Investigation Of The Reliability Of Genomic Selection Using Combined Reference Data Of The Nordic Red Populations

R.F. Brøndum[*], E. Rius-Vilarrasa[**], I. Strandén[***], G. Su[*], B. Guldbrandtsen[*], W.F. Fikse[**], M.S. Lund[*]

Introduction

Genomic Selection is becoming a popular tool in cattle breeding, where selection is based on breeding values predicted directly from dense sets of genetic markers. Each genetic marker from a single nucleotide polymorphism (SNP) panel is potentially in linkage disequilibrium (LD) with a quantitative trait locus. An effect for each allele at the individual marker loci is then estimated by fitting a model to phenotypic data from the reference animals which have both genomic and phenotypic records (or pseudo-observations). The genomic breeding value of a candidate with no phenotypic record is then predicted as the sum of all marker effects. This value is called the direct genomic estimated value (DGV)(Meuwissen et al. (2001)).

Previous studies have shown that the reliability of genomic selection is dependent on the number of animals used to determine the marker effects and the heritability of the trait (Goddard (2008)). This suggests that for a population having a small number of reference animals, there might be little benefit from genomic selection, compared with using the parent average for selection of young bulls. The objective of this study is to investigate the possibility of increasing the reliability of genomic selection in the Nordic Red cattle by pooling reference data from the Danish, Swedish and Finnish Red dairy cattle breeds.

Materials and methods

Data

The dataset included 3735 Danish, Swedish and Finnish Red dairy bulls from 306 half-sib families, born between 1986 and 2005. The bulls were genotyped using Illumina Bovine SNP50 Beadchip (Illumina, San Diego). The genotypic data was edited according the same quality criteria for minor allele frequencies, GC-scores and callrates, but the number of remaining SNPs after the editing varied between reference sets. Phenotypic data were conventional estimated breeding values (EBV) evaluated in 2009. Three separate national studies were conducted for the Danish, Swedish and Finnish bulls. In the Danish study, Swedish animals from Danish/Swedish half-sib families were included in the dataset if the number of Swedish bulls in the family was

* Aarhus University, Faculty of Agri. Sciences, Dep. of Genetics and Biotechnology, DK-8830 Tjele, Denmark
** Swedish University of Agri. Sci., Dep. of Animal Breeding and Genetics, PO Box 7023, 750 07 Uppsala, Sweden
*** MTT, Biometrical Genetics, 31600 Jokioinen, FINLAND

less than five times as large as the number of Danish bulls. A similar approach was applied in the Swedish study for including Danish animals. In the Finnish study, only Finnish animals were included.

To investigate the effect of combining reference data on the reliability of genomic selection, the three reference datasets from the three populations were pooled into one combined reference dataset to estimate marker effects. Marker effects were calculated for five different traits: Protein percentage, Udder health, Female fertility, Milk yield and Maternal Calving.

Statistical Model

In this study, single SNP markers were used as predictors and EBV as response variables, The EBVs were predicted in a joint Nordic model using the same data. Bayesian inference was used to estimate the SNP effects, using the model:

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

where **y** is the vector of conventional EBV, μ is the mean, m is the number of SNP markers, $\mathbf{X}_i$ is the design matrix allocating genotypes to the animals, $\mathbf{q}_i$ is the vector of marker effects, v is a scaling factor, and **e** is the vector of residuals. The prior distributions are assumed as:

$$\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I}) \qquad v_i \sim TN(0, \sigma_v^2) \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e)$$

A more extensive description of the model and priors can be found in Villumsen et al. (2008) and Su et. al (2010).

DGV for each individual is calculated as:

$$DGV_k = \hat{\mu} + \sum_{i=1}^{m} \mathbf{x}_{i,k} \mathbf{q}_i v_i$$

Evaluation of reliability

Reliability was assessed using a five-fold cross validation, where animals were divided into five approximately equally sized subsets according to year of birth . Half-sib families having animals in more than one subset were moved to the subset containing the larger part of the family. Cross-validation was done by successively removing one subset at a time from the whole dataset, and using the left-out subset as a test dataset. To relax dependency between reference data and test data, the sires which had sons in the reference data were removed from the test data. The reliability of DGV was estimated as the within-year squared correlation between EBV and DGV in the five test populations.

Marker effects were estimated using iBay v 1.46 (Janss (2009)). The Gibbs sampler was run as a single chain with 50000 iterations. Samples from the first 10000 iterations were discarded as burn-in, and every 5th sample of the remaining 40000 iterations was saved to estimate parameters in the posterior distribution.

Results and discussion

Squared correlations between EBV and DGV of the five traits for the bulls in test data based on different reference datasets are given in Table 1. Using national reference data, the squared correlations ranged from 0.17 to 0.26 with an average of 0.22 in the Danish Red population, from 0.04 to 0.19 with an average of 0.13 in the Swedish Red population, and from 0.15 to 0.20 with an average of 0.18 in the Finnish Red population. Using the combined reference data, squared correlations were higher than those obtained from national reference data for all traits, except for fertility in the Danish population. Averaged over the five traits, the squared correlation increased by 1% in the Danish Red population, 8% in the Swedish Red population and 7% in the Finnish Red population.

The observed reliabilities in this study are lower than what is theoretically expected for genomic selection and lower than most previously reported from Holstein data . Meuwissen et al. (2001) reported a reliability of 0.72 when using a Bayesian model on simulated data, and (Hayes et al. (2009b); Su et al (2010)) report reliabilities between 0.20 and 0.70 for Holstein data. Earlier results are however obtained from populations with a more homogenous genome compared to the Nordic red breeds. Many sires from different breeds have been introduced and this gives a relatively higher heterozygosity. This could cause different LD patterns in the genome, thus making it harder to estimate precise marker effects.

Table 1: Within year reliabilities of DGV in the nordic red breeds.

Reference	DK	SWE	FIN	Combined			
Test	DK	SWE	FIN	DK	SWE	FIN	Combined
No. of animals	929	1551	1562	778	1395	1562	3735
Protein	0.17	0.11	0.15	0.18	0.21	0.24	0.22
Udder health	0.21	0.19	0.19	0.23	0.25	0.29	0.28
Fertility	0.24	0.18	0.19	0.21	0.27	0.24	0.26
Milk	0.23	0.14	0.20	0.25	0.23	0.29	0.27
Mat. Calving	0.26	0.04	0.16	0.29	0.09	0.21	0.26
Mean	**0.22**	**0.13**	**0.18**	**0.23**	**0.21**	**0.25**	**0.26**

The squared correlations between DGV and EBV when using combined reference data are consistently higher than those obtained using the reference data from a single population, and in line with results from previous studies. VanRaden et al. (2009) showed that the reliability of genomic prediction increased with increasing size of reference data in a Holstein population. In a simulation study Su et al. (2009) reported a considerable improvement of genomic prediction by combining reference data from populations having a common origin, and Hayes et al. (2009a) showed an increase in the accuracy of genomic selection when combining reference data from Australian Holstein and Jersey populations.

Conclusion

The aim of this study was to investigate the effects of including information from different but related populations on the reliability of genomic selection. The results indicate that the reliability of genomic prediction for Red cattle is improved by combining reference data from different Red populations. Application of this method will help breeding companies make more accurate decisions when selecting bulls, and thus increase the genetic gain. Further studies are however needed to investigate the lower than expected benefits of genomic selection.

Acknowledgements

This work was performed in the project "Genomic Selection – from function to efficient utilization in cattle breeding (grant no. 3412-08-02253)", founded by Danish Directorate for Food Fisheries and Agri Business, VikingGenetics, Nordic Genetic Evaluation and Aarhus University.

References

Hayes, B.J., Bowman, P.J., Eggen, A. *et al*. (2009a). *Genet. Sel. Evol*., 41: Art. No. 51.

Hayes, B.J., Bowman, P.J., Chamberlain A.J. *et al.* (2009b). *J. Dairy. Sci*., 92: 433-443.

Su., G., Guldbrandtsen, B., Gregersen, V.R. *et al.* (2010). *J. Dairy. Sci*., 93: 1175-1183.

Su., G., Guo., G., and M.S. Lund (2009). *Abstracts of the.60th EAAP*: 296.

Janss, L. (2009). iBay [1.46]. www.lucjanss.com

Goddard, M.E. (2008). *Genetica* 136: 245-257.

VanRaden, P.M., Van Tassel, C.P., Wiggans, G.R. *et al.* (2009). *J. Dairy. Sci*., 92: 16-24.

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

Villumsen, T.M., Janss, L. and Lund, M.S. (2008). *J. Anim. Breed. Genet.*, 126:3-13.