Genomic-assisted prediction of breeding values: Indications of its effectiveness

DANIEL GIANOLA[1,2]
Department of Animal Sciences, University of Wisconsin, Madison, USA[1]
Institutt for husdyr- og akvakulturvitenskap, UMB[2]

Introduction

A massive quantity of genomic information is now available in livestock. For example, 2.8 million single-nucleotide polymorphisms (SNPs) have been found in the chicken genome. It is natural to consider use of this information as an aid in genetic improvement of livestock or plants. In medicine and agriculture, for example, genomic information could also be used for designing, e.g., diet or plant fertilization regimes that are genotype-specific. The concept is that the distribution of some SNP alleles reflects signals from genes, called quantitative trait loci (QTL), provided that marker and QTL alleles are associated (this is termed linkage disequilibrium, LD). We discuss how the molecular information, such as that conveyed by SNPs, has been employed for marker-assisted prediction of genetic value for quantitative traits in the sense of Meuwissen et al. (2001), Gianola et al. (2003) and Xu (2003). The focus is on inference of genetic value and prediction of phenotypes. The statistical methodology is described briefly, followed by some of our recent applications to chickens.

Statistical aspects

The statistical methodology employed most often consists of fitting a linear regression model to single traits (e.g., milk yield) assuming additive inheritance. There are many more SNP effects than data points, so standard regression does not work. Instead, a Bayesian model is used where SNP effects are random, and assumed to be drawn either from the same distribution with a common variance parameter (standard Bayesian regression), or with a marker-specific variance ("Bayes A" of Meuwissen et al., 2001), or from a mixture of distributions. This mixture has been placed at the level of the marker-variances, allowing for a 0 state with some probability, or at a state given by the value of an unknown, non-zero variance with the complementary probability ("Bayes B"; Meuwissen et al., 2001). An alternative is to use a mixture at the level of the effects. A different class of method is non-parametric regression (Gianola et al., 2006; Gianola and van Kaam, 2008). A non-parametric method may be better if inheritance is non-additive.

Estimates of marker effects are used to predict phenotypic values in the current sample of data (giving estimates of genomic assisted expected performance) or in a future sample which, in animal breeding is typically represented by the progeny generation. Often, the models estimate the SNP effects well, but are less powerful in out-of sample prediction. That is why cross-validation is necessary. Out-of – sample prediction is more difficult, because different LD relationships in the testing and predictive samples may exist (sampling bias), or because models with many parameters reflect vagaries of the data. It is important to find models that are robust under cross-validation.

Mortality in chickens

Gonzalez-Recio et al. (2008a) studied SNP-assisted prediction in broilers. Data were mortality records from birds between 14 and 42 days of age. This trait was scored as 0-1 (alive/dead), and recorded under conditions resembling the environment in commercial farms. The data included 12,167 records on progeny of 200 genotyped sires. Individual bird records were adjusted for environmental and mate effects. The sire-specific means of adjusted records were transformed to logs. Pedigree was tracked six generations back, ending up with 1103 sires in the pedigree file. Sires were genotyped for 5,523 SNPs. Twenty four SNPs, selected with a filter-wrapper algorithm, were used. The filter reduces the SNPs to a smaller number (e.g., 50), by using an information gain measure. In the wrapper step, a naïve Bayesian classifier (using cross-validation prediction accuracy) evaluates each SNP subset's usefulness, eventually arriving at the 24 SNPs with the highest performance. Statistical methods including genomic information (parametric and non-parametric) plus a standard genetic evaluation procedure ignoring markers (E-BLUP) were implemented to analyze sires' adjusted progeny means.

Table 1. Pearson correlations between predicted and actual values of the progeny average of each sire for late mortality in each subset (20% sires predicted in each subset) and by method

Subset	E-BLUP	F_{∞}-metric	Kernel	RKHS	BR
First	0.03	0.26	0.05	*0.27*[a]	0.13
Second	0.18	0.25	0.28	*0.37*	0.12
Third	*0.18*	−0.13	0.06	−0.01	0.17
Fourth	−0.04	0.12	0.13	*0.28*	0.15
Fifth	0.17	0.06	0.23	0.15	*0.25*
Global	0.10	0.08	0.14	*0.20*	0.16

E-BLUP, Bayesian linear model without genomic information; F_{∞}-metric, linear regression on SNPs based in the F_{∞}-metric model; kernel, nonparametric kernel regression with SNPs within sire treated as genomic combinations; RKHS, reproducing kernel Hilbert spaces resgression with SNPs within sire treated as a genomic combination; BR, Bayesian regression on 1000 SNPs based on Xu (2003).

[a]*Higher values indicate more accurate predictions. The highest correlation for each set is in italics.*

As shown in Table 1, results were variable. Globally, the two non-parametric methods (kernel regression and reproducing kernel Hilbert spaces regression, RKHS) had the best out-of-sample predictive performance in cross-validation with 5 randomly created folds. Bayes A (BR) was third, followed by E-BLUP, the currently used methodology for genetic evaluation.

Feed conversion ratio (FCR) in broilers

A one-fold cross-validation with a training and a testing set was carried out (Gonzalez-Recio et al., 2008b, unpublished). The testing set included only sons of sires that were in the training set. Several methods were used to predict phenotypes of animals in the testing set, i.e., first-generation performance. The methods were a standard genetic evaluation, and two methods including all available SNPs (after editing) as predictors in the model. The two methods including genomic information were Bayes A and a RKHS regression. The RKHS regression was also fitted with 400 pre-selected SNP using information gain. Data were average FCR records on progeny of each of 394 sires from a commercial broiler line from Aviagen Ltd. Individual bird FCR records were adjusted for environmental and mate effects. Two data sets (training and testing) were built up. Sires in the training set had more than 20 progenies with FCR records, to have a reliable mean phenotype. Sires in the testing set needed to have sires in the training set with progeny records. Sires in the training and testing sets had an average of 33 (±38) and 44 (±30) progeny, respectively. Sixty-one sires were included in the testing set, and the remaining 333 sires were in the training set. Predictions were calculated from the training set, and accuracy of predicting mean phenotype value of progenies of sons of sires was assessed in the testing set. Genotypes consisted of 4505 SNPs distributed along the genome. SNPs with mono-morphic genotypes or with frequencies lower than 5 % were excluded. Genotypes were built from 3481 out of the 4505 initial SNPs. Pre-selection of SNPs to be included in the analyses was performed using the information gain criterion.

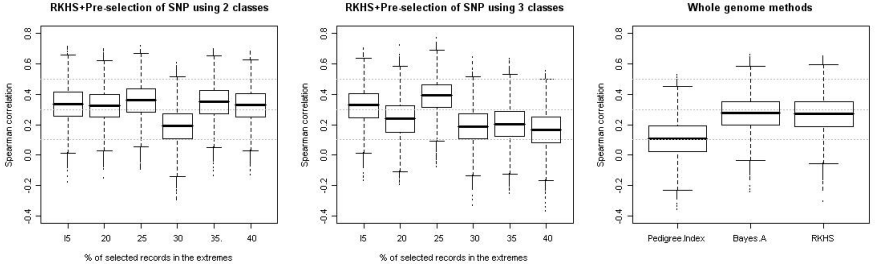

Figure 1. Box plots for the bootstrap distribution of correlations between predicted and observed phenotype in the testing set (progeny) obtained with: RKHS on 400 pre-selected SNPs using 2 or 3 classes to classify sires with different percentiles (left and middle panels, respectively) and methods using pedigree or all available SNPs (right panel).

Results are Figure 1. Bootstrap distributions (form of re-sampling aimed to assess variability) indicated that all 14 models using SNPs outperformed E-BLUP. Cross-validation correlations were generally larger than for mortality, probably because FCR has much less environmental noise.

Concluding remarks

Our results with chickens, USDA results with dairy cattle and French studies with mice indicate that genomic-assisted evaluation can outperform BLUP methodology, which is used in most countries. Much remains to be learned, and considerable research is needed. Some lessons include: 1) Prediction must be treated differently from inference. 2) Simple additive models on SNPs may do well. 3) More time should be spent in cross-validation and less in simulation. 4) Markers have ascertainment problems; simulations ignoring this may give a distorted picture. 5) Genetic complexity cannot be dealt with parametric methods; non-parametric methods are robust. 6) SNP-assisted genetic evaluation works, and seems to outperform BLUP in most cases.

References

Gianola, D., M. Perez-Enciso and M. A. Toro. 2003. On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163: 347-365.

Gianola, D., R. L. Fernando and A. Stella. 2006. Genomic assisted prediction of genetic value with semi-parametric procedures. Genetics 173: 1761-1776.

Gianola, D. and J. B. C. H. M. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289-2303.

Gonzalez-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa and S. Avendano. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178: 2305-2313

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard. 2001. Is it possible to predict the total genetic merit under a very dense marker map? Genetics 157: 1819-1829.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163: 789-801.