Across Breed Multi-Trait Random Regression Genomic Predictions in the Nordic Red Dairy Cattle

M. L. Makgahlela*[12], E. A. Mäntysaari[2], I. Strandén[2], M. Koivula[2], M. J. Sillanpää[1]
U.S. Nielsen[3] and J. Juga[1]

[1]Department of Agricultural Sciences, P. O. Box 27 FI-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
**mahlako.makgahlela@helsinki.fi*

Abstract

Current genomic prediction equations, when carried out in multiple populations with admixed structures ignore structure and assume these populations are uniform. The observed reliabilities of direct genomic breeding values (DGV) for unproven bulls in these populations so far have been low. The current study evaluated reliabilities of DGV in selection candidates using multi-trait random regression model which account for interactions between marker effects and breed of origin in the admixed Nordic Red dairy cattle. Our breed-specific model used breed proportions (BP) as random predictors and deregressed proofs of estimated breeding values (DRP) as response variables weighted by approximated reliability of DRP. Reliabilities were explored as squared correlation between DRP and DGV, weighted by the mean reliability of DRP. Estimated reliabilities were low for milk (0.32) and protein (0.32) and slightly higher (0.42) for fat. Observed reliabilities were similar to those estimated assuming homogenous structure. The Nordic Red cattle is admixed but closely related, thus, the model under investigation may have been unable to differentiate additive genetic effects by breed of origin with a medium dense marker data.

Key words: reliability, genomic breeding values, admixed populations, breed proportions

Introduction

In the past decade, genomic selection as proposed by Meuwissen *et al.* (2001) has revolutionized genetic predictions in livestock populations. Genomic selection (GS), which implies breeding value estimation of unproven bulls based on high-density marker genotypes is the current tool used to identify animals with high genetic merit with higher accuracy. The accuracy of prediction equations used to estimate marker effects, however, depend on a number of factors such as, the extent of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), size and structure of the reference population and the heritability of a trait (Goddard, 2009).

Consequently, GS has been more effective for breeds such as Holstein-Friesian, which have large reference groups and homogeneous genetic structure. In small breeds which are likely to have admixed structures, GS has not been as successful due to lack of data with ideal parameters known to impact predicted accuracies.

Recent genomic prediction studies in dairy cattle attempt to increase the reliability in small breeds by combining data from multiple populations (Hayes *et al.*, 2009; Brøndum *et al.*, 2010, Su *et al.*, 2011). Currently, prediction equations used in these studies assume that data on multiple breeds is homogeneous that is, the marker effects across populations are the same (Hayes *et al.*, 2009; Brøndum *et al.*, 2010; Su *et al.*, 2011). Genomic selection relies on the assumption that phases of LD between genetic markers and QTL are the same in the reference and target population (Calus *et al.*, 2008). Toosi *et al.* (2009) demonstrated that prediction equations derived from one breed do not estimate accurate genomic values when applied to other breeds. Suppose two or more breeds have different LD phase between markers and QTL. The expectation is that marker effects may differ in these breeds. When such populations are combined, extensive LD would exists within breeds, but due to differences in linkage phases between breeds, the underlying LD is represented by the LD across breeds, which on overall, could be less. Assuming a uniform structure for admixed populations may hamper accurate estimation of marker effects across breeds and result in low accuracy of DGV for individuals from other populations.

The population structure of the Nordic Red dairy cattle (RDC) is an admixture of the Danish Red, Swedish Red and Finnish Ayrshire cattle. In addition, the gene pool of each of these 3 populations constitutes fractions from other breeds. Although the population is admixed, when analysing it, current models assume uniform structure (Brøndum *et al.*, 2010; 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 well be improved. Therefore, the objective of this study was to evaluate the predictive ability of a model which account for interactions between marker effects and breed of origin in the estimation of direct genomic values, referred to as, multi-trait random regression in the Nordic Red dairy cattle.

Methods

Phenotype (n = 6,253) and genotype (n = 6,145) data were provided by the Nordic Cattle Genetic Evaluation (NAV) and comprised of bulls born between 1971 and 2006. The genotypic data consisted of 37,995 single nucleotide polymorphic (SNP) markers edited to remove uninformative loci. Published estimated breeding values for milk, fat and protein indices were based on 2010 March NAV routine evaluations. Deregressed proofs (DRP) of estimated breeding values for the traits, calculated as illustrated by Schaeffer (1991) and Jairath *et al.* (1998) were used as dependent variables. DRP were calculated for all bulls in the genotype data, using effective daughter contribution (EDC) as a weight. The reliability of DRP (r^2DRP), which was (estimated as $r^2DRP = \frac{EDC}{EDC+\lambda}$ where $\lambda = \frac{4-h^2}{h^2}$ was required to be at least 20% for the DRP of a bull to be estimated. Heritabilities were obtained from conventional genetic evaluations.

Breed proportions (BP) from ancestral breeds were calculated for 6,145 bulls (Lidauer *et al.*, 2006) from the full Nordic RDC pedigree. There were 13 known breeds in the gene pool of this population. The overall mean BP was calculated for each breed and mainly 3 breeds were major gene contributors with mean BP over 10%. Therefore, 4 main 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 OTHER breed. 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 from 1971 through 2001 and selection candidates were bulls born from 1996 through 2005.

Statistical Analyses

Breed proportions were used as random regression predictors and DRP as response variables weighted by a reliability of DRP which was defined as EDC. Marker data were used to calculate genomic relationship matrix all bulls, using method 1 described by VanRaden (2008), followed by addition of small constant to correct for any possible singularities and finally inverted numerically.

Estimation of direct genomic values

The model used for the estimation of genetic parameters and direct genomic values (DGV):

$$\boldsymbol{y_i = \mu + \sum_{j=1}^{4} c_{ij}\, b_j + \sum_{j=1}^{4} \sqrt{c}_{ij}\, a_{ij} + e_i}\,,$$

where y_i is the DRP of the i^{th} bull; μ is the population intercept; b_j is the fixed regression effect of breed j ($j=1,\ldots,4$); c_{ij} and $\sqrt{\mathrm{c_{ij}}}$ are the

BP and square root of the BP of bull i for breed j respectively, so that $\sum_j c_{ij} = 1$ for all i. For example, for purebreds j: $c_{ij}=1$ and $c_{ik}=0$ for all $k\neq j$; here $a = (a_{ij})$ is a vector with length of 4 times number of bulls so that each bull has a subvector with 4 breed specific elements (a_{i1}, a_{i2}, a_{i3}, a_{i4}); we assume that $\boldsymbol{a}\sim\boldsymbol{N}(\tilde{0}, \boldsymbol{G} \otimes \boldsymbol{G_0})$, where $\tilde{0}$ is a vector of zeros of length 4 times the number of bulls; G is the genomic relationship matrix; G_0 is a 4×4 diagonal matrix of breed specific variances. Covariances between base breeds were assumed to be 0. In the model e_i are random residuals with common variance across breeds, thus $\boldsymbol{e_i}\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_e^2})$ for each i with residual variance σ_e^2.

$$\boldsymbol{E}[\boldsymbol{y_i}] = \boldsymbol{\mu} + \sum_{j=1}^{4} \boldsymbol{c_{ij}}\, \boldsymbol{a_{ij}}$$

DRP were weighted with the weighting factor $\boldsymbol{k_i}$ which was EDC scaled by the variance ratio as $\boldsymbol{k_i} = \frac{\boldsymbol{EDC_i}}{\boldsymbol{\lambda}}$. Scaling was done only to improve the numerical properties of the estimation equations.

Breed-wise variance components were performed biological trait at the time and obtained estimates for G_0 and σ_e^2 were used for DGV prediction. Genetic variances for each trait were obtained as the sum of the product of breed variances and the means of BP. In the estimation of DGV, DRP for the validation bulls were masked by setting their weighting factor $\boldsymbol{k_i}$ to 0.0001 which effectively removed their phenotypic input from the estimation of DGV.

Validation of the model and estimation of reliability of DGV

DGV validation followed Interbull GEBV test, with the exception that allows one dataset to be 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 $\boldsymbol{w_i}$, where $\boldsymbol{w_i} = \frac{\boldsymbol{EDC_i}}{\boldsymbol{EDC_i+\lambda}}$. The coefficient of determination r^2 was then scaled by a constant of the average of $\boldsymbol{w_i}$ as $\boldsymbol{r^2_{DGV}} = \frac{\boldsymbol{r^2_{(DRP,\ DGV)}}}{\boldsymbol{\bar{w}}}$, where the scaling factor $\boldsymbol{\bar{w}}$ was 0.94 for milk and protein and 0.92 for fat. The scaling was intended to account for inaccuracies in the estimation of DRP since our observed value is an indication but not exactly the true value.

Results and Discussion

Estimation of genetic parameters

Relatively high genetic variance is explained by the FAY and NRF breeds (Table 1). Although for the NRF, this may have been partly due to small mean in BP for NRF in the population. Estimates of genetic variance from the data used in the prediction of DGV were ranging from 87.75 for milk to 99.31 for protein and corresponding residual variances were very high and over 2 times the expected variances (Table 2). The resulting variance ratios were large when compared to those from convectional prediction methods. Thus, estimation of variance components may have been biased because only selected proven bulls were used for analysis. In addition, use of DRP may have not recaptured all the genetic variation from the whole data.

Reliability of DGV

The objective of this study was to evaluate reliability of DGV estimated using a breed-specific model. Presented in Table 3 are average EDC (EDC_{avg}), reliability of DRP ($\bar{w}$), coefficient of regression (b_1), coefficient of determination (r^2) and the expected reliability of DGV (r^2_{DGV}) from the validation analysis. The reliabilities estimated from the current model were low for milk and protein, 0.32 and 0.32 respectively, and slightly higher for fat 0.42.

Table 1. Average BP (BP_{avg}) in the data for the four breeds and genetic variances for each breed.

Breed	BP_{avg}	Milk	Protein	Fat
SRB	(0.20)	80.84	97.32	74.05
FAY	(0.46)	90.82	104.91	96.74
NRF	(0.12)	118.218	115.99	89.47
OTHER	(0.22)	70.953	80.481	81.70

Table 2. Estimates of genetic (σ_g^2) and residual (σ_e^2) variances and variance ratios ($\hat{\lambda}$) .

Trait	σ_g^2	σ_e^2	$\hat{\lambda}$
Milk	87.75	3031.68	34.55
Protein	99.31	3068.28	30.89
Fat	87.94	2529.70	28.79

Reliabilities for milk, protein and fat were higher than those reported by Brøndum *et al.* (2010) 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., 2009). In addition to differences in reference population sizes and marker densities, all the other studies used models that assume uniform population structure.

Table 3. The average EDC (EDC_{avg}), average reliability of DRP ($\overline{w}$), coefficients of regression (b_1), coefficients of determination (r^2), and the calculated reliabilities of DGV (r^2_{DGV}) in selection candidates.

Trait	EDC_{avg}	$\overline{w}$	b_1	r^2	r^2_{DGV}
Milk	180.73	0.94	0.79	0.30	0.32
Protein	171.14	0.94	0.81	0.31	0.32
Fat	175.42	0.92	0.94	0.39	0.42

The current model was intended to extract information from the base breeds using BP given that, marker effects will depend on their breed of origin. The Nordic Red cattle population is admixed but closely related lines. Therefore, the current model may have been unable to differentiate additive genetic effects by breed of origin. This limitation may be overcome by alternatively using marker haplotypic effects for each breed instead of single markers or increasing the marker density. Low variance ratios, probably due to selected data on proven bulls, that were used for the estimation of DGV may have resulted in a downward bias in the prediction of DGV. Increasing the reference data may improve estimation of genetic parameters. Lastly, breed proportions for each bull sum to 1, as a result, BP may have been confounded with the genomic relationship matrix.

References

Brøndum, R.F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, F. & Lund, M.S. 2010. Investigation of the reliability of genomic selection using combined reference data of the Nordic Red populations. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Prodcution,* 1–6 August 2010, Leipzig, Germany.

Calus, M.P., Meuwissen, T.H.E., de Roos, A.P. & Veerkamp, R.F. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics 178***,** 553-561.

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.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R. Liu, Z. & Burnside, E.B. 1998. Genetic evaluation for herd life in Canada. *J. Dairy Sci. 81*, 550-562.

Lidauer, M., Mäntysaari, E.A., Strandén, I., Pösö, J., Pedersen, J., Nielsen, U.S., Johansson, K., Eriksson, J.-Å., Madsen, P. & Aamand, G.P. 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.

Meuwissen, T.H.E., Hayes, B.H. & Goddard, M.E. 2001. Accurate prediction of genetic values for complex traits by whole genome resequencing. *Genetics 157,* 1819-1829.

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

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

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.