Effects of Selective Genotyping on Genomic Prediction

A. Ehsani[*†], L. Janss[*], and O. F. Christensen[*]

Introduction

Genomic prediction of breeding values recently became practical due to the sequencing of the bovine genome and the development of SNP arrays for genotyping (Meuwissen et al. (2001); Meuwissen and Goddard (2004); Villumsen (2009)). Among the proposed models and algorithms to predict breeding values based on dense markers, BLUP, Bayes A and Bayes B have been used to analyze simulated data (Meuwissen et al. (2001); Meuwissen and Goddard (2004); Xu 2003; Gianola et al. (2006)).

Genomic selection increases the rate of genetic improvement and reduces cost of progeny testing by allowing breeders to preselect animals that inherited chromosome segments of greater merit (Meuwissen et al. (2001); Schaeffer (2006); VanRaden (2008)). In a traditional progeny-testing scheme in dairy cattle, a random sample of bulls (certainly within family) is progeny tested. In a new genomic "juvenile breeding" scheme where only the better sons of bulls are getting progeny test data, this is no longer true. Pre-selection of data biases parameter estimates (Hung and Lin (2007); Weller et al**. (**2005); Lander and Botstein (1989); Jin et al**. (**2004)) and will not give good estimates of genotype effects, unless explicitly included in an analysis. Selective genotyping refers to genotyping a selected subset of individuals. Selective genotyping can increase the power of detecting QTLs (Lebowitz et al. (1987); Lander and Botstein (1989); Carey and Williamson (1991); Darvasi and Soller (1992)).

In this study the effects of selective genotyping of the reference population on genomic estimated breeding values (GEBVs) has investigated. Different scenarios were compared by computing reliabilities of GEBVs on a test group of animals. To obtain a robust comparison each scenario was repeated 10 times and the mean value of the repetitions is reported.

Materials and Methods

Two populations, one of size 400 and the other with 1750 individuals, were randomly mated for 1000 and 4000 generations, respectively. The next generation produced 8 offspring per individual, creating groups with 3200 and 14000 individuals, respectively. From each group 500 animals were chosen to form validation groups. Four selection processes (High, Medium, Low, and Random) and two levels of selection intensity(50% and 10%) were applied to

[*] Aarhus University, Faculty of Agricultural Sciences, Dept. of Genetics and Biotechnology, Bluchers Alle 20, P.O. Box 50, DK-8830 Tjele, Denmark.

[†] Tarbiat Modares University, Faculty of Agricultural Science, Dept. of Animal Science, P.O.Box: 14115-111, Tehran, Iran.

choose the 500 animals, and to determine the effects of selection on prediction of genomic breeding values. Genomes consisted of 2000 biallelic markers and 100 QTL (also biallelic). The heritability levels (0.1, 0.25, and 0.4) were examined and a recombination rate of 0.02 over the genome. We deleted the loci with minor allele frequency less than 5%. Phenotypes for the training data and genotype for both training and test data were consider to be known.

Statistical Model

A Bayesian approach that assumes a prior distribution of scaling factors for QTL was used. In the analysis all markers were fitted simultaneously as random effects in a Bayesian variable selection model that allows heterogeneous variances for different markers. A Bayesian method which captures the features of BayesA and BayesB but simplifies the computing algorithm (George and McCulloch (1993)) was used to estimate marker effects for genomic prediction. The following model was used:

$$y = 1\mu + \sum_{i=1}^{m} \mathrm{X}_i q_i \nu_i + e,$$

where y is the vector of phenotypic values, μ is the intercept, X_i is a design matrix of the number of alleles with positive effects, m is the number of SNP markers, $\boldsymbol{q}_i$ is the vector of scaled SNP effects of marker *i* with $\boldsymbol{q}_i \sim N(0, 1)$, $\nu_i > 0$ is a scaling factor for the SNP effect i, and e is the vector of residuals with $e \sim N(0, I\sigma_e^2)$. The scaling factor ν_i is assumed to have either a common prior distribution or a mixture prior distribution. Initial analysis showed that using a common prior distribution had better prediction ability than using mixture prior distribution, and therefore we applied a common prior model for prediction. The model was used in the training data to analyze phenotypes. To check the predictions, estimated genomic breeding values were correlated to true breeding values as known from the simulation. The squared correlation indicates the reliability of the estimated genomic breeding values.

Results and discussion

The results showed that for 2 different selection proportions of training data; 50% and 10%, the random selection scenario (*SR*) was the best scenario for prediction of GEBVs in the test data. The worst scenario was selecting individuals with moderate phenotypic values (*SM*). There was no big difference between selection of individuals with high phenotypic values (*SH*) and individuals with low phenotypic values (*SL*).

Table 1 shows that the expected reliabilities for the 50% selection proportion for the trait with heritability of 0.1 were 0.04, 0.1, 0.12 and 0.29, for the trait with $h^2 = 0.25$ were 0.14, 0.29, 0.24, and 0.57, and for trait with $h^2 = 0.4$ were 0.22, 0.44, 0.43, and 0.72 for scenarios *SM*, *SL*, *SH*, and *SR,* respectively.

Table 1: Expected reliability for 50% selection proportion

	h^2= 0.1	h^2= 0.25	h^2= 0.4
SR	0.29±0.002	0.57±0.001	0.72±0.0003
SH	0.12± 0.003	0.24±0.001	0.43±0.004
SL	0.1±0.004	0.29±0.006	0.44±0.002
SM	0.04±0.001	0.14±0.005	0.22± 0.003

Table 2 shows comparison of two scenarios, *SH* and *SR* for the selection proportion of 10%. The expected reliabilities for h^2= 0.1 were 0.02 and 0.15, for h^2= 0.25 were 0.07 and 0.46 and for h^2= 0.4 were 0.09 and 0.65 for scenarios of *SH* and *SR*, respectively.

Table 2: Expected reliability of 10% selection proportion

	h^2= 0.1	h^2= 0.25	h^2= 0.4
SR	0.15±0.005	0.46±0.001	0.65±0.001
SH	0.02±0.002	0.07±0.003	0.09±0.004

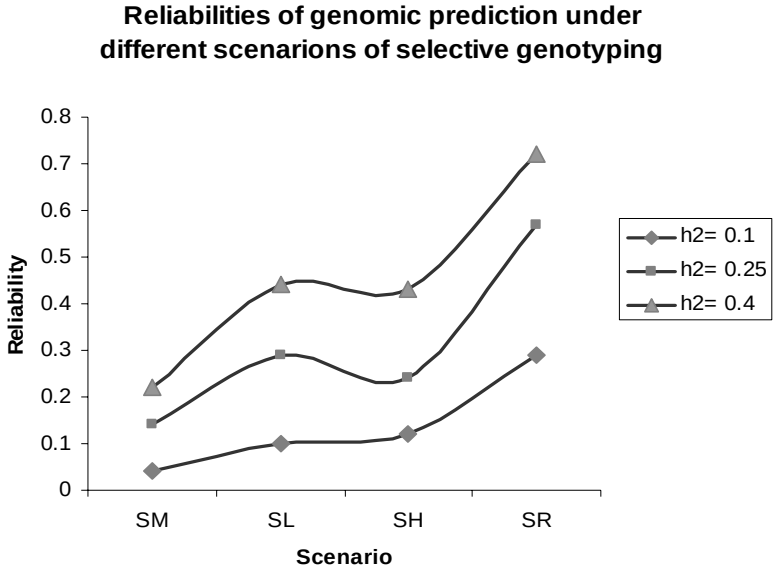

Figure 1: Expected reliability of 50% selection proportion with four different scenarios of selective genotyping: selection of individuals with moderate phenotypic values (*SM*), low phenotypic values (*SL*), high phenotypic values (*SH*) and random selection (*SR*).

Conclusion

As it is clear in figure 1 this study showed that selecting the best individuals (*SH*), doesn't provide good predictions compared with random selection (*SR*). Selective genotyping of individuals for genomic prediction should be based on algorithms that improve the reliability of prediction with emphasis on random selection.

References

Carey, G. and Williamson, J. (1991). *Am. J. Hum. Genet.,* 49:786-796.

Darvasi, A. and Soller, M. (1992). *Theor. Appl. Genet.,* 85:353-359.

George, E. I. and McCulloch, R. E. (1993*). J. Am. Stat. Asso.,* 88:881-889.

Gianola, D., Fernando, R. L., and Stella, A. (2006). *Genetics,* 173:1761-1776.

Huang, B. E. and Lin, D. Y. (2007). *Am. J. Hum. Genet.,* 80:567-576.

Jin, C., Lan, H., Attie, A. D. *et al*. (2004). *Genetics,* 168:2285-2293

Lander, E. S., and Botstein, D. (1989). *Genetics,* 121:185-199.

Lebowitz, R. J., Soller, M., and Beckmann, J. S. (1987). *Theor. Appl. Genet.,* 73:556-562.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). *Genetics,* 157:1819-1829.

Meuwissen, T. H. E., and Goddard, M. E. (2004). *Genet. Sel. Evol.,* 36:261-279.

Schaeffer, L. R. (2006). *J. Anim. Breed. Genet.,* 123:218-223.

VanRaden, P. M. (2008). *J. Dairy Sci.,* 91:4414-4423.

Villumsen, T.M., Janss, L., and Lund, M. S. (2009). *J. Anim. Breed. Genet.,* 126:3-13.

Weller, J. I., Shlezinger, M., and Ron, M. (2005). *Genet. Sel. Evol.,* 37:501-522.

Xu, S. (2003). *Genetics,* 163:789-801.