Investigation of reliability of genomic predictions in the admixed Nordic Red dairy cattle

M. L. Makgahlela*[1, 2], E. A. Mäntysaari[2], I. Strandén[2], M. Koivula[2], U.S. Nielsen[3], M. J. Sillanpää[1, 4, 5] and J. Juga[1]

[1]Department of Agricultural Sciences, P. O. Box 27 FIN-00014 University of Helsinki, Finland; [2]MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics, 31600 Jokioinen, Finland; [3]Danish Agricultural Advisory Service, Udkaersvej 15, 8200 Aarhus, Denmark; [4]Department of Mathematics and Statistics, P.O. Box 68 FIN-00014 University of Helsinki, Finland; [5]Department of Mathematical Science and Department of Biology, P.O. Box 3000 FIN-90014 University of Oulu, Finland

mahlako.makgahlela@helsinki.fi

Abstract

The success of genomic selection (GS) in small breeds which are likely to have admixed structures has been minimal. This is because accuracy of GS depends on the extent of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) and LD depends on the genetic structure of the population and marker density. In the current study, we evaluate reliability of genomic predictions in young unproven bulls, when interactions between marker effects and breed of origin are accounted for in the Nordic Red dairy cattle (RDC). The population structure of the RDC is admixed. Data consisted of animal breed proportions calculated from the full pedigree, deregressed proofs (DRP) of published estimated breeding values (EBV) for yield traits and genotypic data for 37,595 SNP markers. Direct genomic breeding values (DGV) were estimated using 2 models, one accounting for breed-specific effects and other assuming uniform population. Validation reliabilities were calculated as the squared correlation between DRP and DGV ($r^2_{DRP, DGV}$), corrected by the mean reliability of DRP. Using the breed-specific model increased the reliability of DGV by 2% and 3% for milk and protein, respectively, when compared to homogeneous population GBLUP model. The exception was for fat, where there was no gain in reliability. Estimated validation reliabilities were low for milk (0.32) and protein (0.32) and slightly higher (0.42) for fat.

Keywords: reliability, genomic breeding values, admixed breeds, breed proportions

Introduction

Genomic selection has been effectively applied in the prediction of performances in most dairy cattle populations, but the success has not been realized in small breeds which are likely to be admixed. The success of genomic selection depends primarily on the extent of LD between markers and QTL, the number of markers and phenotypic records used in the reference population and the heritability of a trait (Goddard, 2009). Unfortunately, population sizes and structures remain a major limiting factor for increased accuracies in small breeds. Studies have shown that the accuracy of GS in young bulls from small populations could be increased by combining multiple populations in to one reference population (Hayes *et al*., 2009; Brøndum *et al*., 2011). However, when analyzing data on multiple populations, predictions across breeds generally ignore structure and assume that these populations are uniform. This approach may hamper accurate estimation of marker effects across breeds and result in low accuracy of DGV because different breeds may exhibit different QTLs (Toosi *et al*., 2009), allele frequencies vary between populations and also, the extent of LD may not be consistent across breeds (Ewens and Spielman, 2005). Especially for populations which are more diverged.

The population structure of the Nordic Red dairy cattle (RDC) is an admixture of the Danish Red, Swedish Red and Finnish Ayrshire cattle. Furthermore, the gene pools of each of these 3 populations contain fractions from other breeds. Although the population is admixed, current predictions ignore structure and assume a genetically homogeneous population (Brøndum *et al*., 2011; Su *et al*., 2011). If interactions between marker effects and breed of origin were to be included in the model, the accuracy of GS in this admixed population may be improved. Therefore, the objective of the current study was to estimate direct genomic estimated breeding values using a breed-specific model and compare its reliability with a model which assumes homogeneous population in the Nordic Red dairy cattle.

Methods

Phenotype (n = 6,253) and genotype (n = 6,145) data were available for the Nordic RDC bulls. Genomic information for bulls born between 1971 and 2006 were provided by the Nordic genomic selection project. These data was edited to remove uninformative loci and thereafter consisted of 37,995 single nucleotide polymorphic (SNP) markers for each bull. Published estimated breeding values (EBV) for milk, fat and protein indices were based on 2010 March NAV (Nordic Cattle Genetic Evaluation) routine evaluations. Deregressed proofs (DRP) for all genotyped bulls were calculated using effective daughter contribution (EDC) as a weight (Schaeffer, 2001). In order to estimate DRP for the bull, the reliability of the DRP was required to be at least 20%.

Breed proportions (BP) from ancestral breeds were from the full Nordic RDC pedigree (Lidauer *et al*., 2006). There were 13 known breeds in the population. The overall mean BP was calculated for each breed and mainly 3 breeds had higher mean BP, over 10%. Therefore, 4 breeds were defined as Swedish Red (SRB), Finnish Ayrshire (FAY), Norwegian Red (NRF), and the remaining breeds with BP less than 10% were put together into breed OTHER. After phenotypic and BP data were merged, there were 4,142 records in the data. These bulls were divided into the reference population of 3,330 and selection candidates with 812 bulls. The reference population included bulls born between the years 1971 and 2001 and selection candidates were bulls born from 1996 to 2005 and had not been evaluated during NAV 2005 routine evaluations.

Statistical Analyses

The DRP were used as response variables, BP as covariables for random regression and the analysis were weighted by the reliability of DRP which was defined as EDC. The genomic relationship matrix **G** for 4,142 bulls was calculated from the genotypic data using method 1 as described by VanRaden (2008). All diagonal elements of the **G** matrix were multiplied by 1.01 to correct any possible singularities.

Estimation of direct genomic values

The general structure of the mixed effect model in matrix notation can be represented as:

where is a *n* x 1 vector of DRP; is the general mean; **1** is a unit vector; $\mathbf{C}_i$ is an *n* x *n* diagonal matrix with BP for all bulls in breed *i* on the diagonal and $\mathbf{S}_i$ is square root of $\mathbf{C}_i$, here square roots of BP were used to equalize the proportion of genetic variance accounted for by breeds; is a 4 by 1 vector of fixed breed effects; is a vector of random breed specific animal genetic effects ordered by animals within breed, and **e** is an *n* x 1 vector of

random residual terms with common across breeds. For , with $\mathbf{G}_0$ being diagonal matrix with country genetic variances in the diagonal.

In order to compare the models, we also fitted a GBLUP model, which assumes homogeneous population:

,

where **Z** is n by n_g incidence matrix relating genotyped animals to DRP; **g** is a vector of additive genetic effects. It is assumed that ; **G** is the n_g x n_g genomic relationship matrix; is the genetic variance across breeds and **e** is the random residual with common across breeds.

Breed-wise variance components were estimated using the breed specific model for each trait at the time and obtained estimates for and were used for DGV prediction. Average genetic variance for each trait was obtained as the sum of the product of breed variances and the means of BP in the data. In the estimation of DGV, DRP for the validation bulls were removed from the phenotypic data. However, they received estimated DGV based on their genomic relationships with animals in the reference population. DGV values for all animals were obtained as the sum of the product of animals' base breed DGV and the BP in that breed.

Validation of the model and reliability of DGV

The validation approach of DGV generally followed Interbull GEBV test, with the exception that one dataset is used for both prediction and validation (Mäntysaari *et al*., 2010). The reliability and unbiasedness of DGV was assessed as the regression of DRP on DGV for selection candidates, weighted by ——, the reliability of DRP, where ——, with the heritability used in NAV evaluations (see Interbull 2008). The coefficient of determination r^2 was then scaled by a constant of the average of *w*, where the scaling factor was 0.94 for milk and protein and 0.92 for fat.

Results and discussion

Figure 1 illustrates trends in breed proportions for bulls registered in Finland. In this population, average BP between 1980 and 1996 were over 70% for FAY and about 20% for NRF. After 1996 FAY gene fractions has declined steadily to about 50% as SRB and CAY proportions have gradually increased. NRF proportion has been steady about 15% to 20% during the whole period.

Average genetic variances ranged from 87.75 for milk to 99.31 for protein but corresponding residual variance estimates were very high being over 2 times the expected (Table 1). The resulting variance ratios were high at 34.55, 30.89 and 28.79, respectively, for milk, protein and fat when compared to the original ratios obtained from conventional evaluations.

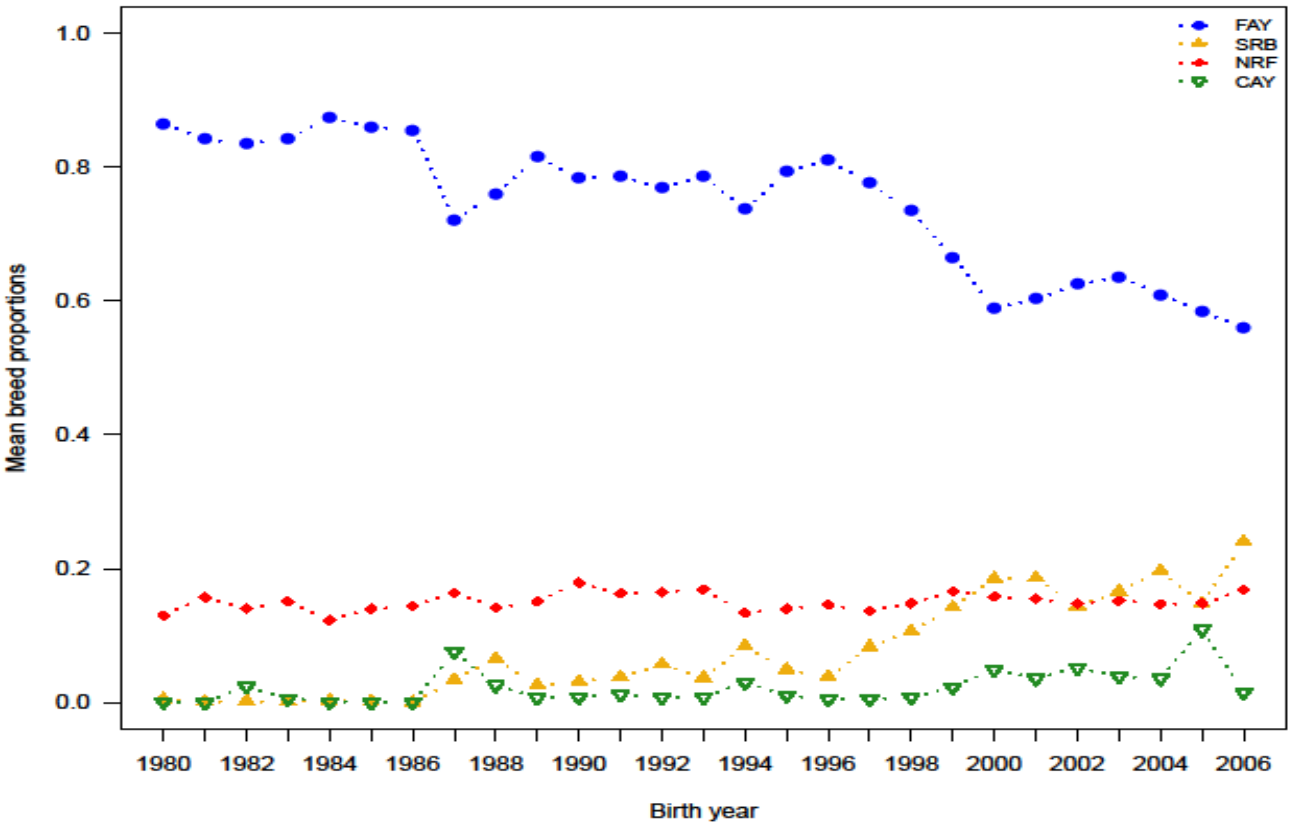
Table 1 The overall mean genetic ( ) and residual ( ) variances and variance ratios ( ) estimated from the genomic data by trait

Trait			
Milk	87.75	3031.68	34.55
Protein	99.31	3068.28	30.89
Fat	87.94	2529.70	28.79

Validation reliability of DGV

The objective of this study was to explore reliability of DGV estimated using a breed-specific model. Table 2 shows the reliability of DRP ( ), coefficient of regression (b_1), coefficient of determination (r^2) and validation reliability of DGV ( ). The validation reliabilities estimated from the breed-specific model were 2% and 3% higher for milk and protein,

Figure1 Average breed proportions by birth year for bulls registered in Finland



respectively, than those estimated using GBLUP model. The exception was however for fat, where there was no gain from using a breed-specific model. Validation reliabilities of DGV were low for milk and protein (0.32), and slightly higher for fat (0.42). The validation reliabilities for all traits were higher than those reported by Brøndum *et al*. (2011) using Bayesian model, similar to those observed by Su *et al*. (2011) with GBLUP in the same population but lower than reported elsewhere for other breeds (Hayes *et al*., 2009). In addition to differences in reference population sizes and marker densities, all the other studies used models that assume uniform population structure.

Current genomic predictions in admixed populations ignore information of true origin of genes and assume admixed populations are homogeneous. Our model assumed that an animal has a certain probability to carry marker effects from the base breeds and therefore, combines QTL from different breeds simultaneously in the prediction of DGV. The observed validation reliabilities suggest that the predictive ability of our model was comparable but not notably better than models that ignore breed-specific effects. Therefore, further investigation on analyses that utilize breed-specific marker associations need to be developed.

Table 2 The reliability of DRP (r_{DRP}), regression coefficients (b_1), coefficients of determination (r^2) and validation reliabilities (r^2_{DGV}) for both models in selection candidates

		GBLUP model			**Multi-breed Random regression**		
Trait		b_1	r^2		b_1	r^2	
Milk	0.94	0.78	0.28	0.30	0.79	0.30	0.32
Protein	0.94	0.82	0.28	0.29	0.81	0.31	0.32
Fat	0.92	0.94	0.39	0.43	0.94	0.39	0.42

References

Brøndum, R.F., Rius-Vilarrasa, E., Strandén, I., Su G., Guldbrandtsen, B., Fikse, W.F., Lund, M.S. (2011) Reliabilities of genomic predictions using combined reference data of the Nordic Red cattle populations. *J. Dairy Sci*., 94, 4700-4707.

Ewens, W.J., Spielman, R.S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet*., 57, 455-464.

Goddard, M. (2009) Genomic selection: prediction of accuracy and maximization of long term response. *Genetica*. 136, 245-257.

Hayes, B.J., Bowman P.J., Chamberlain, A.C. (2009) Accuracy of genomic breeding values in multi-breed dairy cattle population. *Genet. Sel. Evol*. 41, 51.

Interbull. (2008) National genetic evaluation system. Accessed 02 June 2010. http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm.

Lidauer M., E.A. Mäntysaari, I. Strandén, J. Pösö, J. Pedersen, U.S. Nielsen, K. Johansson, J.-Å. Eriksson, P. Madsen, G.P. Aamand. (2006) Random Heterosis and Recombination Loss Effects in a Multibreed Evaluation for Nordic Red Dairy Cattle. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, 13-18 August 2006, Belo Horisonte, Brasil.

Mäntysaari E.A., Liu, Z., VanRaden, P. (2010) Interbull validation test for genomic evaluations. *Interbull Bulletin*. 41, 17-22.

Schaeffer, L.R. 2001. Multiple trait international bull comparisons. *Liv. Prod. Sci*. 69, 145-153.

Su G., Madsen P., Nielsen U.S., Mäntysaari E.A., Aamand G.P., Christensen OF., Lund M.S. (2011) Genomic prediction for the Nordic Red cattle using one-step and selection index blending approaches.

Toosi A., Fernando R.L., Dekkers J.C.M. 2009. Genomic selection in admixed and crossbred populations. *J. Anim. Sci*. 88, 32-46.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci*., 91, 4414–4423.