Genomic Selection in the Nordic Countries

M. S. Lund[1] and G. Su[1]
[1]Aarhus University, Faculty of Agricultural Sciences, Genetics and Biotechnology, Foulum, Denmark.

1. Introduction

Genomic selection, which use dense markers in linkage disequilibrium with QTL alleles affecting the traits (Meuwissen and Goddard 2001), is now possible due to high throughput genotyping for thousands of SNP markers through the whole genome. Therefore, genetic evaluation can be performed as soon as DNA can be obtained, which allows selection in both genders early in life.

Several statistical models and algorithms have been proposed to predict breeding values based on dense markers (e.g. Meuwissen and Goddard 2001; Xu 2003; Gianola 2006). Moreover, several simulation studies have shown that Bayesian models with prior mixture distributions for variances of chromosome segment effects have better prediction ability (Meuwissen 2001, Lund *et al*).

So far investigations on genomic selection in the literature are generally limited to simulation studies, and are rarely based on real data. The objectives of this study were to assess the predictive ability of the models with different shrinkage intensities of chromosome segment effects, and to investigate the reliability of genomic estimated breeding values for all index traits in the Nordic total merit index, based on the data from the Danish and Swedish Holstein populations.

2. Material and methods

2.1 Reference population and phenotypic data

2012 Holstein bulls from 125 half-sib families born during years from 1986 to 2002 were chosen to be genotyped. From all halfsib families with more than 5 sons with official breeding values the sires and sons were genotyped. A maximum of 25 sons were genotyped within one halfsib family. If more were available, 25 were selected randomly.

The bulls were genotyped using Bovine 50K Illumina™ iSelect SNP chip. The marker data were edited using the following criteria: 1) the locus was deleted if the minor allele frequency less then 5%, or the proportion of typed animals at this locus was less than 95%, or typing accuracy was less than 60%; 2) the individual was deleted if the proportion of typed loci was less than a score of 0.65. After the editing, There were 2012 bulls and 38055 SNP markers available.

Conventional pedigree based EBV were used as response variable to estimate SNP effects. The EBV and their reliability for the genotyped bulls were obtained from official evaluations in 2008. In total 17 index traits were analyzed in this study.

2.2 Bayesian stochastic search variable selection model

In this study, a single SNP marker were used as predictors. Conventional EBV were used as response variables to estimate SNP effects. A linear model applying Bayesian Gibbs sampling algorithm (Janss, 2008) was used to estimate SNP effects,

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{m} \mathbf{X}_i \mathbf{q}_i v_i + \mathbf{e}$$

where $\mathbf{y}$ is pedigree based EBV, μ is the intercept, m is the number of SNP loci, $\mathbf{q}_i$ is the vector of scaled effects (scaled by standard deviation) of SNP at locus i with $\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I})$, v_i ($v_i > 0$) is the scale factor (standard deviation) for SNP effect at locus i, and $\mathbf{e}$ is the vector of residual with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$. Accordingly, the effects of SNP at locus i is the product of v_i and $\mathbf{q}_i$.

Scale factor v_i were assumed to have either a common prior distribution or mixture prior distributions. A common prior distribution, which slightly differentiated between small and large effects of SNP, was assumed as a truncated normal distribution,

$$v_i \sim TN(0, \sigma^2_v), \quad v_i > 0$$

Mixture prior distributions, which intensively differentiate between small and larger effect of SNP, assumes that a large proportion (π_0) of SNP have small effect, and small proportion (π_1) of SNP have large effect. This is achieved by assuming that prior distribution of v_i is either a truncated normal distribution with a small variance (σ^2_{v0}) or a truncated normal distribution with large variance (σ^2_{v1}),

$$v_i \sim \pi_0\, TN(0, \sigma^2_{v0}) + \pi_1\, TN(0, \sigma^2_{v1})$$

The prior distribution of μ, σ^2_v and σ^2_{v1} were assumed to be improper uniform distributions, while σ^2_{v0} was fixed at a small value. In this study, σ^2_{v0} was set to 0.01 for all traits.

GEBV for individual k is defined as the sum of predicted effects of SNP over all loci,

$$GEBV_k = \hat{\mu} + \sum_{i=1}^{m} \mathbf{x}_{i(k.)} \mathbf{q}_i v_i$$

The effect of shrinkage intensity on accuracy of GEBV was investigated using five scenarios: 1) mixture prior of scale factors with π_1=5%, 2) π_1=10%, 3) π_1=20%, 4) π_1=50%, and 5) common prior of scale factors for all loci.

2. 3 Cross validation

The models with different priors for scale factors and the accuracy of GEBV were evaluated using a five-fold cross validation. In the cross validation, five subsets were created from the whole data, each left two year's records out (subset 1 without bulls born in 1993 and 1994, subset 2 without 1995 and 1996, and so on). Each of the five subsets was used as training data to estimated SNP effect, and the corresponding "leave out" data as test data to predict GEBV. To relax the dependency of test data on training data, sires of bulls were excluded from the test data. The five test datasets comprised a total test data including 1548 bulls.

Five scenarios of prior distribution for scale factors (standard deviations, v_i) of SNP effects were evaluated by analysing four index traits (protein, yield index, udder health, and female fertility). Model predictive ability was assessed by R^2 between GEBV and conventional EBV in the five-fold cross validation. The best model (which was common prior distribution in this study) was used to analyse all the 17 index traits.

Reliability of GEBV was assessed by R^2 between GEBV and conventional EBV in the cross validation for those bulls which had a reliability 0.99 for yield EBV. These bulls were proven bulls with a large number of daughter records. Their EBV should be very close to true breeding values: Therefore the R^2 between GEBV and conventional EBV for those bulls could be a good measurement of reliability of GEBV.

The analyses were carried out using IBAY package (Janss, 2008). The Gibbs sampler was run as a single chain with a length of 50,000 samples. Convergence was monitored by graphical inspection. The first 20,000 samples were discarded as burn-in. Every 10th sample of the remaining 30,000 was saved to estimate the features of the realised posterior distribution.

3. Results

3.1 Reliabilities

For the bulls with reliability of yield index EBV equal to or larger than 99%, the EBV should be very close to true breeding values. Therefore, R^2 between GEBV and EBV for this group of bulls is used as a measurement of reliability of GEBV (Table 1). Accordingly, the lowest reliability of GEBV was found in body conformation (0.301). The reliabilities for the other 16 traits ranged from 0.409 for other disease to 0.731 for fat percentage. The average reliability of GEBV over all traits was 0.513. Although the heritabilities for these

traits differed considerably, the difference in reliabilities of GEBV between the traits were relatively small. In addition, the reliability of GEBV was not strongly associated with heritability. For example, the reliability of GEBV for fertility which had a low heritability was as high as those for production traits.

Table 1. Within-year R^2 between EBV and GEBV for bulls with reliability of EBV for yield index equal to or higher than 0.99.

Trait	R^2 between EBV and GEBV
Mastitis	**0.5**
Other diseases	**0.4**
Female fertility	**0.56**
Calving maternal	**0.43**
Calving direct	**0.5**
Longevity	**0.5**
Feet and legs	**0.46**
Body conformation	**0.30**
Mammary system	**0.45**
Milking speed	**0.48**
Temperament	**0.5**
Milk-index	**0.62**
Fat-index	**0.54**
Fat%	**0.73**
Protein-index	**0.63**
Protein%	**0.52**
Yield index	**0.55**

4. Discussion

Results provide an efficient base for genomic selection in Danish and Swedish Holstein Fresian populations. Reliability of GEBV was measured as R^2 between GEBV and EBV for the bulls with 0.99 reliability of yield index EBV, based on a cross validation. The reliabilities of EBV for these bulls were very high for all the traits (ranged from 87% to 99%), indicating that the EBV of those bulls could be very close to true breeding values. Therefore, R^2 between GEBV and EBV based on these bulls are expected to be a good measurement of the reliability of GEBV. Accordingly, the reliabilities were moderately high (ranged from 0.409 for health to 0.731 for fat percentage), except for body score which had a reliability of 0.301. The average reliability of GEBV over all the traits was 0.513. There are very limited reports on the reliability of GEBV based on real data.

Although the heritabilities for the 17 traits were quite different, the differences in reliability of GEBV between these traits were relatively small. Since SNP effects were estimated from EBV in this study, the influence of heritability on GEBV was through its influence on reliability of EBV. Given accurate estimates of SNP effect, GEBV would be independent of heritability. The less dependency on heritability indicates that genetic evaluation based on GEBV would be more beneficial for the traits with low heritability. As a consequence it becomes easier to obtain a balanced genetic progress between functional traits with low heritability and production traits. This is particularly important in the Nordic breeding schemes which include more functional traits than most other countries.

The reliabilities of GEBV in this study are moderately high. On average over 17 traits, the reliability is 0.513. The figure is about twice of the reliability of parent average. It indicates that genomic selection can greatly improve the accuracy of pre-selection for young bulls, compared with traditional selection based on parent average. According to the results from this study, VikingGenetics has started to use GEBV to pre-select young bulls for progeny test in Danish Holstein population.

In 2009 the Nordic HF reference population will increase from 2000 to about 4000 progeny tested bulls. This should result in more reliably estimated SNP effects and therefore higher reliabilities for GEBV. At the same time GEBV models will be developed for Jersey, FAY, SRB, and RDM. For the individual breeds the reference populations will be smaller, ranging from 1000 to 1600. The many genetic links between the three Nordic red breeds should allow for a common model for the red breeds. We expect, however, that this requires a more complex modeling of the genetic structures within and between the breeds.

5. Acknowledgement

We thank the Danish Cattle Federation, Swedish Dairy Association and Nordic Cattle Genetic Evaluation for providing phenotypic data. This work was performed in the project 'DNA based selection to improve disease resistance, fertility, calf survival and production in Danish dairy cattle' funded by the Danish Directorate for Food, Fisheries and Agri Business grant no. 3401-04-00853, Viking Genetics, Danish Cattle Association, and Aarhus University.

References

Gianola, D., Fernando, R. & Stella, A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics 173*, 1761–1776.

Janss, L. 2008. iBay manual version 1.44. http://www.lucjanss.com.

Lund, M.S., Sahana, G., de Koning, D.J., Su, G. & Carlborg, Ö. 2008. Comparison of analyses of the QTLMAS XII common dataset: I Genomic selection. *J. Anim. Breed. Genet.* (in press).

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819–1829.

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