsymbol{x}_i'\boldsymbol{\beta},\sigma_e^2\right)$ and a conditional (given λ) prior on $\boldsymbol{\beta}$ that is a product of p independent, zero mean, double-exponential (DE) densities [40]. The double-exponential (or Laplace) distribution has a convenient hierarchical representation as a mixture of scaled Gaussian densities (e.g., [41]), i.e.:

$$\begin{aligned}\beta_j\sim DE(\beta_j|\lambda)&=\frac{\lambda}{2}e^{-\lambda|\beta_j|}\\&=\int_0^{\infty}\left[\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\left(\beta_j^2/2\sigma_j^2\right)}\right]\left[\frac{\lambda^2}{2}e^{-\lambda^2/2\sigma_j^2}\right]d\sigma_j^2.\end{aligned}$$

Convenient priors for the parameters of the Bayesian LASSO (BL) model have been suggested by [27] as:

$$\begin{aligned}p\left(\beta,\sigma_\varepsilon^2,\tau^2,\lambda^2|H\right)&=p\left(\boldsymbol{\beta}|\sigma_\varepsilon^2,\boldsymbol{\tau^2}\right)p(\sigma_\varepsilon^2)p\left(\tau^2|\lambda\right)p\left(\lambda^2|\alpha_1,\alpha_2\right)\\&=\left[\prod_{j=1}^{p}N(\beta_j|0,\tau_j^2\sigma_\varepsilon^2)\right]\chi^{-2}\left(\sigma_\varepsilon^2|d.f.,S\right)\\&\quad x\left[\prod_{j=1}^{p}\exp\left(\tau_j^2|\lambda\right)\right]G\left(\lambda^2|\alpha_1,\alpha_2\right)\end{aligned}$$

where H is a set of hyper-parameters. Here, $p(\boldsymbol{\beta}|\sigma_\varepsilon^2,\boldsymbol{\tau^2})=\prod_{j=1}^{p}N\left(\beta_j|0,\tau_j^2\sigma_\varepsilon^2\right)$ is the product of p normal densities with zero mean and variance $\tau_j^2\sigma_\varepsilon^2$ relative to each marker effect j. Further $p\left(\sigma_\varepsilon^2|d.f.,S\right)$ is a scaled inverted chi-square distribution $\chi^{-2}\left(\sigma_\varepsilon^2|d.f.,S\right)$ with *d.f.* degrees of freedom and scale parameter S; $\exp(\tau_j^2|\lambda)$ is an exponential distribution, and $p(\lambda^2|\alpha_1,\alpha_2)$ is a Gamma distribution with parameters α_1 and α_2. The parameter λ, also called smoothing parameter, plays a central role in the model as it controls the trade-off between goodness of fit and model complexity [39]. As its value approaches 0, the solution approximates a least squares solution; a large value of λ induces a sharper prior on β and, consequently, stronger shrinkage. Compared to Bayesian Ridge Regression, this model has the advantage of assigning a higher density to markers with zero effects, which seems biologically plausible [27].

The model was fitted to the training set in all scenarios considered. Inferences were based on a Gibbs sampling chain with 70,000 samples after a burn-in of 5,000. The parameters of the prior distribution were $S_\varepsilon=d.f._\varepsilon=S_u=d.f._u=1$, and $\alpha_1=1.2$ and $\alpha_2=10^{-5}$. The package BLR [42] developed for the R software was used for the analysis. Fitted models were then used to predict phenotypes in the testing set, and their predictive ability was assessed by the correlation between measured and predicted phenotypes, and by the PMSE.

Reproducing Kernel Hilbert spaces regression

The RKHS theory was introduced by Aronszajn [43] and has been applied in statistics and machine learning (e.g., Support Vector Machines) fields for many years; foundations are provided in [44]. This semi-parametric approach was proposed by Gianola et al. [19,45] for regressing phenotypes on genotypes. The RKHS method has the property of having an infinite space of functions for searching the dependency between input and target variables, and the space is defined by the measure of distance used (in this case the type of kernel), without any additional assumptions on gene action or functional form. The method can be seen as a combination of the classical additive genetic model with an unknown function of markers, which is inferred nonparametrically, and has the potential of capturing complex interactions without explicitly modeling them [45]. To map the relationship between inputs (genotypes) and targets (phenotypes), a collection of functions defined in a Hilbert space (say $f\in H$) is used, from which an element, $\hat{f}$, is chosen based on some criterion (e.g. penalized residual sum of squares or posterior density) [20]. The optimization problem for obtaining the estimates of RKHS is:

$$\hat{f}=\underset{f\in H}{\arg\min}\left\{l(f,y)+\lambda\|f\|_H^2\right\},$$

where $l(f,y)$ is a loss function representing a measure of goodness of fit; $\|f\|_H^2$ is the squared norm of f, related to

model complexity, and λ controls the trade-off between goodness of fit and model complexity.

According to the Moore-Aronszajn theorem [43], each RKHS is associated to a unique positive definite kernel. In RKHS, the markers are used to build a covariance or similarity matrix that measures distances between genotypes. Here, $Cov(g_i, g_{i'}) \sim K(\mathbf{x_i}, \mathbf{x_{i'}})$, with $\mathbf{x_i}$ and $\mathbf{x_{i'}}$ representing vectors containing genotypes for the ith and i'th individuals, and $K(.,.)$ is the Reproducing Kernel (RK) related to a positive definite function [20].

The Kernel matrix (K) employed here was a Gaussian kernel, i.e. $K(\mathbf{x_i}, \mathbf{x_{i'}}) = \exp\{-h \times d_{ii'}\}$, where h is a bandwidth parameter and $d_{ii'} = \sum_{k=1}^{p}(x_{ik} - x_{i'k})^2$ represents an element of the matrix of squared Euclidean distances among the individuals in the sample. The choice of h is a model selection issue and must consider the observed distribution of $d_{ii'}$. In this study we used "kernel averaging" (multi-kernel fitting) as an automatic way of choosing the kernel based on the sample median of $d_{ii'}$ as described by [46]. Hence, $h = a \times q_{0.5}^{-1}$ in which a was −5, −1 and −1/5, and $q_{0.5}$ is the sample median of $d_{ii'}$ for the three kernels used for kernel averaging. In this model, the genotypic values were the sum of three components, $g = f_1 + f_2 + f_3$, with $p(f_1, f_2, f_3 | \sigma^2_{\alpha,1}, \sigma^2_{\alpha,2}, \sigma^2_{\alpha,3}) = N(f_1 | 0, K_1\sigma^2_{\alpha,1}) N(f_2 | 0, K_2\sigma^2_{\alpha,2})\ N(f_3 | 0, K_3\sigma^2_{\alpha,3})$. The variance parameters for these components were treated as unknown and assigned identical and independent scaled inverse chi-square prior distributions with degrees of freedom and scale parameters equal to $df = 5$ and $S = (\mathrm{var}(y)/2 \times (df - 2))$, respectively. Posterior distribution samples were obtained with a Gibbs sampler as described by de los Campos et al. [20]. Inferences were based on 50,000 samples after 5,000 samples of burn-in.

Bayesian regularized artificial neural networks

A Bayesian Regularized Artificial Neural Network (BRANN) is a feed-forward network implemented with a maximum a posteriori approach in which the regularizer is the logarithm of the density of a prior distribution [47]. This model assigns a probability distribution to the network weights and biases, so that predictions are made in a Bayesian framework and generalization is improved over predictions made without Bayesian regularization. Details are in [48].

A basic feed-forward network uses initial weights and biases and transforms input information (in this case, genotype codes) through each given connected neuron in the hidden layer using an activation function. Information is then sent to the neuron in the output layer using another activation (transformation) function generating the output or predicted value. Next, the results are backpropagated (non-linear least-squares) in order to update weights and biases using derivatives. Therefore, no assumptions about the relationship between genotypes (input) and phenotypes (target) are made in this model. After training, outputs are calculated as:

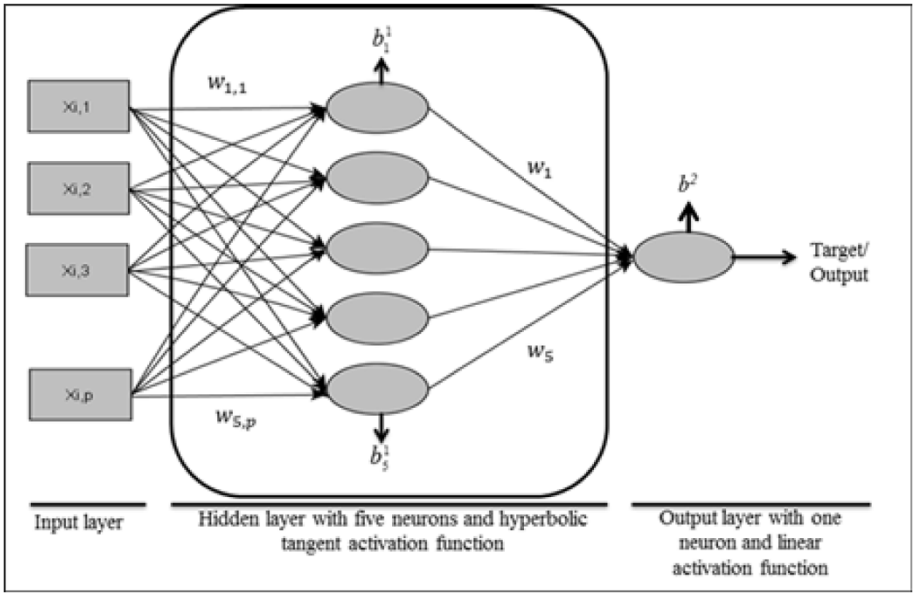

Figure 1 Artificial Neural Network architecture with two layers containing 5 neurons in the hidden layer and one neuron in the output layer. The $x_{i,p}$ are the inputs for each animal i, and p is the number of SNPs; the $w_{k,j}$ are the weights where k is the hidden layer neuron indicator and j is the index for SNP; b_k^l are the hidden layer biases, where k and l are the indexes for neurons and layers, respectively, and b^2 is the output neuron bias.

$$\hat{y}_i = g\{\sum_{k=1}^{s} w_k f(\sum_{k=1}^{R} w_{k,i} \underset{\sim}{x}_i + b_k^l) + b^2\},$$

where $\hat{y}_i$ is the predicted phenotype for an individual and $\underset{\sim}{x}_i$ are the input genotypes; g and f are the activation functions for output and hidden layers, respectively; w_k and $w_{k,i}$ are the weights from neurons of the hidden to the output neuron, and from the input to the hidden neurons, respectively, and b_k^1 and b^2 are the biases of the two layers. Training is the process by which the weights are modified in light of the data while the network attempts to produce an optimal outcome [25]. After training, the network can then be used to predict unknown phenotypes from individuals with genotype information.

In BRANN, in addition to the loss function given by the sum of squared errors, a penalty to large weights is also included in order to have a smoother mapping (regularization). The objective function is:

$$f = \gamma E_D(D|\underset{\sim}{w}, M) + \alpha E_w(\underset{\sim}{w}|M),$$

where $E_D(D|\underset{\sim}{w}, M)$ is the sum of squares of residuals in which D is the data (input data and target variable), $\underset{\sim}{w}$ are the weights and M is the architecture of the neural network. Further, $E_w(\underset{\sim}{w}|M)$ is known as weight decay which is calculated as the sum of squares of weights of the network, and α and γ are the regularization parameters that control the trade-off between goodness of fit and smoothing.

The posterior distribution of w given α, γ, D and M is [49]:

$$P(w|D, \alpha, \gamma, M) = \frac{P(D|w, \gamma, M)P(w|\alpha, M)}{P(D|\alpha, \gamma, M),}$$

where $P(D|w, \gamma, M)$ is the likelihood function, $P(w|\alpha, M)$ is the prior distribution on weights under the chosen architecture, and $P(D|\alpha, \gamma, M)$ is the normalization factor.

To assess overfitting, network architectures and number of epochs (iterations) were tested in a first step. A network containing 5 neurons in the hidden layer with a tangent sigmoid function and 1 neuron in the output layer with a linear function was used after such tests (Figure 1). The number of epochs was set to 30. Results were the average of 20 repetitions of the analysis with different randomly generated starting values. As an attempt to improve generalization, use of early stopping was tested for regularization, but Bayesian regularization worked better. The software MATLAB [50] was used for the analysis. The predictive ability was also assessed by correlation between estimated and measured phenotypes, and by PMSE, as it was for BL and RKHS.

Availability of supporting data

The data set supporting the results of this article is available in the http://gscan.well.ox.ac.uk/.

Abbreviations

BMI: Body mass index; BL: Bayesian LASSO; BRANN: Bayesian Regularized Artificial Neural Networks; BW: Body weight; CV: Cross-validation; LASSO: Least angle shrinkage selection operator; LD: Linkage disequilibrium; PMSE: Prediction mean squared error; RKHS: Reproducing kernel Hilbert spaces regression; SNP: Single nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VF, HO, DG, MS and GR contributed to the concept, design, execution, and interpretation of this work. VF conducted the statistical analyses and drafted the first version of the manuscript. HO assisted with the neural network implementation. VF, HO, DG, MS and GR read and approved the final manuscript.

Acknowledgements

Financial support by the Wisconsin Agriculture Experiment Station, by COBB-Vantress, Inc. (Siloam Springs, AR) and by the National Council of Scientific and Technological Development (CNPq, Brazil) is acknowledged. We also would like to extend our thanks to The Welcome Trust Centre for Human Genetics for making the mice data available.

Author details

[1]Department of Animal Sciences, University of Wisconsin, Madison 53706, USA. [2]Department of Animal Sciences, Biometry and Genetics Branch, University of Yuzuncu Yil, Van 65080, Turkey. [3]Department of Animal Sciences, Federal University of Jequitinhonha and Mucuri Valleys, Minas Gerais, Brazil.

Received: 4 September 2014 Accepted: 10 December 2014
Published online: 29 December 2014

References

1. Goddard ME, Hayes BJ: **Genomic selection.** *J Anim Breed Genet* 2007, **124**(6):323–330.
2. Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM: **Predicting unobserved phenotypes for complex traits from whole-genome SNP data.** *PLoS Genet* 2008, **4**(10):e1000231.
3. Weigel KA, De Los Campos G, Vazquez AI, Rosa GJM, Gianola D, Van Tassell CP: **Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle.** *J Dairy Sci* 2010, **93**(11):5423–5435.
4. De Los Campos G, Gianola D, Allison DB: **Predicting genetic predisposition in humans: the promise of whole-genome markers.** *Nat Rev Genet* 2010, **11**(12):880–886.
5. Vazquez AI, De Los Campos G, Klimentidis YC, Rosa GJ, Gianola D, Yi N, Allison DB: **A comprehensive genetic approach for improving prediction of skin cancer risk in humans.** *Genetics* 2012, **192**(4):1493–1502.
6. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**(4):1819–1829.
7. Gianola D: **Priors in whole-genome regression: the Bayesian alphabet returns.** *Genetics* 2013, **194**(3):573–596.
8. Mulder HA, Calus MPL, Druet T, Schrooten C: **Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle.** *J Dairy Sci* 2012, **95**(2):876–889.
9. Jimenez-Montero JA, Gianola D, Weigel K, Alenda R, Gonzalez-Recio O: **Assets of imputation to ultra-high density for productive and functional traits.** *J Dairy Sci* 2013, **96**(9):6047–6058.
10. Habier D, Fernando RL, Dekkers JC: **Genomic selection using low-density marker panels.** *Genetics* 2009, **182**(1):343–353.
11. Weigel KA, De Los Campos G, Gonzalez-Recio O, Naya H, Wu XL, Long N, Rosa GJM, Gianola D: **Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers.** *J Dairy Sci* 2009, **92**(10):5248–5257.
12. Dassonneville R, Brondum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, Lund MS, Ducrocq V, Su G: **Effect of imputing markers from a low-density**

chip on the reliability of genomic breeding values in Holstein populations. *J Dairy Sci* 2011, **94**(7):3679–3686.
13. Moser G, Khatkar MS, Hayes BJ, Raadsma HW: **Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers.** *Genet Sel Evol* 2010, **42**:37.
14. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**(2):210–223.
15. Calus MP, Veerkamp RF, Mulder HA: **Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework.** *J Anim Sci* 2011, **89**(7):2042–2049.
16. Sun CY, Wu XL, Weigel KA, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D: **An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle.** *Genet Res* 2012, **94**(3):133–150.
17. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA: **Genomic evaluations with many more genotypes.** *Genet Sel Evol* 2011, **43**:10.
18. Mackay TF: **The genetic architecture of quantitative traits: lessons from Drosophila.** *Curr Opin Genet Dev* 2004, **14**(3):253–257.
19. Gianola D, van Kaam JB: **Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits.** *Genetics* 2008, **178**(4):2289–2303.
20. de Los CG, Gianola D, Rosa GJ: **Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation.** *J Anim Sci* 2009, **87**(6):1883–1887.
21. De Los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J: **Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods.** *Genet Res* 2010, **92**(4):295–308.
22. Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, Gonzalez-Recio O: **Radial basis function regression methods for predicting quantitative traits using SNP markers.** *Genet Res* 2010, **92**(3):209–225.
23. Gonzalez-Camacho JM, de Los CG, Perez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J: **Genome-enabled prediction of genetic values using radial basis function neural networks.** *Theor Appl Genet* 2012, **125**(4):759–771.
24. Gianola D, Okut H, Weigel KA, Rosa GJ: **Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat.** *BMC Genet* 2011, **12**:87.
25. Okut H, Gianola D, Rosa GJ, Weigel KA: **Prediction of body mass index in mice using dense molecular markers and a regularized neural network.** *Genet Res* 2011, **93**(3):189–201.
26. Heslot N, Yang HP, Sorrells ME, Jannink JL: **Genomic selection in plant breeding: a comparison of models.** *Crop Sci* 2012, **52**(1):146–160.
27. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: **Predicting quantitative traits with regression models for dense molecular markers and pedigree.** *Genetics* 2009, **182**(1):375–385.
28. Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, Crossa J, Manes Y, Dreisigacker S: **Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat.** *G3* 2012, **2**(12):1595–1605.
29. Howard R, Carriquiry AL, Beavis WD: **Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures.** *G3-Genes Genomes Genetics* 2014, **4**(6):1027–1046.
30. Legarra A, Robert-Granie C, Manfredi E, Elsen JM: **Performance of genomic selection in mice.** *Genetics* 2008, **180**(1):611–618.
31. Berry DP, Kearney JF: **Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection.** *Animal* 2011, **5**(8):1162–1169.
32. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME: **Imputation of missing genotypes from sparse to high density using long-range phasing.** *Genetics* 2011, **189**(1):317–327.
33. Vazquez AI, Rosa GJ, Weigel KA, De Los Campos G, Gianola D, Allison DB: **Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins.** *J Dairy Sci* 2010, **93**(12):5942–5949.
34. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J: **Genome-wide genetic association of complex traits in heterogeneous stock mice.** *Nat Genet* 2006, **38**(8):879–887.
35. Mott R: **Finding the molecular basis of complex genetic variation in humans and mice.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1467):393–401.
36. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JN, Mott R, Flint J: **Genetic and environmental effects on complex traits in mice.** *Genetics* 2006, **174**(2):959–984.
37. Usai MG, Goddard ME, Hayes BJ: **LASSO with cross-validation for genomic selection.** *Genet Res* 2009, **91**(6):427–436.
38. Browning BL, Browning SR: **A fast, powerful method for detecting identity by descent.** *Am J Hum Genet* 2011, **88**(2):173–182.
39. Tibshirani R: **Regression shrinkage and selection via the Lasso.** *J R Stat Soc Ser B-Methodological* 1996, **58**(1):267–288.
40. Park T, Casella G: **The Bayesian Lasso.** *J Am Stat Assoc* 2008, **103**(482):681–686.
41. Rosa GJM, Padovani CR, Gianola D: **Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation.** *Biom J* 2003, **45**(5):573–590.
42. Perez P, de Los CG, Crossa J, Gianola D: **Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R.** *Plant Genome* 2010, **3**(2):106–116.
43. Aronszajn N: *Introduction to the theory of Hilbert spaces.* Stillwater, Okla: Reasearch sic Foundation; 1950.
44. Wahba G: **Society for Industrial and Applied Mathematics.: Spline models for observational data.** In *CBMS-NSF Regional Conference series in applied mathematics 59.* Philadelphia, Pa: Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104); 1990. 1 electronic text (xii, 169 p.).
45. Gianola D, Fernando RL, Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures.** *Genetics* 2006, **173**(3):1761–1776.
46. Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ: **Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers.** *Genetics* 2010, **186**(2):713–724.
47. Bishop CM: *Pattern recognition and machine learning.* New York: Springer; 2006.
48. Mackay DJC: **Bayesian Interpolation.** *Neural Comput* 1992, **4**(3):415–447.
49. MacKay DJC: *Information theory, inference, and learning algorithms.* Cambridge, UK; New York: Cambridge University Press; 2003.
50. Demuth HB, Beale MH, MathWorks Inc: *Neural network toolbox for use with MATLAB : user's guide.* Natick, Mass: MathWorks; 2001.