Comparative assessment of methods for estimating genomic relationships and their use in predictions in an admixed population

M. L. Makgahlela^{1,3,†}, I. Strandén^{1,3}, U.S. Nielsen^{4}, M. J. Sillanpää^{1,2,5,6}, J. Juga^{1} and E. A. Mäntysaari^{3}

Departments of ^{1}Agricultural Sciences and ^{2}Mathematics and Statistics, University of Helsinki, Finland;
^{3}MTT Agrifood Research Finland, 31600 Jokioinen, Finland;
^{4}Danish Agricultural Advisory Service, Udkaersvej 15, 8200 Aarhus, Denmark;
^{5}Departments of Mathematical Sciences and ^{6}Biology, University of Oulu, Finland

Abstract

Different genomic relationship (**G**) estimators were compared within and across populations in an admixed population. By assessing relationship coefficients separately for different populations, this study found that scaling **G** with current data allele frequencies across breeds increased coefficients for individuals in distant related populations, when compared to using breed allele means calculated from breed proportions. The latter however shifted most relationships towards zero or less. The predictions of direct estimated genomic values (DGV) were unaffected regardless of allele frequencies used. Relationship coefficients that combine genomic information and polygenic variation from the pedigree slightly increased the validation reliability of DGV in the current population.

Keywords: allele frequencies, genomic relationships, genomic predictions, admixed population

Introduction

The accurate estimation of relationships plays a crucial role in any genetic estimation of breeding values (EBV). Traditional relationships calculated from the pedigree are based on expected average identity by descent sharing between individuals (Malécot, 1948) and have been applied successfully within the framework of mixed model equations (MME) for BLUP estimation of EBV. Conversely, with the increasing availability of dense genetic markers, both pedigree-based relationships (**A**) and realized relationships calculated from marker data (**G**) are available for many important animals (Hayes *et al.*, 2009). Marker-derived relationships have more variation between closely related animals because they can show realized differences in genotypes between animals. Methods to compute **G** have been proposed for genotyped animals only (VanRaden, 2008) and when non-genotyped animals are included in the same evaluation (Legarra *et al.*, 2009; Christensen & Lund, 2010).

For accurate estimation of genetic variation, relationships are defined relative to a base population where individuals are unrelated. The challenge in calculating **G** remains the unavailability of base population allele frequencies. So, in practice, currently genotyped population allele frequencies are used to make **G**, thus genotyped animals define the base population. The use of current population allele frequencies within a breed may not have major practical implications in genomic-BLUP. In the context of structured populations, the effect of using across-breed allele frequencies to scale **G** may lead to bias in the estimation of relationships. Thus, due to breeds within a multi-breed population, the current population scaling would set the average relationship across breeds to zero. However, the average relationship within breeds may be non-zero. An alternative would be to use the average of breed-specific frequencies (VanRaden *et al.*, 2011). But this would not be possible in populations that constitute mainly crossbred animals. The Nordic Red dairy cattle (RDC) comprises of three sub-populations. Majority of animals (98%) in the Nordic RDC are composite of breeds. Estimation of breed-specific allele frequencies remains a major challenge. In this study, we assess alternative methods for calculating **G** and compare predictions resulting from using different **G** matrices.

Materials and Methods

The data used were 38194 genotypes of single nucleotide polymorphism (SNP) markers for 4106 bulls. The entire RDC pedigree (>4 million animals) was used to calculate breed proportions (BP) for the bulls (Lidauer *et al.*, 2006). A more detailed description about the final 4 breeds is provided by Makgahlela *et al.* (2011). Breeds used in this study were the Swedish red (SRB), Finnish Ayrshire (FAY), Norwegian red (NRF) and the remaining breeds with BP less than 10% were combined into breed "OTHER". The pruned pedigree for genotyped bulls contained 22300 animals.

Pedigree relationships (**A**) of genotyped bulls were estimated using the pruned pedigree. Genomic relationships (**G**) were computed following approaches demonstrated by VanRaden (2008) and modifications of these methods to adapt to the admixed structure of the current population. Following VanRaden (2008) method one and using observed allele frequencies, the genomic relationship matrix **G** (named **GOF**) was computed as $\mathbf{GOF} = \mathbf{ZZ'}/\mathrm{k}$. Element of animal *i* for marker *j* in **Z** is 0-2p_j, 1-2p_j or 2-2p_j if an individual carries 0, 1, or 2 copies for the second allele, respectively. The p_j is the frequency of the second allele at SNP marker *j* and $\mathrm{k} = 2\sum_j \mathrm{p_j}(1-\mathrm{p_j})$. The use of across breeds' observed allele frequencies defines the base population to be genotyped animals with estimates of relationships approximately zero on average. Diagonals of **GOF** were multiplied by a factor 1.01 to make **GOF** positive-definite.

Two genomic matrices computed using breed allele means, named **GBM** and **GBM2**, were obtained by modifying methods one and two, respectively by VanRaden (2008). Matrix **GBM** was calculated as $\mathbf{GBM} = \mathbf{MM'}/\mathrm{k}$ where, with same notation as in Z, elements of M are 0-2p_{ij}, 1-2p_{ij} or 2-2p_{ij}. However, now p_{ij} is expected allele frequency of the j^{th} marker for individual *i* with known base breed proportions. For each breed, the allele frequencies were computed by a simple multiple regression of genotypes on breed proportions. For the **GBM2** matrix, the columns of **M** were scaled by the standard deviation of the expected marker effects, now the elements of **M*** were $\frac{-2\mathbf{p_{ij}}}{\sqrt{2\mathbf{p_{ij}}(1-\mathbf{p_{ij}})}}$, $\frac{1-2\mathbf{p_{ij}}}{\sqrt{2\mathbf{p_{ij}}(1-\mathbf{p_{ij}})}}$ and $\frac{2-2\mathbf{p_{ij}}}{\sqrt{2\mathbf{p_{ij}}(1-\mathbf{p_{ij}})}}$. In order to improve the stability, the expected allele frequencies less than 0.1 or greater that 0.9 were set to these values. Then finally, relationships were obtained as $\mathbf{GBM2} = \mathbf{M^*M^{*\prime}}/\mathrm{m}$, where m is the number of markers.

Alternative matrices **GAOF** and **GABM2** were computed by combining **GOF** and **GBM2** with pedigree-based matrix **A**, using 20% weight on **A**. To make **GOF** to the same scale as **A**, **GOF** was scaled to have the same average of diagonal elements, i.e. $\mathbf{GOF}^* = \mathbf{GOF}\frac{\sum_i \mathbf{A_{ii}}}{\sum_i \mathbf{GOF_{ii}}}$. To express **A** and **GBM2** relative to the same unique ancestral population, **GBM2** was scaled following Wright's *F*-statistics (F_{st}) as illustrated in detail by Meuwissen *et al.* (2011).

Phenotypes were individual daughter deviations (IDD) for milk, protein and fat, obtained from March 2010 official Nordic RDC genetic evaluations (NAV). The IDD are actual cow performances adjusted for fixed effects, non-genetic random effects and genetic effects of the cow's dam (Mrode & Swanson, 2004), and were computed via animal model deregression from the 305d combined EBVs (Mäntysaari *et al.*, 2011). For the validation of methods, the data were split into sets of 3300 and 806 bulls for training and validation, respectively. The training data had older bulls that were evaluated for the first time during 2005 NAV routine evaluation.

Statistical Analyses

Variance components estimation and DGV predictions were analyzed separately for each matrix using ASReml 3.0 (Gilmour *et al.*, 2009) and MiX99 (Lidauer & Strandén, 1999), respectively under the following GBLUP model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e},$$

where **y** is a vector of IDD for daughters of bulls in the reference data set, **X** and **Z** are design matrices allocating records to **b** and **a**, respectively, **b** is a vector of fixed general mean and breed regression effects, **a** is vector of breeding values for all genotyped bulls and **e** is a vector of residuals. We assumed that $\mathbf{e}\sim N(0, \mathbf{R}\sigma_e^2)$ where R is a diagonal of $\frac{1}{\text{EDC}}$. It is assumed that $\mathbf{a}\sim N(0, \mathbf{G}\sigma_a^2)$, where **G** is the genomic matrix and σ_a^2 is the additive genetic variance. Models that used **GMB2** and **GABM** included fixed breed regressions. Predicted values for all animals in this case were obtained as the sum of the animals' DGV and fixed breed regression solutions.

Results

Statistics of the diagonal elements of the pedigree ($\mathbf{A}_{\text{ii}}$)-1 and genomic data ($\mathbf{G}_{\text{ii}}$)-1 from different genomic estimators are presented in Table 1. The average $\mathbf{A}_{\text{ii}}$ was greater in the Finnish bulls (0.016) and smaller in the Danish bulls (0.007). However, these averages were *vice versa* for $\mathbf{G}_{\text{ii}}$ using observed allele frequencies from **GOF**. The mean of diagonals from **A** and **GOF** were close to zero across breeds, SWE and FIN but was 0.136 for DNK from **GOF**. The averages of diagonal elements were zero or less in all populations for **A**, **GBM**, and **GBM2**. In all cases, the tendencies observed for diagonal elements were also clear for pair-wise relationships (results not shown). Pedigree relationships had higher correlations with **GOF** across breeds (0.70) and in the Finnish bulls (0.82). **GBM** had higher correlation with **A** in the Danish population (0.86), and both estimators had a correlation of 0.78 in the Swedish population.

Figure 1 illustrates histograms for diagonal elements from different relationship estimators. There were 155 and 145 animals with $\mathbf{GOF}_{\text{ii}}$ greater than 0.6 in the combined and DNK population, respectively. Diagonals of **GBM2** were mainly in the category less than one for all animals in all populations. Thus, with **GBM2**, individuals appeared to be less homozygous.

Table 1. Statistics of diagonal elements from **A** and different **G** matrices within and across populations. Values given are deviations from 1.00.

	Mean	Minimum	Maximum
Across populations			
A	0.012	0.000	0.135
GOF	0.019	-0.129	0.379
GBM	-0.051	-0.254	0.310
GBM2	-0.242	-0.387	0.093
Danish population			
A	0.007	0.000	0.109
GOF	0.136	-0.027	0.328
GBM	-0.040	-0.173	0.310
GBM2	-0.233	-0.339	0.093
Swedish population			
A	0.008	0.000	0.081
GOF	0.006	-0.129	0.184
GBM	-0.043	-0.226	0.234
GBM2	-0.238	-0.387	0.029
Finnish population			
A	0.016	0.000	0.135
GOF	-0.021	-0.123	0.157
GBM	-0.062	-0.217	0.283
GBM2	-0.250	-0.377	0.077

Table 2 shows the correlations between EBV and DGV from different estimators in the validation data set. The estimation of breeding values from **GOF**, **GBM** and **GBM2** converged to similar solutions with correlations equal to 1.0. This includes DGVs from matrices that combined **A** and **G** information. Thus, DGV for animals were similar regardless of allele frequencies used to build **G**. The correlations between EBV and DGV were 66% from matrices with genomic information only and increased to more than 70% when weighted genomic information was combined with polygenic variation **GAOF** and **GABM2**. The validation reliabilities of DGV for all traits (not shown) were also similar between **G** matrices, however there was a slight increase in reliabilities when genomic information and pedigree data were combined.

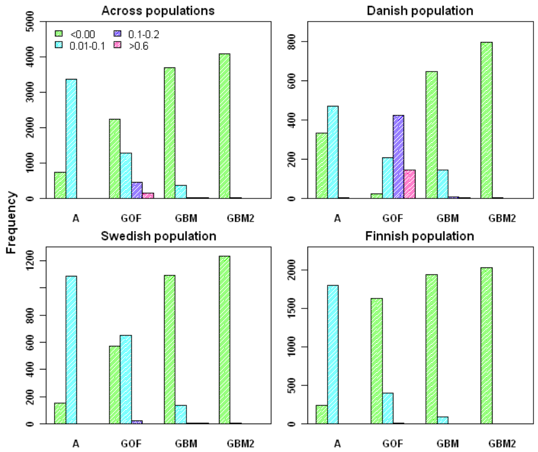

Figure 1. Histograms of (diagonal elements)-1 for **A** and **G** estimators within and across populations

Table 2. The correlations between EBV and DGV for milk in the validation bulls

	GOF	**GBM**	**GBM2**	**GAOF**	**GAB M2**
A	0.67	0.67	0.66	0.76	0.76
GOF		1.00	1.00	0.98	0.98
GBM			1.00	0.98	0.98
GBM2				0.98	0.98
GAOF					1.00
GBM2					

Discussion

The comparison of three genomic relationship estimators in a structured population using different allele frequencies showed that when populations are combined, simple observed allele frequencies across breeds tend to overestimate relationships among individuals from populations that are distantly related to the mean allele frequency. The mean allele frequency across breeds was strongly influenced by the Swedish and Finnish population since these breeds have more animals in the founder population and are more related genetically (Brøndum *et al.*, 2011; Makgahlela *et al.*, 2011). As a result, and, in contrast to pedigree relationships, the Danish population appeared to be the most inbred and related. This is unexpected because this population has been found to be more admixed than the other two, due to years of crossbreeding (Brøndum *et al.*, 2011; Makgahlela *et al.*, 2011). Thus, relationships from **GOF** have been increased between Danish bulls and were decreased for other populations. The use of estimated breed allele means from breed proportions averaged relationships similarly in all populations, although relationship coefficients were shifted more towards zero or less.

The predictions of DGV were unaffected regardless of which allele frequencies were used to calculate relationships. This has been observed previously (VanRaden, 2008; Forni *et al.*, 2011). The prediction of DGV has been found to be insensitive provided a common fixed general mean is included in the model (Strandén & Christensen, 2011). In this case, inclusion of fixed breed regressions for **GBM2** and **GABM2** has transmitted breed means back into the prediction of DGV.

In many implementations of genomic evaluations, the inclusion of polygenic effect has been found to be beneficial (Sullivan & VanRaden, 2009; Van Doormaal *et al.*, 2009). It can improve the accuracy of DGV, and reduce the bias in DGVs. In the current study, when the **G** and **A** relationship matrices were combined, it was expected that the proper scale of **G** would improve the predictions. However, we saw no difference in predictive value between DGV calculated using **GAOF** and **GABM2**. This could be because the evaluation data only included genotyped animals. The **GABM2** matrix could be more useful in single-step evaluations where most animals are evaluated by pedigree relationship matrix and through their relationships to genotyped animals.

References

Brøndum, R.F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W.F. and Lund, M.S. 2011. *Journal of Dairy Science* 94, 4700-4707.

Christensen, O. F. & Lund, M. S. 2010. *Genetic Selection Evolution* 42, 2.

Forni, S., Aguilar, I. and Misztal, I. 2011. *Genetic Selection Evolution* 43, 1.

Gilmour, A.R., Gogel, B.J., Cullis, B.R. and Thompson, R. 2009. ASREML User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Hayes, B. J., Visscher, P. M. and Goddard, M. E. 2009. *Genetics Research* 91, 47-60.

Legarra, A., Aguilar, I. and Misztal, I. 2009. *Journal of Dairy Science* 92, 4656-4663.

Lidauer, M. & Strandén, I. 1999. Fast and flexible program for genetic evaluation in dairy cattle. International workshop on high performance computing and new statistical methods in dairy cattle breeding. *Interbull Bulletin. 20*, 20.

Lidauer, M., Mäntysaari, E.A., Strandén, I., Pösö, J., Pedersen, J., Nielsen, U.S., Johansson, K., Eriksson, J.-Å., Madsen, P. and 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 Horizonte, Brazil.

Malécot, G. 1948. Les Mathématiques de l'hérédité. Masson. Paris.

Makgahlela, M. L., Mäntysaari E. A., Strandén I., Koivula M., Sillanpää M. J., Nielsen U.S. and Juga, J. 2011. Across Breed Multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *Interbull Bulletin. 44*, 42-46.

Meuwissen, T.H.E., Luan, T. and Woolliams, J.A. 2011. *Journal of Animal Breeding and Genetics* 128, 429-439.

Mäntysaari, E.A., Koivula M., Strandén, I., Pösö, J. and Aamand, G.P. 2011. Estimation of GEBVs using deregressed individual cow breeding values. *Interbull Bulletin. 44*, 26-29

Mrode, R. A. & Swanson G. J. T. 2004. *Livestock Production Science* 86, 253-260.

Strandén, I. & Christensen, O. F. 2011. *Genetic Selection Evolution* 43, 25.

Sullivan, P.G. & VanRaden, P.M. 2009. Development of genomic GMACE. *Interbull Bulletin. 40*, 157-161.

Van Doormaal, B.J., Kistemaker, G.J., Sullivan, P.G., Sargolzaei, M. and Schenkel, F.S. 2009. Canadian Implementation of genomic evaluations. *Interbull Bulletin. 40*, 214-217.

VanRaden, P.M. 2008. *Journal of Dairy Science* 91, 4414-4423.

VanRaden, P. M., Olson, K. M., Wiggans, G. R., Cole, J. B. and Tooker, M. E. 2011. *Journal of Dairy Science* 94, 5673-5682.